DOI: 10.18372/2310-5461.59.17950
УДК 621.391

О. Ю. Лавриненко, канд. техн. наук
Національний авіаційний університет
orcid.org/0000-0002-7738-161X
e-mail: oleksandrlavrynenko@gmail.com;

Д. І. Бахтіяров, канд. техн. наук, доц.
Національний авіаційний університет
orcid.org/0000-0003-3298-4641
e-mail: bakhtiiaroff@tks.nau.edu.ua;

Г. Ф. Конахович, д-р техн. наук, проф.
Національний авіаційний університет
orcid.org/0000-0002-6636-542X
e-mail: heorhii.konakhovych@npp.nau.edu.ua;

В. Є. Курушкін, канд. техн. наук
Національний авіаційний університет
orcid.org/0009-0000-4411-0509
e-mail: vitaliy.kurushkin@npp.nau.edu.ua

АНАЛІЗ ЕФЕКТИВНОСТІ СИСТЕМИ ГОЛОСОВОЇ ІДЕНТИФІКАЦІЇ НА ОСНОВІ MFCC ТА GMM-SVM ЗА УМОВ ВПЛИВУ ЗАВАД У КАНАЛІ ЗВ'ЯЗКУ

Вступ

На сьогодні вразливість приватних осіб, організацій і держави щодо загроз в інформаційній безпеці особливо зростає при використанні інформаційних мереж – як загального користування, так і корпоративних. Цьому сприяє також тенденція до розподіленого опрацювання даних, що розширюється та пов'язана з використанням дистанційного режиму і телекомунікаційних технологій, зокрема, розширюється сфера діяльності співробітників і залучених осіб поза відповідною організацією. Дедалі більших масштабів набувають кримінальні напрями комп'ютерної діяльності, до цих напрямів можна віднести комп'ютерне шахрайство, несанкціонований доступ до інформації, підробку комп'ютерної інформації, несанкціоноване перехоплення даних та інші види злочинних дій. У зв'язку з цим надважливим завданням стає створення та застосування нових ефективних методів і засобів захисту інформації в галузі інформаційно-телекомунікаційних мереж [1].

Розвиток нових методів і засобів забезпечення інформаційної безпеки покликаний насамперед запобігти загрозам доступу до інформаційних ресурсів сторонніх осіб, які не мають доступу. Для вирішення цього завдання необхідна наявність ідентифікаторів і створення процедур ідентифікації для всіх користувачів. Сучасні ідентифікація та автентифікація включають в себе різні системи і способи біометричної ідентифікації особи. Розвиток систем ідентифікації особи, що ґрунтуються на біометричних вимірах, пов'я-заний із цілим комплексом переваг: такі системи надійніші, оскільки біометричні показники складніше підробити; сучасна мікропроцесорна техніка робить біометричні методи зручнішими порівняно зі звичайними методами ідентифікації, і зрештою, їх значно простіше піддавати автоматизації вимірювань [2].

Однією з найпоширеніших біометричних характеристик людини є її голос, що має набір індивідуальних особливостей, які відносно легко піддаються вимірюванню, наприклад, частотний спектр мовного сигналу (МС). До переваг голосової ідентифікації відносяться також зручність застосування і використання, досить невисока вартість пристроїв, що застосовуються для ідентифікації, наприклад мікрофон.

Можливості ідентифікації особи за голосовими даними охоплюють вельми широкий спектр завдань, що виділяє їх серед інших біометричних систем. Перш за все, голосова ідентифікація досить давно і широко використовується в різних системах розмежування доступу до фізичних об'єктів та інформаційних ресурсів. Голосова ідентифікація є частиною окремого наукового напряму – акустичної теорії мовлення. Перспективним видається її нове застосування в телекомунікаційних системах для дистанційної взаємодії з різними інтернет-сервісами. Як приклад, у мобільному зв'язку за допомогою голосу можна здійснювати управління послугами, при-

чому впровадження голосової ідентифікації сприяє захисту від шахрайства [3].

Особливе місце ідентифікація особи за голосом посідає при розслідуванні злочинів, зокрема у сфері комп'ютерної інформації, та при формуванні доказової бази такого розслідування. У цих випадках часто виникає необхідність проведення ідентифікації невідомого голосового запису. Проведення голосової ідентифікації – важливе практичне завдання під час пошуку підозрюваного за записом голосу зробленого в телекомунікаційних мережах. Визначення таких характеристик за голосом диктора, як стать, вік, національність, діалект, емоційне забарвлення мови, також є важливими у сфері криміналістики та антитерористичних дій. Результати ідентифікації важливі при проведенні фоноскопічних експертиз, при здійсненні експертного криміналістичного дослідження на основі теорії криміналістичної ідентифікації [4].

