УДК 004.522

КОМП‘ЮТЕРНА СИСТЕМА КОНТРОЛЮ ДОСТУПУ З ВИКОРИСТАННЯМ АВТОРИЗАЦІЇ ЗА ГОЛОСОМ

М. І. Горбійчук, Р. Р. Соловій

Івано-Франківський національний технічний університет нафти і газу, вул. Карпатська, 15, м. Івано-Франківськ, 76019

Розглянуті основні методи і алгоритми ідентифікації особи за голосом, а також розроблено комп‘ютерну систему контролю доступу, яка дозволяє вдосконалити стандартні парольні системи шляхом запровадження механізмів голосової авторизації. Удосконалено алгоритм виділення фрагментів голосу в аудіофайлах. Розроблено схему зберігання голосових даних користувачів з метою їх подальшого використання під час авторизації, підсистему голосової авторизації та розроблено відповідне програмне забезпечення комп‘ютерної системи.
Ключові слова: голос, ідентифікація особи за голосом, голосова авторизація, контроль доступу, вектор ознак голосу.

Рассмотрены основные методы и алгоритмы идентификации личности по голосу, а также разработана компьютерная система контроля доступа, которая позволяет усовершенствовать стандартные парольные системы путем внедрения механизмов голосовой авторизации. Усовершенствован алгоритм выделения фрагментов голоса в аудиофайлах. Разработана схема речевых данных пользователей с целью их дальнейшего использования при авторизации, подсистема голосовой авторизации и разработано соответствующее программное обеспечение компьютерной системы.
Ключевые слова: голос, идентификация личности по голосу, голосовая авторизация, контроль доступа, вектор признаков голоса.

The article is devoted to techniques and algorithms for voice identification and development of computer access control system that allows to improve the standard password system by establishing mechanisms for voice authorization. Improved algorithm for the selection of fragments of the voice in the audiofiles. Designed the scheme of storing users voice data for further use during the authorization and designed the access control system for voice authorization, which allows to determine the fact of belonging language user-defined sequence and investigated the performance of the developed hardware and software.
Keywords: voice, speaker identification, voice authorization, access control, voice feature vector.

Вступ. У міру розвитку комп'ютерних технологій та у зв'язку з широким розповсюдженням комп'ютерів і використанням їх в різних сферах життя особливо гостро постає питання обмеження доступу користувачів до конфіденційної інформації. Основним способом авторизації користувача в більшості комп‘ютерних систем є вказівка його логіна та пароля. Проте такі системи авторизації мають суттєві недоліки, пов‘язані з конфіденційністю облікових даних користувачів. Так, наприклад, користувачі часто забувають паролі, зберігають їх у невідповідному місці, або в найгіршому випадку їх можуть просто викрасти. Згідно статистики багатьох компаній більшість дзвінків у службу підтримки пов'язана із забутими або втраченими паролями. Тому проблема надійного розпізнавання і підтвердження особи є актуальною.

Сьогодні все частіше стандартні парольні системи захисту замінюються або доповнюються біометричними системами ідентифікації особи.

Ідентифікація особи за голосом – це один з видів біометричної аутентифікації, що грунтується на індивідуальних особливостях людського голосу і мови. Основними перевагами таких систем є низька ціна, високий рівень безпеки, зручність для користувачів, доступність, простота використання, можливість віддаленого доступу.

Використання систем ідентифікації особи за голосом є найбільш зручним для користувача способом аутентифікації, який грунтується на індивідуальних фізіологічних особливостях мовного апарату людини. Такі системи авторизації дозволяють вирішити проблеми, пов‘язані з конфіденційністю облікових даних

користувачів, розпізнаванням і підтвердженням особи в комп‘ютерних системах.

Якщо для функціонування інших біометричних систем безпеки необхідне спеціальне дороговартісне обладнання (наприклад, сканер відбитків пальців, сканер райдужної оболонки ока тощо), або ж для них характерний складний процес відбору та аналізу біометричних зразків (наприклад, аналіз ДНК), то все, що потрібно для функціонування системи ідентифікації особи за голосом - це спеціальне програмне забезпечення і мікрофон, підключений до комп’ютера чи іншого мобільного пристрою.

Унікальність людського голосу зумовлена індивідуальною формою та розмірами ротової і носової порожнин, горла, органів дихання, формою грудної клітки з дихальними м'язами і діафрагмою, гортані з голосовими зв'язками, положення язика тощо [1,2,3]. Крім того, вимова кожної людини індивідуальна і визначається соціальними та психічними факторами. Сформовані в юності особливості мови, інтонації стають звичними і майже не змінюються протягом усього життя. Тому виділяють два види ознак голосу: низькорівневі (зумовлені анатомічною будовою мовного апарату) і високорівневі (пов'язані з манерою вимови).

