УДК 681.327.12

М.М. БИКОВ, В.В. КОВТУН, М.С. ФУРМАН
Вінницький національний технічний університет

МЕТОД ПРЕДСТАВЛЕННЯ ОЗНАК У АВТОМАТИЗОВАНІЙ СИСТЕМІ РОЗПІЗНАВАННЯ МОВЦЯ КРИТИЧНОГО ЗАСТОСУВАННЯ

У статті запропоновано метод екстрагування інформативних ознак для автоматизованої системи розпізнавання мовців критичного застосування з параметрів прихованих шарів глибокої нейронної мережі з bottleneck-шаром, адаптованої до індивідуальних особливостей мовця і акустичного оточення за рахунок використання і-векторів. Також запропоновано алгоритм навчання акустичних моделей на основі глибоких нейромереж з використанням екстрагованих інформативних ознак.

Ключові слова: автоматизована система розпізнавання мовців критичного застосування, розпізнавання образів, кепстральний аналіз, суміш гаусових розподілів, приховані марковскі моделі, глибокі нейромережі.

M.M. Bykov, V.V. KOVTUN, M.S. Furman
Vinnytsia National Technical University

FEATURES REPRESENTATION METHOD FOR THE AUTOMATIC SPEAKER RECOGNITION SYSTEM OF CRITICAL USE

The article suggests a method of informative features extracting for an automatic speaker recognition system of critical use from the parameters of hidden layers of a deep neural network with a bottleneck layer adapted to the individual features of the speaker and acoustic environment due to the use of i-vectors. Also, an algorithm for learning acoustic models based on deep neural networks with the use of extracted informative features is proposed. The authors considered the work of the deep neural network as a complex recognition system, which carries out a cascade of nonlinear transformations of incoming informative features and makes it the implementation of the classification procedure. Theoretical results demonstrate the possibility of using options neurons hidden layer deep neural network trained as independent informative features for recognition classifier with reduced sensitivity to perturbations in the input vectors to the increasing number of hidden layer neural network. The authors also substantiated the approaches to obtaining acoustic models of PMM based on mixtures of Gaussian distributions or on the basis of deep neural networks, the first of which potentially more accurately describes the input signal, and the second is more flexible to those present in the disturbance signal. The efficiency of the theoretical results proposed by the authors is proved experimentally.

Keywords: automatic speaker recognition system of critical use, pattern recognition, cepstral analysis, a Gaussian mixture models, hidden markov models, deep neural networks.

Вступ

Систему автоматизованого розпізнавання мовців критичного застосування [1] в загальному вигляді можна віднести до широкого класу інформаційних систем, основною характеристикою якості яких є надійність, достовірність і безпека. Надійність – властивість системи зберігати в часі у встановлених межах значення всіх параметрів, що характеризують здатність виконувати необхідні функції в заданих режимах і умовах застосування.

Серед актуальних сучасних систем розпізнавання мовців переважають текстонезалежні системи [2, 3], основані на спільному використанні сумішей гаусових розподілів (Gaussian Mixture Model, СГР) [4–8] та універсальної фонової моделі (Universal Background Model, УФМ) [9]. Основні відмінності систем полягають у застосуванні різних методів зниження розмірності отримуваних інформативних ознак і методів класифікації на основі параметрів СГР-моделей мовців. Ефективність таких систем залежить від тривалості мовних сигналів і є чутливою до акустичного оточення, що робить їх обмежено придатними до використання для створення систем розпізнавання мовця критичного застосування.

Одним з варіантів реалізації стійкішої системи розпізнавання мовців є комбінування можливостей системи незалежного від особливостей мовлення розпізнавання мови і системи розпізнавання мовця, основаної саме на виділенні особливостей мовлення. На етапі навчання такої системи створюється статистична модель мовного сигналу, здатна описувати послідовність звуків парольної фрази, а також особливості їх проголошення конкретним мовцем. Можливість створення такої системи на основі прихованих ПММ підтверджується результатами роботи [10, 11]. В роботі показано, що емісійні розподіли станів ПММ-СГР моделі парольної фрази можна апроксимувати на основі адаптації статистичної моделі голосу мовця, яка навчається класичними методами текстонезалежної ідентифікації [3]. Недоліком таких систем є надзвичайно висока ресурсоємність, адже фактично одночасно повинні працювати дві комплексні системи – розпізнавання мови і розпізнавання мовця, а також чутливістю СГР моделей до присутності у модельованому сигналі завад і суттєвій залежності ПММ до якості розмітки мовного матеріалу. Вирішення цих проблем обмежує активне упровадження відповідних систем розпізнання.

Постановка задачі дослідження

Специфіка використання автоматизованих систем розпізнавання мовців критичного застосування вимагає збереження заданих якісних показників роботи системи не зважаючи на умови використання і акустичне оточення. Існуючий математичний апарат сумішей гаусових розподілів дозоляє описати мовний сигнал із заданою точністю [12], проте не дозволяє достатньо ефективно узагальнити отриманий опис для задачі розпізнавання множини мовців. Існуючий математичний апарат глибоких нейромереж навпаки дозволяє в процесі навчання отримати узагальнені акустичні моделі, втрачаючи при цьому інформацію про індивідуальні особливості мовлення. Синтез методу об'єднання переваг описаних математичних апаратів для застосування у задачі розпізнавання мовця є актуальною задачею, яку розв'язано у представленій статті.

Виділення інформативних ознак із мовного сигналу

Створення комплексних систем розпізнавання образів, до класу яких відноситься і автоматизована система розпізнавання мовця критичного застосування, починається виділенням із запису мовного сигналу (фонограми) інформативних для розпізнавання мовця ознак. Найбільш повно описують мовний сигнал кепстральні коефіцієнти, зокрема, Мел-кепстральні коефіцієнти (Mel-Frequency Cepstral Coefficients, MFCC) [13, 14].