Мета дослідження. Метою даної роботи є розроблення простої комп‘ютерної системи контролю доступу за голосом, що дасть змогу вдосконалити стандартні парольні системи та вирішити проблему конфіденційності облікових даних користувачів шляхом запровадження механізмів голосової авторизації.

Основна частина. Результат ідентифікації особи за голосом повністю залежить від вхідних даних, математичних алгоритмів та обчислювальної потужності. Під вхідними даними розуміють зразок голосу особи, отриманий за допомогою запису з мікрофона. Якість такого зразка залежить від типу пристрою введення (наприклад, професійний мікрофон або мобільний телефон) і навколишнього середовища (гучна вулиця або тихе приміщення). Математичні алгоритми використовуються для того, щоб порівняти отриманий голосовий зразок із зразками в базі даних. Під обчислювальною потужністю розуміють швидкість і якість обробки біометричних ознак користувача, що залежить від апаратних особливостей системи.

Системи ідентифікації особи за голосом поділяються на два основних види: текстозалежні і текстонезалежні. У текстозалежних методах особа повинна сказати одну і ту ж парольну фразу під час навчання системи і під час розпізнавання голосу. Текстонезалежні системи можуть ідентифікувати особу незалежно від того, що вона сказала.

На сьогоднішній день існує декілька підходів до ідентифікації людини за голосом, які базуються на аналізі структури голосового сигналу. Процедура обробки голосового сигналу полягає в використанні короткочасного аналізу, тобто сигнал розбивається на фрагменти (фрейми) фіксованого розміру. Потім до кожного вікна застосовуються алгоритми виділення ознак. Більшість популярних систем ідентифікації використовують як вектори ознак мел-частотні кепстральні коефіцієнти (MFCC) або коефіцієнти лінійного передбачення (LPCC) [2]. Дані методи грунтуються на виділенні векторів ознак голосового сигналу з урахуванням особливостей сприйняття звуку людським вухом. Ще одним методом аналізу голосових фрагментів є аналіз формантних частот. Аналіз формантних частот є одним найдавніших методів ідентифікації особи за голосом, а ідентифікацію особи при формантному підході найчастіше проводять на голосних звуках, в яких можна ефективно виділити форманти [2, 4, 5].

Завдання ідентифікації особи за голосом полягає у перевірці статистичної гіпотези, яку можна сформулювати наступним чином [6]. Припустимо, що невідома особа вимовила вислів X і представляється користувачем S. Дві протилежні гіпотези, в такому випадку матимуть вигляд:

$$\begin{cases} H_0 : \textit{вислів } X \textit{ вимовив користувач } S, \\ H_1 : \textit{вислів } X \textit{ вимовив не користувач } S. \end{cases} \quad (1)$$

Комп‘ютерна система повинна визначити, яку з двох гіпотез необхідно прийняти і відповідно прийняти рішення про те, чи надавати користувачу S доступ до ресурсів системи.

Роботу системи контролю доступу можна розділити на чотири етапи:

- нормалізація вхідного мовного сигналу;
- виділення характерних ознак голосу;
- побудова моделі джерела мови (режим навчання);
- прийняття за побудованою моделлю та новою вхідною послідовністю одної з двох гіпотез (1).

Важливим етапом голосової аутентифікації є попередня підготовка вхідного голосового

сигналу до виділення характерних ознак. Одним з основних методів нормалізації вхідного голосового сигналу є видалення із аудіофайлу фрагментів, що не містять голос (фрагменти тишини). Для цього вхідний звуковий сигнал (рис. 1) проходить через детектор голосової активності (Voice Activity Detector, VAD), що дозволяє виділити фрагменти голосу і цим зменшити навантаження на комп'ютерну систему та збільшити її швидкодію за рахунок зменшення кількості зайвих обчислень над ділянками аудіофайлу, що не містять корисної інформації про фізіологічні особливості мовного апарату особи.