Їх отримання почнемо з фільтрації запису фонограми для її спектрального вирівнювання і пригнічення низькочастотного дрейфу: $y_t = x_t - ax_{t-1}$, де x_t – вхідний сигнал, y_t – сигнал після фільтрації, $a \in 0.9 \div 1$ – коефіцієнт фільтрації (зазвичай дорівнює 0,97). Далі розіб'ємо звукового сигналу на часові відрізки (фрейми) тривалістю 20 мс з постійним кроком 10 мс. Описані далі дії виконуються для кожного фрейму сигналу. Здійснюємо дискретне перетворення Фур'є $Y_k = \sum_{t=0}^{T-1} \omega_t y_t e^{-\frac{2\pi i}{T} kt}$, де T – кількість відліків у фреймі, $k = 0,1,\ldots,T/2$, ω_t – вагова функція, яка призначена для зменшення крайових ефектів, викликаних розбиттям сигналу на фрейми. У системах, пов'язаних із обробленням записів мовних сигналів, зазвичай використовується вагова функція Хеммінга $\omega_t^{hamm} = 0.54 - 0.46\cos\left(\frac{2\pi t}{T-1}\right)$, $t = 0,1,\ldots,T-1$. Створимо множину трикутних фільтрів, рівномірно розташованих за Мел-шкалою. Перетворення частот за Мел-шкалою здійснюється за відношенням $B(f) = 1125\ln\left(1 + \frac{f}{700}\right)$, а зворотне перетворення – за формулою $B^{-1}(b) = 700\left(e^{\frac{b}{1125}} - 1\right)$. Для $m = 0,1,\ldots,M-1$ трикутних фільтрів $H_m(k)$ опишемо як

$$H_m(k) = \begin{cases} 0, & k < f(m-1), k > f(m+1), \\ \dfrac{k - f(m-1)}{f(m) - f(m-1)}, & f(m-1) \le k \le f(m), \\ \dfrac{f(m+1) - k}{f(m-1) - f(m)}, & f(m) \le k \le f(m+1), \end{cases}$$

де $f(m-1)$, $f(m)$ і $f(m+1)$ – початок, середина і кінець трикутного вікна m-го фільтра, так що $f(m) = \frac{T}{F_s} B^{-1}\left(B(f_{low}) + (m+1)\frac{B(f_{high}) - B(f_{low})}{M+1}\right)$, де f_{low} і f_{high} – нижня і верхня межі аналізованого частотного діапазону, F_s – частота дискретизації сигналу. Далі обчислимо логарифм енергії спектра для згенерованої множини трикутних фільтрів $E_m = \ln\left(\sum_{k=0}^{T/2} |Y_k|^2 H_m(k)\right)$, $m = 0,1,\ldots,M-1$, і виконаємо дискретне косинусне перетворення для обчисленого логарифму енергії E_m:

$$c_n = \sum_{m=0}^{M-1} E_m \cos\left(\frac{\pi(m+0.5)n}{M}\right), \; n = 0,1,\ldots,M, \tag{1}$$

Вектор c_n містить множину кепстральних коефіцієнтів. Для опису мовного сигналу зазвичай використовують перші 10–15 з них. Для моделювання динаміки мовного сигналу вектор c_n доповнимо векторами перших d_n і других a_n похідних відповідно:

$$d_n = \frac{\sum_{l=1}^{L} l(c_{n+l} - c_{n-l})}{2\sum_{l=1}^{L} l^2}, \; a_n = \frac{\sum_{l=1}^{L} l(d_{n+l} - d_{n-l})}{2\sum_{l=1}^{L} l^2}. \tag{2}$$

Формування ПММ-СГР моделі для автоматизованої системи розпізнавання мовця критичного застосування

Врахуємо варіативність мовного сигналу в часовому просторі у створюваній системі застосувавши приховані марковські моделі, для чого визначимося із кількістю станів моделі N, множиною станів $S = \{S_1, S_2, \ldots, S_N\}$ (стан моделі у момент часу t позначатимемо q_t), множиною значень, що спостерігаються, які можуть породжуватися моделлю (спостереження в момент часу t позначатимемо o_t), розподілом імовірностей переходів між станами $A = \{a_{ij} = P(q_t = S_j | q_{t-1} = S_i)\}$, $i, j = 1,2,\ldots,N$, розподілом

імовірностей спостережень у стані S_j $P\left(o_t \middle| S_j\right)$, $j=1,2,\ldots,N$ і початковим розподілом імовірностей станів $\pi=\{\pi_i=P(q_1=S_i)\}$, $i=1,2,\ldots,N$.

У задачі розпізнавання мовця стани ПММ найчастіше моделюють фонеми парольної фрази (зазвичай 3 стани на фонему), під спостереженням йдеться про вектор інформативних ознак, а для з'ясування наскільки добре стан марковської моделі описує поточний фрейм мовного сигналу (імовірність емісії) застосуємо моделі гаусових сумішей. В цьому випадку, щільність розподілу імовірностей емісії задається сумішшю гаусових розподілів

$$b_i(o_t)=\sum_{m=1}^{M}\frac{c_{i,m}}{(2\pi)^{D/2}\left|\sum_{i,m}\right|^{1/2}}e^{-\frac{1}{2}(o_t-\mu_{i,m})^T\sum_{i,m}^{-1}(o_t-\mu_{i,m})}, \quad (3)$$

де набір параметрів b_i включає значення ваг суміші $c_{i,m}$, вектори математичних очікувань гаусіан $\mu_{i,m}$ і коваріаційні матриці гаусіан $\sum_{i,m}$.