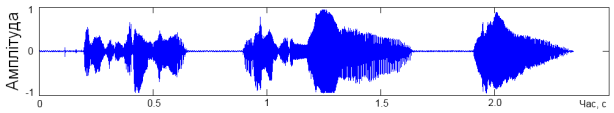

Рисунок 1 – Вхідний аудіо сигнал

Для виділення фрагментів голосу використано алгоритм, який грунтується на статистичній оцінці вхідного аудіосигналу. Зазвичай всі записані за допомогою мікрофона аудіофайли починаються з фрагмента, що не містить голосу (тишина). Зумовлено це тим, що в процесі авторизації чи реєстрації в системі користувач реагує на початок запису фрагментів голосу з деяким запізненням. Даний початковий фрагмент аудіо файлу, тривалістю 100-200 мс, є вхідним параметром алгоритму і містить інформацію про зовнішнє середовище, в якому відбувається запис зразків голосу. Згідно даного підходу значення амплітуди аудіосигналу розглядаються як випадкові величини, які залежать від великої кількості факторів, кожен з яких вносить невеликий внесок, а згідно центральної граничної теореми такі випадкові величини (в даному випадку амплітуда сигналу) мають розподіл близький до нормального з параметрами (μ, σ).

Згідно правила «трьох сигма» практично всі значення нормально розподіленої випадкової величини лежать в інтервалі $x-3\sigma; x+3\sigma$.

На наступному етапі всі значення амплітуди сигналу, для яких не виконується таке правило, розглядаються як голос, а всі інші – як фрагменти тишини. Такий підхід має власні недоліки пов'язані з коректністю виділення фрагментів голосу в аудіофайлах. Згідно цього підходу фрагменти голосу починаються там, де значення амплітуди знаходяться за межами інтервалу $x-3\sigma; x+3\sigma$. Проте, у процесі експериментального дослідження було встановлено, що в аудіофайлах часто трапляються ділянки з дуже незначними відхиленнями, або їх кількість невелика (5-10 значень амплітуди, що еквівалентно тривалості приблизно 0,000625 с при частоті дискретизації 8000 Гц). Такі ділянки аудіофайлу не можна розглядати як фрагмент голосу, оскільки середня тривалість звучання букв в слові становить десятки мілісекунд, слів - сотні мілісекунд, або навіть і декілька секунд.

Таким чином, може виникнути ситуація, коли в аудіофайлі тривалістю 30-60 с знайдеться декілька десятків або сотень таких фрагментів, які послідовно йдуть один за одним і при видаленні пауз утвориться відносно великий фрагмент аудіофайлу, який алгоритм помилково вважатиме фрагментом голосу, хоча насправді він не нестиме жодної корисної інформації про особу. Тому було прийнято рішення про удосконалення такого підходу наступним чином. Аудіо файл розбивається на рівні фрагменти тривалістю 25-30 мс, а далі кожне значення амплітуди поточного фрагменту оцінюється згідно правила «трьох сигма». Для кожного фрейму створюється тимчасовий масив значень типу boolean, який містить тільки значення «true» або «1», якщо правила «трьох сигма» виконується, і «false» або «0», якщо така умова не виконується. На наступному етапі здійснюється обчислення ймовірності появи елемента з значенням «true» P_1 та ймовірності появи значення «false» P_0. Ймовірності обчислюються шляхом знаходження відношення кількості появи того, чи іншого значення до загальної кількості значень в масиві (довжина масиву, або кількість значень амплітуди в фрагменті). Якщо значення P_1 менше деякого порогового значення α, то вважається, що даний фрагмент містить голос, а якщо ні, то тишину.

Значення α в визначалось шляхом експериментального дослідження результатів роботи алгоритму виділення фрагментів голосу в аудіо файлах і за результатами експериментів було прийнято значення α рівне 0,65. Параметр α можна інтерпретувати таким чином: якщо 65% значень амплітуди аудіосигналу в фрагменті знаходиться за межами інтервалу $x-3\sigma; x+3\sigma$, то система приймає рішення про те, що фрагмент містить голос, а у протилежному випадку – фрагмент містить паузи (тишина).

Після нормалізації сигналу необхідно виділити ознаки, які характеризують особливості мовного апарату конкретного користувача. У сфері цифрової обробки сигналів (DSP, Digital Signal Processing) і, зокрема,

розпізнавання мови та ідентифікації особи за голосом, найбільш активне застосування знайшли так звані мел-частотні кепстральні коефіцієнти (MFCC, Mel-frequency cepstral coefficients). Основна ідея методу MFCC полягає в максимальному наближенні інформації, що надходить на вхід системи до інформації, що надходить в слуховий аналізатор людського мозку.

Мовний сигнал спочатку являє собою масив значень амплітуд, отриманий за допомогою дискретизації вихідного аналогового сигналу з певною частотою Fs за допомогою аналогово-цифрового перетворювача (АЦП) звукової карти. Наступним кроком є поділ вхідного сигналу на фрейми, як правило, тривалістю 25-30 мс [1]. Фрейми перекривають один одного на 25-70%. Перекриття фреймів використовується для того, щоб компенсувати втрати інформації на початку і кінці кожного фрейму, які відбуваються у результаті застосування віконної функції на наступному кроці алгоритму.