Нехай $q_1^T=(q_1,q_2,\ldots,q_T)$ – послідовність станів СГР-ПММ, $o_1^T=(o_1,o_2,\ldots,o_T)$ – послідовність спостережень. Імовірність породження СГР-ПММ послідовності спостережень o_1^T для послідовності станів q_1^T визначається як

$$P\left(o_1^T \middle| q_1^T\right)=\prod_{t=1}^{T}b_{q_t}(o_t)=\prod_{t=1}^{T}\sum_{m=1}^{M}\frac{c_{q_t,m}}{(2\pi)^{D/2}\left|\sum_{q_t,m}\right|^{1/2}}e^{-\frac{1}{2}(o_t-\mu_{q_t,m})^T\sum_{q_t,m}^{-1}(o_t-\mu_{q_t,m})}. \quad (4)$$

Імовірність появи послідовності станів q_1^T є добутком імовірностей переходів між станами ПММ, тобто

$$P\left(q_1^T\right)=\pi_{q_1}\prod_{t=1}^{T-1}a_{q_tq_{t+1}}. \quad (5)$$

Тоді спільну імовірність появи послідовності спостережень o_1^T і послідовності станів o_1^T моделі представимо добутком імовірностей (4) і (5):

$$P\left(o_1^T,q_1^T\right)=P\left(o_1^T,q_1^T\right)P\left(q_1^T\right) \quad (6)$$

Повна імовірність появи послідовності спостережень o_1^T для даної моделі опишемо як

$$P\left(o_1^T\right)=\sum_{q_1^T}P\left(o_1^T,q_1^T\right). \quad (7)$$

Імовірність (7) можна обчислити за алгоритмом прямо-зворотного ходу [15] за час, пропорційний T. Далі здійснюватимемо навчання СГР-ПММ із врахуванням критерію максимальної правдоподібності (Maximum Likelihood) [16], налаштовуючи параметри моделі по заданій послідовності спостережень так, щоб збільшити імовірність появи цієї послідовності спостережень для модифікованої моделі. Таке навчання можна здійснити, наприклад, EM-алгоритмом. При достатній кількості параметрів моделі гаусових сумішей описують розподіл імовірностей із необхідною точністю.

Застосування глибокої нейромережі для утворення акустичної моделі мовця

Глибокі нейромережі (Deep Neural Network) [17,18] також дозволяють обчислити імовірності емісій (3). Глибокою нейромережею називають штучну нейромережу із двома або більше прихованими шарами. Нехай є глибока нейромережа із $L+1$ шарами. Позначимо вхідний шар як o, в вихідний шар як L. Для вхідного і прихованих шарів виконується залежність

$$v^l=f\left(z^l\right)=f\left(W^lv^{l-1}+b^l\right),\ 0<l<L, \quad (8)$$

де $W^lv^{l-1}+b^l\in R^{N_l}$, $v^l\in R^{N_l}$, $W_l\in R^{N_l\times N_{l-1}}$, $b^l\in R^{N_l}$ і $N_l\in R$ – відповідно, вектор індукованого локального поля, вектор активації, матриця ваг, вектор зміщення і кількість нейронів шару l; $v^0=o$ – вектор спостереження, або вектор ознак, $N_0=D$ – розмірність вектора ознак; $f(\ldots)$ – функція активації, яка поелементно застосовується до вектору індукованого локального поля. Зазвичай функція активації є сигмоїдною $\left(\sigma(z)=\frac{1}{1+e^{-z}}\right)$. Для задач класифікації кожен вихідний нейрон відповідає за $i\in\{1,2,\ldots,C\}$ класів, де $C=N_L$ – кількість класів. Значення i-го вихідного нейрона обчислюється за формулою

$$v_i^L = P_{DNN}(i|0) = soft\max_i(z^L) = \frac{e^{z_i^L}}{\sum_{j=1}^{C} e^{z_j^L}} \quad (9)$$

і інтерпретується як імовірність того, що спостереження o належіть класу i.

Подавши на вхід ГНМ вектор спостережень o, можна обчислити значення на виході ГНМ, врахувавши множину параметрів $\Theta = \{W, b\} = \{W^l, b^l | 0 < l \le L\}$, описану у рівнянні (8), для послідовного обчислення векторів активації шарів з 1 по $L-1$ включно, і за допомогою рівняння (9) використати отриманні значення для задач класифікації. Навчатимемо нейромережу налаштовуючи параметри $\Theta = \{W, b\}$ за наявними навчальними даними $S = \{(o^m, y^m) | 0 \le m < M\}$, де M – кількість наборів даних, o^m і y^m – вектори спостережень і бажаний вихідний вектор для m-го набору вхідних даних. У нашій задачі вектор y є розподілом імовірностей акустичних класів і для оцінювання якості роботи нейромережі доцільним є застосування критерію мінімізації взаємної ентропії (Cross-Entropy, CE)

$$J_{CE}(W, b, S) = \frac{1}{M} \sum_{m=1}^{M} J_{CE}(W, b; o^m, y^m), \quad (10)$$

де $J_{CE}(W, b; o, y) = -\sum_{i=1}^{C} y_i \log v_i^L$, а $y_i = P_{emp}(i|o)$ є емпіричною імовірністю того, що спостереження o належить класу i, а v_i^L – ця ж імовірність, обчислена ГНМ.

Результати досліджень [18] дозволяють стверджувати, що комбінація прихованих шарів глибокої нейромережі може виконувати функцію вилучення інформативних ознак із вхідних даних. Приховані шари ГНМ реалізують композицію простих нелінійних перетворень вхідних даних, що дозволяє утворювати комплексні закономірності, а вихідний softmax-шар виконує функції простої логлінійної класифікації. Тоді, для випадку моделювання акустичних даних, глибока нейромережа виконує двоетапний процес обчислення апостеріорної імовірності $P(s|x)$. На першому етапі вхідний вектор ознак x трансформується у вектор v^{L-1} шляхом $L-1$ нелінійних перетворень, здійснюваних $L-1$ прихованими шарами глибокої нейромережі. На другому етапі відбувається обчислення апостеріорної імовірності $P\left(s\middle|v^{L-1}\right)$ із використанням логлінейної моделі. Отже, приховані шари глибокої нейромережі екстрагують із вхідних ознак внутрішні закономірності, які ефективно класифікуються логлінейною моделлю на вихідному шарі, а процеси навчання класифікатора і екстрагування ознак відбуваються одночасно.

Нехай на вхідний шар ГНМ подано вектор ознак $x = v^0$, до якого додано збурення δ^0. Тоді поріг активації $v^l = \sigma(W^l v^{l-1} + b^l)$ для l-го $(l = 1,2,\ldots,L-1)$ прихованого шару зміниться на величину $\delta^l = \sigma(W^l(v^{l-1} + \delta^{l-1}) + b^l) - \sigma(W^l v^{l-1} + b^l) \approx diag(\sigma'(W^l v^{l-1} + b^l))(W^l)^T \delta^{l-1}$, де $\sigma(z)$ – вектор, кожна компонента якого це сигмоїда від відповідної компоненти вектора z, $diag(z)$ – діагональна матриця, на діагоналі якої розташовані компоненти вектора z. Продовжуючи перетворення отримаємо $\sigma'(W^l v^{l-1} + b^l) = \sigma(W^l v^{l-1} + b^l) \circ (1 - \sigma(W^l v^{l-1} + b^l)) = v^l \circ (1 - v^l)$, де символом $\circ$ позначено операцію поелементного добутку векторів. Отже, збурення δ^l можна оцінити як

$$\|\delta^l\| \approx \|diag(\sigma'(W^l v^{l-1} + b^l))(W^l)^T \delta^{l-1}\| \le \|diag(\sigma'(W^l v^{l-1} + b^l))(W^l)^T\| \|\delta^{l-1}\| = \|diag(v^l \circ (1 - v^l))(W^l)^T\| \|\delta^{l-1}\| \quad (11)$$

З виразу (11) випливає, що мале збурення у вхідних даних буде зменшуватися з кожним прихованим шаром. Отже, внутрішні закономірності, які екстрагуються прихованими шарами глибокої нейромережі з вхідних даних, стають менш чутливими до збурень у вхідних даних із зростанням кількості прихованих шарів. Отже, на прихованих шарах глибокої нейромережі, близьких до вхідного шару, екстрагуються низькорівневі ознаки, які презентують локальні шаблони, чутливі до незначних змін вхідних ознак. Відповідно, ознаки, які екстрагуються на прихованих шарах глибокої нейромережі, близьких до вихідного шару, носять більш абстрактний характер і стають інваріантними до змін вхідних ознак, зо робить ці ознаки менш цікавими для задачі розпізнавання мовця. Проте, слід також встановити поріг величини збурень вхідних даних, перевищивши який глибока нейромережа вже не зможе коректно інтерпретувати вхідні дані.

Метод екстрагування ознак глибокої нейромережею із урахуванням ПММ-СГР акустичної моделі

Враховуючи вище сказане, є можливість використати як інформативні ознаки для навчання акустичної моделі вектори активацій одного з прихованих або вихідного шарів глибокої нейромережі. Наприклад, перетворені відповідним способом імовірності фонем, які утворюються вихідним шаром

нейромережі з одним прихованим шаром, можна застосувати як вектор ознак для навчання СГР-ПММ акустичної моделі, але отриманий таким чином вектор параметрів буде дуже великої розмірність. Зробити отримуваний вектор параметрів компактнішим можна увівши до архітектури глибокої нейромережі bottleneck-шар [17] – прихований шар із невеликою кількістю нейронів із, зазвичай, лінійними функціями активації, який розташовується між прихованими шарами глибокої нейромережі. Параметри нейронів bottleneck-шару можна об'єднати із традиційними ознаками (наприклад, кепстральними коефіцієнтами) і, застосувавши до утвореного вектору параметрів відомі процедури компактифікації даних (наприклад, сингулярний розклад (Singular Values Decomposition, SVD) [1]), використати отриманий компактний вектор параметрів для точнішого навчання СГР-ПММ акустичних моделей. Автори припускають, що параметри bottleneck-шару глибокої нейромережі, які забезпечили отримання оптимальної СГР-ПММ акустичної моделі мовлення, будуть інформативні для і для задачі розпізнавання мовців. Причому ефективність bottleneck-ознак можна варіювати деталізуючи СГР-ПММ акустичну модель. На основі вищенаведених міркувати синтезовано послідовність операцій для виділення інформативних ознак з bottleneck-шару глибокої нейромережі для задачі розпізнавання мовців:

1. Виділення кепстральних коефіцієнтів (1)–(2) для навчання СГР-ПММ моделі.
2. Навчання СГР-ПММ моделі (3).
3. Вибір ознак для навчання глибокої нейромережі (ці ознаки можуть відрізнятися від використаних для навчання СГР-ПММ моделі) і приведення вхідних даних для навчання глибокої нейронної мережі до нульового середнього і одиничної дисперсії.
4. Ініціалізація навчання глибокої нейромережі із L прихованими шарами за критерієм мінімізації взаємної ентропії (10).
5. Створення *i*-векторів для навчальної бази мовців і приведення *i*-векторів до нульового середнього і одиничної дисперсії.
6. Розширення вхідного шару навченої на етапі 4 глибокої нейромережі з ініціалізацією відповідних коефіцієнтів матриці ваг нульовими значеннями.
7. Донавчання глибокої нейромережі з розширеним вхідним шаром за ознаками, до яких на кожному фреймі додано *i*-вектор, який відповідає цьому фрейму фонограми. При цьому у цільову функцію введено доданок $R(W)$, який штрафує відхилення ваг W^l моделі, що навчається, від значень ваг $\widetilde{W}^l$ вихідної моделі:

$$R(W) = \lambda \sum_{l=1}^{L+1} \left\| vec\left(W^l - \widetilde{W}^l\right) \right\|^2 = \lambda \sum_{l=1}^{L+1} \sum_{i=1}^{N_l} \sum_{j=1}^{N_{l-1}} \left(W_{ij}{}^l - \widetilde{W}_{ij}{}^l\right)^2, \quad (12)$$

де $vec(W)$ – вектор, отриманий злиттям всіх стовпців матриці W, λ – величина штрафу.