У більшості задач цифрової обробки немає можливості досліджувати сигнал на нескінченному інтервалі. Обмеження інтервалу аналізу рівносильне добутку вихідного сигналу на віконну функцію [1, 2, 7]. Віконна функція використовується для того, щоб уникнути неприродних розривів в мовних фрагментах звукового файлу і відповідно спотворень в спектрі аудіосигналу [3]. Множення вихідного сигналу на значення віконної функції дозволяє послабити значення амплітуди на обох кінцях поточного фрейму і цим запобігти різкій зміні значень в кінцевих точках. Мова йде про те, що алгоритм швидкого перетворення Фур'є [6] передбачає, що сигнал є неперервним і періодичним, а в даному випадку сигнал ділиться на фрейми фіксованої довжини і для того, щоб уникнути спотворень в спектрі використовується віконна функція.

Як віконна функція використане вікно Хеммінга, оскільки воно найчастіше використовується в задачах розпізнавання мови та ідентифікації особи за голосом [2, 3]:

$$w_n = 0,54 - 0,46 \cdot \cos\left(2\pi \frac{n}{N-1}\right), n = 0,..., N-1,$$

де N – довжина вікна.

На рис. 2 зображено вхідний сигнал та його спектрограму. На горизонтальній осі спектрограми представлено час в секундах, по вертикальній осі – частота, а колір кожної точки зображення визначається значенням амплітуди на певній частоті в конкретний момент часу t.

Для кожної людини при вимові звуків, букв характерний власний набір (комбінація) частот і завдяки цьому людське вухо здатне відрізняти один звук від іншого і, зокрема, голоси різних людей. Для того, щоб виділити даний набір частот для кожного фрейму вхідного аудіосигналу використовується мел-частотний аналіз. Згідно даного методу отримане представлення сигналу в частотній області розбивають на діапазони за допомогою банку (гребінки) трикутних фільтрів. Межі фільтрів розраховують в шкалі Мел, яка є результатом досліджень здатності людського вуха до сприйняття звуків на різних частотах [2, 4].

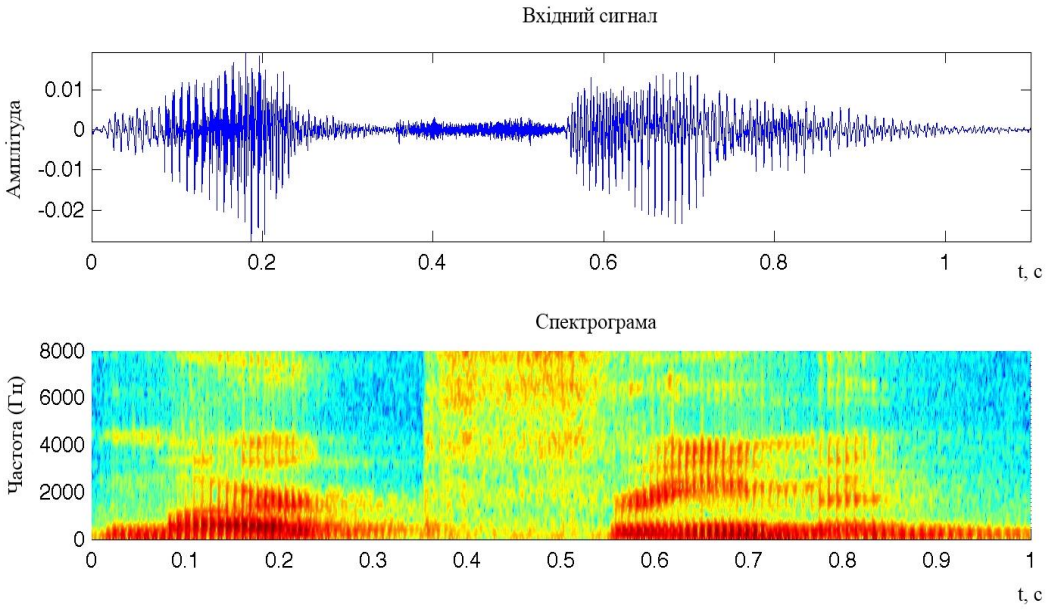

Рисунок 2 – Спектрограма вхідного сигналу

Перетворення в мел-шкалу здійснюється за формулою:

$$B(f) = 1127 \cdot \ln\left(1 + \frac{f}{700}\right).$$

Вважається, що інформація, яку несуть низькочастотні компоненти мовного сигналу є більш важливою в порівнянні з високочастотними компонентами і тому мел-шкала лінійна до 1 кГц і логарифмічна вище 1 кГц, тобто на низьких частотах, фільтри застосовуються лінійно, тоді як на високих частотах – логарифмічно. Дані фільтри нерівномірно розташовані на осі частот, тому у таких фільтрів більше в ділянках спектру з низьким рівнем частот (до 1 кГц) і менше області високих частот (понад 1 кГц) [2, 6, 8]. Фільтри застосовуються до квадратів модулів коефіцієнтів перетворення Фур'є, а отримані значення логарифмуються:

$$e_m = \ln\left(\sum_{k=0}^{N} |X_k|^2 H_{m,k}\right), m = 0,..., N_{FB-1},$$

де N_{FB} – кількість фільтрів (найчастіше використовують близько 24 фільтрів [2]), $H_{m,k}$ – вагові коефіцієнти отриманих фільтрів.

Такий підхід дозволяє частково виключити складові шуму в частотній області, оскільки найбільш важливі частоти людського голосу знаходяться в діапазоні від 70 Гц до 3400 Гц.

Останній етап процесу виділення ознак голосу полягає в застосуванні дискретного косинусного перетворення (DCT, Discrete Cosine Transform), результатом якого буде множина мел-частотних кепстральних коефіцієнтів (MFCC), які і будуть елементами векторів ознак голосу особи:

$$c_i = \sum_{m=0}^{N_{FB-1}} e_m \cos\left(\frac{\pi i(m+0.5)}{N_{FB}}\right), i = 0,..., N_{MFCC},$$

де e_m – логарифмовані значення коефіцієнтів перетворення Фур‘є, NMFCC – кількість коефіцієнтів (розмір векторів ознак).

Отримання кепстральних коефіцієнтів можна також здійснювати за допомогою алгоритму IFFT (інверсне швидке перетворення Фур‘є), проте в даному випадку використовується алгоритм DCT, який є більш ефективним, оскільки не використовує роботу з комплексними числами [2, 4, 6].

У результаті для кожного фрагменту з вихідного мовного сигналу отримуємо скінченну множину мел-частотних коефіцієнтів кепстра $C_i=(C_1;C_2;... C_N)$, яка містить N елементів і є вектором характерних ознак голосу конкретного користувача. В даній роботі використовувалось значення N=13.

На рис. 3 представлено множину векторів ознак голосу для конкретної особи. Кожен вектор містить 13 мел-частотних кепстральних коефіцієнтів. Всього на графіку зображено 50 векторів ознак: по осі OX міститься порядковий номер коефіцієнта, а по осі OY його значення.

Розпізнавання за голосом відрізняється від інших систем тим, що в даному випадку предметом розпізнавання є процес, а не статичне зображення як у випадку з розпізнаванням відбитків пальців або райдужної оболонки ока. Тому найчастіше зразок голосу представляється не у вигляді єдиного вектора ознак, а у вигляді послідовності векторів ознак, кожен з яких описує характеристики невеликої ділянки мовного сигналу [3, 4, 9]. Послідовність векторів, отримана після етапу обробки сигналу, використовується для побудови моделі голосу особи.

Основним параметром, який використовується для ідентифікації користувача є міра подібності двох звукових фрагментів (вхідний зразок та зразок в базі даних) [2, 3, 5]. Отримані в процесі авторизації вектори ознак розглядається як неточна копія векторів ознак збережених в базі даних системи.

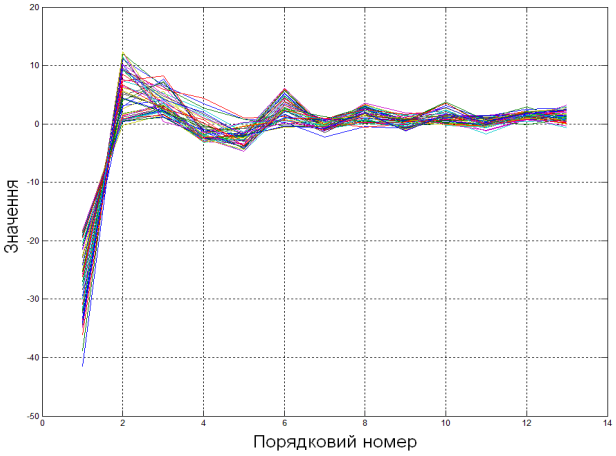

Рисунок 3 – Множина векторів ознак голосу

Метою наступного етапу є побудова моделі голосу користувача на основі унікального для нього вектора характерних ознак голосу. Для вирішення цієї задачі було використано підхід на основі моделей гаусових сумішей (GMM, Gaussian Mixture Model). Це найбільш поширений підхід, заснований на апараті математичної статистики. GMM-моделі широко використовуються в області машинного

навчання, інтелектуального аналізу даних (data mining) та при розпізнаванні графічних образів. Крім того, цей підхід можна використовувати в текстонезалежних системах, тобто при авторизації користувач говорить не одну й ту саму парольну фразу кожен раз, а різну. Даний метод не вимагає значних обчислювальних ресурсів і володіє достатньою точністю розпізнавання.