1. Розбиття обраного прихованого шару l глибокої нейромережі на два шари за правилом $v^l = f\left(W^l v^{l-1} + b^l\right) \approx f\left(W_{out}^l \left(W_{bn}^l v^{l-1} + 0\right) + b^l\right)$. В результаті отримаємо перший шар – шар із лінійною функцією активації, матрицею ваг W_{bn}^l і нульовим вектором зсувів, і другий шар – нелінійний шар з матрицею ваг W_{out}^l і вектором зсувів b^l, із розмірністю вихідного шару. Для розбиття застосуємо сингулярний розклад матриці ваг $W^l = USV^T \approx \widetilde{U}_{bn}\widetilde{V}_{bn}^T = W_{out}^l W_{bn}^l$, де індекс bn означає знижену розмірність. Отже, вихідна глибока нейромережа із L прихованими шарами перетворюється на глибоку нейромережу із $L+1$ прихованими шарами та лінійним bottleneck-шаром l.
2. Донавчання глибокої нейромережі із bottleneck-шаром із штрафом (11) на відхилення ваг від ваг вихідної моделі.
3. Відкидання шарів глибокої нейромережі, що слідують за bottleneck-шаром.
4. Використання отриманої нейронної мережі із вихідним bottleneck-шаром для формування інформативних ознак.

I-вектор, введений з п.5 алгоритму, призначений для персоналізації СГР-ПММ акустичної моделі для можливості її використання у задачі розпізнавання мовця. Вектор акустичних ознак x_t утворюється з моделі гаусових сумішей з діагональними коваріаційними матрицями, названої універсальною фонової моделлю (Universal Background Model, УФМ), яка навчається за великим обсягом фонограм: $x_t \sim \sum_{k=1}^{K} c_k N\left(\mu_k(0); \Sigma_k\right)$. Вектор акустичних ознак $x_t(s)$, що належить мовцю s, утворюється із адаптований до цього мовця моделі гаусових сумішей $x_t(s) \sim \sum_{k=1}^{K} c_k N\left(\mu_k(s); \Sigma_k\right)$. Отже, індивідуальний для мовця s *i*-вектор $w(s)$ визначається із лінійної залежності $\mu_k(s) = \mu_k(0) + T_k w(s)$ між залежними від мовця математичними очікуваннями $\mu_k(s)$ і незалежними від мовця математичними очікуваннями $\mu_k(0)$. Таким чином, *i*-вектор кодує відмінність щільності розподілу імовірностей акустичних ознак, оціненої за фонограмою, від еталонної. Додавання до вектора акустичних ознак *i*-вектора, обчисленого за фрагментом

фонограми, що відповідає певному мовцю, забезпечує адаптацію СГР-ПММ акустичної моделі до індивідуальних особливостей мовця і акустичного оточення.

Інтерпретація результатів роботи глибокої нейромережі

Отримані на основі ПММ-ГНМ моделювання імовірності необхідно інтерпретувати в контексті призначення автоматизованої системи розпізнавання мовців критичного застосування. Для цього необхідно сформулювати вирішальне правило, яке на основі імовірностей, генерованих акустичними моделями та універсальними фоновими моделями, видаватиме номер мовця, якому належить фонограма із записом парольної фрази $\hat{\omega}$. Правило сформулюємо так:

$$\hat{\omega} = \arg\max_{\omega} P\left(\omega|x\right) = \arg\max_{\omega} \frac{P\left(x|\omega\right)P\left(\omega\right)}{P\left(x\right)} = \arg\max_{\omega} P\left(x|\omega\right)P\left(\omega\right), \quad (12)$$

де максимум береться по всіх можливим ланцюжкам слів парольної фрази ω, а $x = (x_1, x_2, \ldots, x_T)$ є набором векторів ознак сигналу, що використовується для розпізнавання СГР-ПММ системою, або набір векторів комплексних ознак для ГНМ-ПММ систем. Імовірність $P(\omega)$ є еталонною імовірністю спостереження ланцюжка слів парольної фрази ω, а імовірність $P(x|\omega) = \sum_q P(x|q,\omega)P(q|\omega) \approx \max_{q|\omega} \pi(q_0)\prod_{t=1}^{T} a_{q_{t-1}q_t} \prod_{t=0}^{T} P(x_t|q_t)$ генерується акустичної моделлю. Для пошуку максимально правдоподібною послідовності станів прихованої марковської моделі використовується алгоритм Вітербі [3]. Відзначимо, що у ГНМ-ПММ системі глибока нейромережа видає апостеріорну імовірність $P(q_t|x_t)$, тоді як нам необхідна імовірність $P(x_t|q_t)$. Але її можна отримати скориставшись теоремою Байєса [17]: $P(x_t|q_t) = \frac{P(q_t|x_t)P(x_t)}{P(q_t)}$, де імовірністю $P(x_t)$ в контексті задачі розпізнавання мовця можна знехтувати тому, що ця імовірність не залежить від послідовності слів у парольній фразі ω, а $P(q_t)$ – апріорна імовірність спостереження q_t, яку обрахуємо із навчальних даних $P(q_t) = {}^{N_{q_t}}\!/_{N}$, де N_{q_t} – кількість фреймів, що містять парольну фразу, N – всі фрейми навчальних даних.

Для балансування імовірностей, генерованих акустичними і універсальними фоновими моделями, використаємо ваговий коефіцієнт λ, тоді правило (12) виглядатиме як $\hat{\omega} = \arg\max_{\omega}\left(\ln P(x|\omega) + \lambda \ln P(\omega) - \omega_{penalty} n(\omega)\right)$, де ω_{penalty} – штраф за розпізнавання нового слова (щоб уникнути розбиття довгих слів на велику кількість коротких), $n(\omega)$ – кількість слів у ланцюжку ω.

Отже, на основі вищенаведених теоретичних результатів авторами синтезовано автоматизовану систему розпізнавання мовця критичного застосування на основі

Планування експерименту та аналіз результатів

Для емпіричного оцінювання ефективності отриманих теоретичних результатів авторами здійснено програмну реалізацію запропонованих алгоритмів у вигляді текстозалежної автоматизованої системи розпізнавання мовця критичного застосування із реалізацію вище запропонованого ПММ-ГНМ підходу до отримання акустичних моделей мовців.

В якості бази еталонних записів, які піддавалися на вхід створеної системи, використано записи із безкоштовної бази даних NOIZEUS – спеціалізованої бази даних Школи інжинірингу та комп'ютерних наук Еріка Джонсона при Університеті Техасу в Далласі, США, яка використовується для дослідження алгоритмів покращення звуку і складається з 30 речень англійської розмовної мови, вимовлених трьома чоловіками та трьома жінками (по 5 на кожного диктора, частота дискретизації записів складає 25 кГц, але задля додавання шуму була зменшена до 8 кГц) та записів типових побутових та техногенних шумів. В ході експерименту автоматизовану систему розпізнавання мовців критичного застосування навчали як записами чистих парольних фраз, так і парольними фразами із додаванням шумів. В першому випадку навчальна вибірка містила 18 парольних фраз, у другому – 576, де до чистого сигналу додавався штучний шум з рівнями шум/сигнал 0 дБ, 5 дБ, 10 дБ, 15 дБ відповідно. Для навчання моделей використано 60% обсягу бази аудіозаписів, в яку увійшли екземпляри записів без шумів та із різним рівнем шум/сигнал (5, 10, 15 дБ) відповідно. Тестувальна вибірка склала решту 40% аудіозаписів. Ідентифікація мовця вважалася вірною якщо модель і фонограма із парольною фразою належали одному мовцеві і розпізнаний системою лінгвістичний вміст парольної фрази збігалася із еталонним. Проведено дві етапи досліджень: перший етап передбачав оцінювання імовірності правильного розпізнавання мовців за індивідуальними особливостями мовлення із сталою структурою парольної фрази для всіх мовців, але із варіювання рівнем шумів у фонограмах, другий етап досліджень передбачав варіювання структури парольної фрази без зміни акустичного середовища.

В якості базової для експериментів взято глибоку нейромережу *dnn_b* із 6 прихованими шарами по 2048 нейронів із сигмоїдними функціями активації, яку навчено на 36-мірних ознаках, отриманих з базової СГР-ПММ моделі *gmm_b* з [12], взятих з 11 фреймів фонограм (центральний фрейм і по 5 фреймів зліва і справа відносно центрального). Адаптована модель глибокої нейромережі *dnn_a* навчено на вхідних ознаках

базової моделі, доповнених *i*-вектором, зі штрафом 10^{-8} на відхилення ваг від значень базової моделі. I-вектори довжиною 100 значень отримано для навчальних і тестових записів на основі УФМ із 512 гаусіанами, яку навчено на 12-мірних MFCC-ознаках (1), доповнених першими і другими похідними (2). У адаптовану модель глибокої нейромережі із застосуванням сингулярного розкладу матриці ваг 3-го прихованого шару додано лінійний шар з 80 нейронів. Отримана в результаті bottleneck-нейромережа використовувалася для ініціалізації навчання моделі *dnn_a* +. Навчання проводилося зі штрафом 10^{-8} на відхилення ваг від значень нейронної мережі *dnn_a*. Ця глибока нейромережа після видалення шарів, наступних за bottleneck-шаром, використовувалася для побудови 80-мірних оптимізованих ознак. На оптимізованих ознаках навчено СГР-ПММ модель *gmm_o* з тією ж кількістю гаусіан і зв'язаних станів, що і в базової моделі *gmm_b*. Також оптимізовані ознаки, отримані з 31 фрейму, проріджених через 5 фреймів (тобто [-15 -10 -5 0 5 10 15], де 0 – центральний фрейм), використані для навчання ГНМ-ПММ моделі *dnn_o* з 4 прихованими шарами по 2048 нейронів ss сигмоїдними функціями активації та ініціалізацією навчання із застосуванням обмежених машин Больцмана [17].

Результати тестування створених моделей для двох запланованих етапів експериментів наведено на рис. 1.