Алгоритм GMM приймає на вхід послідовність векторів ознак голосу, отриманих за допомогою алгоритму M_{FCC} і використовує їх для створення моделі голосу особи, або GMM-моделі. Метою даного методу є моделювання функції щільності розподілу значень елементів векторів ознак голосу. У цьому методі складний розподіл моделюється за допомогою зваженої суми Гауссіанів [4]:

$$p\left(x|\lambda\right) = \sum_{i=1}^{K} w_i \cdot p_i\left(x|\lambda_i\right),$$

де λ – параметри моделі голосу особи, K – кількість компонентів моделі, x – множина векторів характерних ознак голосу, p_i – щільність розподілу i-ї компоненти, w_i – ваги компонентів такі, що $\sum_{i=1}^{K} w_i = 1$.

Розмірність кожного з розподілів D збігається з розмірністю вектора характерних ознак, а щільність розподілу кожної компоненти – це D-мірний Гауссіан [4]:

$$p_i\left(x|\lambda_i\right) = \frac{1}{(2\pi)^{D/2} \cdot \left|\sum_i\right|^{1/2}} \times \\ \times exp\left(-\frac{1}{2}\left(x-\mu_i\right)^T \sum_i^{-1}\left(x-\mu_i\right)\right),$$

де μ_i – вектори математичних сподівань, $\sum_i$ – матриця коваріації.

На рис. 4 показаний графік Гауссіану, який побудований з використання даних отриманих для певного фрагменту голосу.

Найчастіше в системах, що реалізують цю модель, використовується діагональна матриця коваріації з елементами на головній діагоналі σ^2 [2, 3, 4, 7].

Кожен користувач в системі буде представлений власною моделлю з параметрами λ:

$$\lambda = \{w_i, \mu_i, \sum_i\},\ i = \overline{1, K}.$$

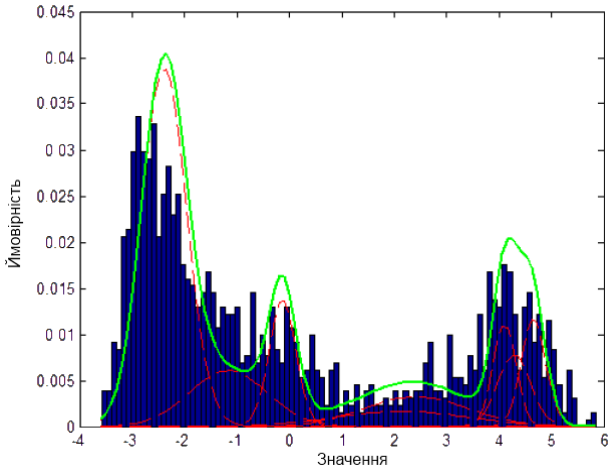

Рисунок 4 – Апроксимація функції щільності розподілу кепстрального коефіцієнта C_i

Отже, для побудови моделі голосу особи необхідно визначити вектори математичних очікувань, матриці коваріації і ваги компонентів. Для визначення параметрів даної моделі існує кілька методів, найбільш поширеним з яких є метод максимальної правдоподібності [2,4]. Завдання методу полягає у знаходженні по заданих навчальних даних таких параметрів моделі, при яких функція правдоподібності моделі досягає максимуму. Для послідовності з T навчальних векторів $X = \{x_1, ..., x_T\}$ функція правдоподібності (GMM likelihood) може бути записана так:

$$p(x|\lambda) = \prod_{t=1}^{T} p(x_t|\lambda). \qquad (2)$$

Дану задачу найчастіше вирішують за допомогою алгоритму EM (Estimation-Maximization) [4]. На вхід подається навчальна послідовність векторів ознак $X = \{x_1, ..., x_T\}$, параметри моделі ініціалізуються початковими значеннями, а потім на кожній ітерації алгоритму відбувається переоцінка параметрів у результаті отримують параметри $\bar{\lambda}$ нової моделі (2), такі що $p(x|\lambda) \le p(x|\bar{\lambda})$. Далі отримані параметри $\bar{\lambda}$ моделі (2) стають початковими для наступної ітерації. Процес отримання параметрів λ моделі голосу називається тренуванням, або навчанням.