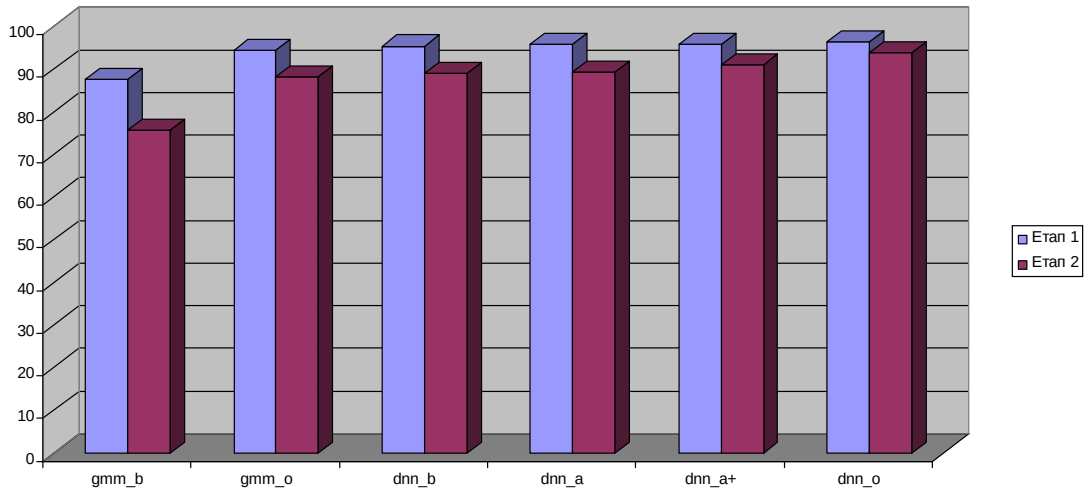

Рис. 1. Залежність імовірності правильного розпізнавання мовців автоматизованою системою розпізнавання мовців критичного застосування від обраної акустичної моделі для 1 та етапів досліджень

Висновки

В даній роботі автори поглибили інтеграцію математичного апарату сумішей гаусових розподілів для використання у системах розпізнавання мовців критичного застосування інтегрувавши його із математичним апаратом прихованих марковських моделей, що дозволило використати для розпізнавання особи мовця не лише фізіологічні особливості мовлення, які моделювалися СГР, а і лінгвістичну інформацію, описувану ПММ. Також авторами обґрунтовано підходи до отримання акустичних моделей ПММ на основі сумішей гаусових розподілів або на основі глибоких нейромереж, перший з яких потенційно точніше описує вхідний сигнал, а другий гнучкіший до присутніх у сигналі збурень.

Автори розглянули роботу глибокої нейромережі як комплексної системи розпізнавання, яка здійснює каскад нелінійних перетворень вхідних інформативних ознак і звершує його здійсненням процедури класифікації. Отримано теоретичні результати, які доводять можливість використання параметрів нейронів прихованих шарів навченої глибокої нейромережі як самостійних інформативних ознак для розпізнавання зі зниженням чутливості класифікатора до збурень у вхідному векторів зі зростанням кількості прихованих шарів нейромережі.

Для зменшення розмірності отриманих із прихованих шарів глибокої нейромережі інформативних ознак із збереженням їх ефективності запропоновано здійснювати перетворення архітектури глибокої нейромережі із застосуванням bottleneck-шару, параметри якого отримуються із застосуванням сингулярного розподілу.

Розроблено метод екстрагування інформативних ознак з параметрів прихованих шарів глибокої нейронної мережі з bottleneck-шаром, адаптованої до індивідуальних особливостей мовця і акустичного оточення за рахунок використання *i*-векторів. Також запропоновано алгоритм навчання акустичних моделей на основі глибоких нейромереж з використанням екстрагованих інформативних ознак. Ефективність запропонованих авторами теоретичних результатів доведено експериментально.

Література

1. Ковтун В.В. Оцінювання надійності автоматизованих систем розпізнавання мовців критичного застосування / М.М. Биков, В.В. Ковтун // Вісник Вінницького політехнічного інституту. – 2017. – № 2. – С. 70–76.

2. Брагина Е.К. Современные методы биометрической аутентификации: обзор, анализ и определение перспектив развития / Е. К. Брагина, С. С. Соколов // Вестник АГТУ. – 2016. – № 61. – С. 40–45.

3. Матвеев Ю.Н. Технологии биометрической идентификации личности по голосу и другим

модальностям / Ю.Н. Матвеев // Вестн. МГТУ им. Н. Э. Баумана. Сер. „Приборостроение“. – 2012. – № 3(3). – С. 46–61.

4. A short tutorial on Gaussian Mixture Models [Електронний ресурс]. – Режим доступу : http://www.computerrobotvision.org/2010/tutorial_day/GMM_said_crv10_tutorial.pdf.

5. Kenny P., Boulianne G., Ouellet P., Dumouchel P. Speaker and Session Variability in GMM-Based Speaker Verification // IEEE Transact. on Audio, Speech, and Language Processing. 2007. Vol. 15, N 4. P. 1448–1460.

6. Vogt R. J., Lustri C. J., Sridharan S. Factor Analysis Modelling for Speaker Verification with Short Utterances // Proc. Of Speaker and Language Recognition Workshop “Odyssey-2008”. Stellenbosch, South Africa, 2008. P. 1–5.

7. Kanagasundaram A., Vogt R., Dean D.B., Sridharan S., Mason M. W. I-vector based speaker recognition on short utterances // Proc. of 12th Annual Conf. of International Speech Communication Association (INTERSPEECH 2011). Firenze Fiera, Florence, 2011. P. 2341–2344.

8. Kanagasundaram A., Vogt R.J., Dean D. B., Sridharan S. PLDA based speaker recognition on short utterances // Proc. of Speaker and Language Recognition Workshop “Odyssey-2012”. Singapore, 2012. P. 28–33.

9. Larcher A.O., Bonastre J.-F., Mason J. S. D. From GMM to HMM for embedded password-based speaker recognition // Proc. 16th Europ. Signal Processing Conf. (EUSIPCO-2008). Lausanne, Switzerland, 2008. P. 1–5.

10. Juang B. H., Rabiner L. R. Hidden Markov Models for Speech Recognition // Technometrics. 1991. Vol. 33, № 3. P. 251–272.

11. Subramanya A., Zhengyou Z., Surendran A. C., Nguyen P., Narasimhan M., Acero A. A Generative-Discriminative Framework using Ensemble Methods for Text-Dependent Speaker Verification // Proc. of the Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP-2007). Honolulu, Hawaii, USA, 2007. Vol. 4. P. 225–228.

12. Ковтун В.В. Використання множини мікрофонів у автоматизованій системі розпізнавання мовця критичного застосування / М.М. Биков, В.В. Ковтун // Вісник Вінницького політехнічного інституту. – 2017. – № 3. – С. 67–73.

13. Paliwal K.K., Atal B.S. Frequency related representation of speech // Proc. EUROSPEECH, pp. 65–68 Sep. (2003).

14. Fukuda T., Takigawa M., Nitta T. Peripheral features for HMMbased speech recognition // Proc. ICASSP, 1: pp. 129–132 (2001).

15. Baum L. Statistical inference for probabilistic functions of finite state Markov chains / L. Baum, T. Petrie // Ann. Math. Statist. – 1966. – Vol. 37, no. 6. – P. 1554–1563.

16. Dempster A. Maximum-likelihood from incomplete data via the EM algorithm / A. Dempster, N. Laird, D. Rubin // J. Roy. Stat. Soc. Ser. B. – 1977. – Vol. 39, no. 1. – P. 1–38.

17. Grézl F. Adaptation of multilingual stacked bottle-neck neural network structure for new language / F. Grézl, M. Karafiát, K. Veselý // Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – 2014. – P. 7654–7658.

References

1. Kovtun V.V. Otsiniuvannia nadiinosti avtomatyzovanykh system rozpiznavannia movtsiv krytychnoho zastosuvannia / M.M. Bykov, V.V. Kovtun // Visnyk Vinnytskoho politekhnichnoho instytutu. – 2017. – № 2. – S. 70–76.

2. Brahyna E.K. Sovremennye metody byometrycheskoi autentyfykatsyy: obzor, analyz y opredelenye perspektyv razvytyia / E. K. Brahyna, S. S. Sokolov // Vestnyk AHTU. – 2016. – № 61. – C. 40–45.

3. Matveev Yu.N. Tekhnolohyy byometrycheskoi ydentyfykatsyy lychnosty po holosu y druhym modalnostiam / Yu.N. Matveev // Vestn. MHTU ym. N. E. Baumana. Ser. „Pryborostroenye“. – 2012. – № 3(3). – S. 46–61.

4. A short tutorial on Gaussian Mixture Models [Elektronnyi resurs]. – Rezhym dostupu : http://www.computerrobotvision.org/2010/tutorial_day/GMM_said_crv10_tutorial.pdf.

5. Kenny P., Boulianne G., Ouellet P., Dumouchel P. Speaker and Session Variability in GMM-Based Speaker Verification // IEEE Transact. on Audio, Speech, and Language Processing. 2007. Vol. 15, N 4. P. 1448–1460.

6. Vogt R. J., Lustri C. J., Sridharan S. Factor Analysis Modelling for Speaker Verification with Short Utterances // Proc. Of Speaker and Language Recognition Workshop “Odyssey-2008”. Stellenbosch, South Africa, 2008. P. 1–5.

7. Kanagasundaram A., Vogt R., Dean D.B., Sridharan S., Mason M. W. I-vector based speaker recognition on short utterances // Proc. of 12th Annual Conf. of International Speech Communication Association (INTERSPEECH 2011). Firenze Fiera, Florence, 2011. P. 2341–2344.

8. Kanagasundaram A., Vogt R.J., Dean D. B., Sridharan S. PLDA based speaker recognition on short utterances // Proc. of Speaker and Language Recognition Workshop “Odyssey-2012”. Singapore, 2012. P. 28–33.

9. Larcher A.O., Bonastre J.-F., Mason J. S. D. From GMM to HMM for embedded password-based speaker recognition // Proc. 16th Europ. Signal Processing Conf. (EUSIPCO-2008). Lausanne, Switzerland, 2008. P. 1–5.

10. Juang B. H., Rabiner L. R. Hidden Markov Models for Speech Recognition // Technometrics. 1991. Vol. 33, № 3. P. 251–272.

11. Subramanya A., Zhengyou Z., Surendran A. C., Nguyen P., Narasimhan M., Acero A. A Generative-Discriminative Framework using Ensemble Methods for Text-Dependent Speaker Verification // Proc. of the Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP-2007). Honolulu, Hawaii, USA, 2007. Vol. 4. P. 225–228.

12. Kovtun V.V. Vykorystannia mnozhyny mikrofoniv u avtomatyzovanii systemi rozpiznavannia movtsia krytychnoho zastosuvannia / M.M. Bykov, V.V. Kovtun // Visnyk Vinnytskoho politekhnichnoho instytutu. – 2017. – № 3. – S. 67–73.

13. Paliwal K.K., Atal B.S. Frequency related representation of speech // Proc. EUROSPEECH, pp. 65–68 Sep. (2003).

14. Fukuda T., Takigawa M., Nitta T. Peripheral features for HMMbased speech recognition // Proc. ICASSP, 1: pp. 129–132 (2001).

15. Baum L. Statistical inference for probabilistic functions of finite state Markov chains / L. Baum, T. Petrie // Ann. Math. Statist. – 1966. – Vol. 37, no. 6. – P. 1554–1563.

16. Dempster A. Maximum-likelihood from incomplete data via the EM algorithm / A. Dempster, N. Laird, D. Rubin // J. Roy. Stat. Soc. Ser. B. – 1977. – Vol. 39, no. 1. – P. 1–38.

17. Grézl F. Adaptation of multilingual stacked bottle-neck neural network structure for new language / F. Grézl, M. Karafiát, K. Veselý // Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – 2014. – P. 7654–7658.

Рецензія/Peer review : 20.09.2017 р. Надрукована/Printed :27.10.2017 р.
Рецензент: д.т.н., проф. О.В. Бісікало