Переоцінка параметрів відбувається за такими формулами [2]:

– обчислення апостеріорних ймовірностей (Estimation-step):

$$p(i \mid x_t, \lambda) = \frac{w_i p_i(x_t)}{\sum_{k=1}^{K} w_k p_k \left(x_t \right)};$$

– обчислення нових параметрів моделі (Maximization-step):

$$\mu_i = \frac{\sum_{t=1}^{T} p(i \mid x_t, \lambda) \cdot x_t}{\sum_{t=1}^{T} p(i \mid x_t, \lambda)},$$

$$\sum_i = \frac{\sum_{t=1}^{T} p(i \mid x_t, \lambda)(x_t - \mu_i)(x_t - \mu_i)^T}{\sum_{t=1}^{T} p(i \mid x_t, \lambda)}.$$

Після етапу навчання в системі створюється готова модель голосу з параметрами λ, яка відповідає конкретній особі. Тепер для тестової послідовності векторів ознак голосу (наприклад, під час авторизації) можна обчислити значення функції правдоподібності $p(S \mid H_0)$ згідно з співвідношенням (1). На наступному етапі необхідно обчислити значення функції $p(S \mid H_1)$, яке дорівнює ймовірності того, що вислів X вимовив не користувач S. Одним з методів вирішення цієї задачі є використання універсальної фонової моделі.

Універсальна фонова модель (Universal Background Model, UBM) [3, 4, 8] – це модель, основне завдання якої полягає в тому, щоб охарактеризувати всі можливі джерела мови в системі. Дана модель створюється аналогічно до моделі звичайних користувачів, проте навчається на великій кількості зразків голосу від безлічі джерел мови, а сам процес навчання займає відносно тривалий час.

Тоді значення функції правдоподібності обчислюється за такою формулою:

$$p(S \mid H_1) = p(X \mid \lambda_{UBM}),$$

де λ_{UBM} – параметри універсальної фонової моделі; X – це множина векторів ознак особи, що намагається авторизуватися.

Підхід на основі GMM-моделей показує хороші результати в процесі ідентифікації / верифікації особи за голосом лише у випадку великої кількості вхідних даних на стадії навчання (60-90 с для кожної особи) [1, 2, 4], тому рекомендована тривалість аудіозаписів під час реєстрації становить не менше 1 хвилини.

На основі отриманої моделі голосу користувача та деякого порогового значення Q_s система повинна прийняти рішення про надання доступу для користувача, який намагається авторизуватися. Залежно від конкретних завдань сформований результат може надходити або на виконавчий механізм, або в підсистему авторизації.

Процедура прийняття рішення може виглядати так [4, 5]:

$$Рішення = \begin{cases} H_0, L < Q_s, \\ H_1, L \geq Q_s, \end{cases}$$

де L – міра подібності моделей голосу зареєстрованого користувача і користувача, що намагається авторизуватися:

$$L = \frac{\log p(X \mid H_0)}{\log p(X \mid H_1)},$$

$$\log p(X \mid \lambda) = \frac{1}{T} \sum_{t=1}^{T} \log p(X \mid \lambda).$$

Величина порогу входження Q_s для користувача S визначається із зразків для навчання, які отримуються в процесі реєстрації в системі. Варто зауважити, що кількість помилок при ідентифікації особи за голосом залежить не тільки від порогу входження Qs, методів виділення характерних ознак голосу чи методів побудови моделі голосу користувача, але й від зовнішніх факторів таких як навколишнє середовище (шум, ревербація), спотворення мікрофона, завади каналу зв‘язку тощо.

У загальному випадку комп‘ютерна система міститиме такі компоненти: мікрофон (для запису голосових зразків), підсистема фільтрів та виявлення голосової активності, підсистема розпізнавання голосових відбитків, база голосових відбитків та виконавчий механізм. Структурна схема розробленої комп‘ютерної системи показана на рис. 5.

Комп‘ютерна система працює в двох режимах: режим навчання та режим авторизації.

В режимі навчання користувач попередньо реєструється в системі, записавши зразки власного голосу, які аналізуються системою з метою виділення характерних ознак. На основі виділених ознак будується шаблон (модель голосу), який дозволяє ідентифікувати користувача в майбутньому.

В режимі авторизації користувач намагається ввійти в систему, пред’являючи ідентифікатор у вигляді логіну та зразок голосу. Система аналізує даний зразок, порівнює з

зразками представленими в режимі навчання і намагається ідентифікувати особу за голосом. Якщо особу вдалося розпізнати, то система приймає рішення про надання доступу.

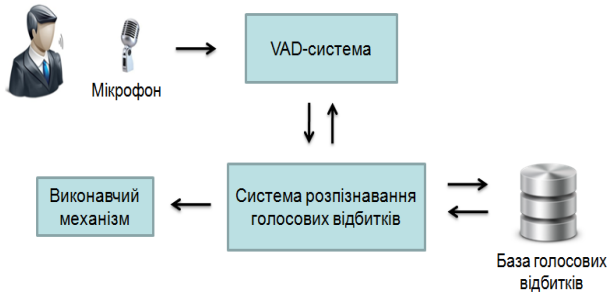

Рисунок 5 – Структура комп‘ютерної системи

Одним з найбільш ефективних способів підвищення рівня безпеки системи є використання механізмів багатофакторної аутентифікації, яка ґрунтується на одночасному використанні декількох факторів аутентифікації (знання секрету, володіння предметом, фізичні особливості особи), що значно підвищує захищеність системи. Тому було прийнято рішення про запровадження механізмів багатофакторної аутентифікації шляхом введення додаткового пароля. Тобто комп‘ютерна система матиме два рівні захисту (порівняння біометричних даних та перевірка пароля) і це є однією з її переваг у порівнянні з аналогічними системами. В момент реєстрації в системі користувач матиме змогу додатково вказати пароль, який потрібно буде вводити в процесі авторизації. Проте, дана опція є необов‘язковою і користувач може використовувати для авторизації тільки голос.

Одне з головних завдань, яке ставилося в процесі розробки програмного забезпечення системи, полягає в забезпеченні можливості одночасного використання сервера голосової аутентифікації декількома клієнтськими програмами. Для цього розроблено спеціальний формат передачі повідомлень між клієнтом і сервером та розроблено спеціальну модель зберігання даних зареєстрованих користувачів в сховищі даних App Engine DataStore.

ВИСНОВКИ

В результаті проведених досліджень розглянуто популярні методи і алгоритми ідентифікації особи за голосом, а також розроблено структурну схему та архітектурне рішення комп‘ютерної системи контролю доступу, яка дозволяє вдосконалити стандартні парольні системи шляхом запровадження механізмів голосової авторизації. Вдосконалено алгоритм виділення фрагментів голосу в аудіофайлах. Розроблено схему зберігання голосових даних користувачів з метою їх подальшого використання під час авторизації, підсистему голосової авторизації, що дозволяє визначити факт приналежності мовної послідовності заданому користувачу та розроблено відповідне програмне забезпечення.

Системи голосової аутентифікації можуть широко використовуватись в таких сферах, як банківські послуги, при підтвердженні банківських операцій, в системах «розумний дім» тощо.

1. Ю.Н. Хитрова. Применение речевой биометрии в системах ограничения доступа. [Електронний ресурс] Режим доступу: http://www.e-expo.ru/docs/sp/bat/data/media/18_ru.pdf. 2. Е.А. Первушин. Обзор основных методов распознавания дикторов. [Електронний ресурс] Режим доступу: http://msm.univer.omk.su/jrn2324/pervushinOverview.pdf. 3. Л.Р. Рабинер. Цифровая обработка речевых сигналов / Л.Р. Рабинер, Р.В. Шафер. - М.: Радио и связь, 1981. 4. Reynolds D. A. Speaker identication and verication using Gaussian mixture speaker models / D. A. Reynolds. – Helsinki: Speech Commun, 1995. 5. Kinnunen T. Spectral Features for Automatic Text-independent Speaker Recognition / T. Kinnunen. – Helsinki: Gaudeamus, 2013. 6. Bishop C. Pattern Recognition and Machine Learning / C. Bishop. – New-York: Business Media, 2006.7. Campbell J.P. Speaker Recognition: A Tutorial // Proceedings of the IEEE. - 1997. - Vol. 85. - P. 1437-1462. 8. Садыхов Р.Х. Модели гауссовых смесей для верификации диктора по произвольной речи / Р.Х. Садыхов, В.В. Ракуш. – Минск: Белкнига, 2003. 9. Przybocki M. The NIST 1999 Speaker Recognition Evaluation – An Overview // Digital Signal Processing. - 2000. - Vol. 10.

Поступила в редакцію 31.10.2014р.

Рекомендували до друку: докт. техн. наук, проф. Мещеряков Л.І. та докт. техн. наук, проф. Пістун Є.П.