Істотний інтерес представляє розвиток методів голосової ідентифікації для суміжних напрямів, а саме, для нових мовних технологій, пов'язаних із розпізнаванням усного мовлення, управлінням комп'ютерними системами за допомогою голосових команд тощо.

Аналіз останніх досліджень та постановка проблеми

Незважаючи на широку застосованість і перераховані вище переваги, використовувані методи ідентифікації особи за голосовими даними мають низку серйозних недоліків, проаналізувати які і спрямована ця наукова стаття. До них відносяться, перш за все, невисока розрізняльна здатність методів і значний відсоток помилок як першого роду (помилково відкинуті особи, які мають право на допуск), так і найнебезпечнішого другого роду (особи, яких помилково допускають до конфіденційної інформації, а права на допуск до неї не мають). Особливо ускладнює ситуацію проведення ідентифікації в реальних умовах, що супроводжуються набором несприятливих зовнішніх факторів [5, 6].

Ідентифікація особи за голосом, що проводиться в реальних умовах, зустрічається з такими серйозними труднощами. По-перше, при такій ідентифікації виникають всілякі апаратні спотворення і завади, зумовлені особливостями апаратури і пристроїв для запису, обробки і зберігання інформації. По-друге, на МС неминуче накладаються зовнішні акустичні шуми, які можуть істотно спотворювати індивідуальні інформативні характеристики. З огляду на це системи ідентифікації, що демонстрували досить високу ефективність у лабораторних умовах, під час аналізу мовної інформації із зовнішніми шумами можуть показати надійність, значно нижчу. Нарешті, у низці завдань доводиться проводити ідентифікацію у вельми складних умовах накладення голосів кількох дикторів, зокрема з близькими акустичними характеристиками. Зазначимо, що дослідження можливостей голосової ідентифікації для цього найскладнішого випадку практично не проводилися [7, 8].

Таким чином, проведення голосової ідентифікації охоплює комплекс технічних, алгоритмічних і математичних методів, що охоплюють усі етапи, починаючи із запису голосу і закінчуючи класифікацією голосових даних. Розглянуті труднощі та недоліки приводять до висновку, що подальший розвиток систем голосової ідентифікації потребує розроблення нових підходів, спрямованих на опрацювання великих масивів експериментальних МС, їхній ефективний аналіз і надійну класифікацію. Це свідчить про актуальність досліджень зі створення нових математичних методів оброблення, аналізу та класифікації голосових даних, які б забезпечували надійність і достовірність ідентифікації особи в умовах впливу шумової обстановки та завад у каналі зв'язку інформаційно-телекомунікаційних мереж. Саме для аналізу та вирішення вищевикладених наукових проблем спрямоване дане дослідження [9].

Запропонована система голосової ідентифікації особи на основі MFCC та GMM-SVM

Запропонована система голосової ідентифікації особи має два режими роботи: режим навчання і режим розпізнавання (тестування). Ці режими входять в структурну схему системи голосової ідентифікації особи (рис. 1), завдання, якої полягає у виконанні наступних етапів: 1) поділ МС на часові кадри та виділення ділянок активної мови зі знаходженням значень зміни короткочасної енергії та кількості перетинів нуля між суміжними кадрами МС (англ. Short-Time Energy and Zero-Crossing Rate, STE-ZCR); 2) адаптивна вейвлетфільтрація МС (англ. Adaptive Wavelet Thresholding, AWT) для вирішення задачі шумоочищення, де необхідно провести адаптивну генерацію мікролокальних порогів, що дасть змогу зменшити вплив адитивного шуму на чисту форму МС та виділення ознак розпізнавання, де як інформативні ознаки розпізнавання МС під час автоматичної ідентифікації особи за голосом використовуються мел-частотні кепстральні коефіцієнти (англ. Mel-

Frequency Cepstral Coefficients, MFCC), які засновані на двох ключових поняттях – кепстр та мел-шкала.

Подальша поведінка системи залежить від режиму роботи. Якщо система знаходиться в режимі навчання, отримані на етапі виділення ознаки розпізнавання MFCC, зберігаються в базу даних MFCC. При знаходженні системи в стані розпізнавання, набір ознак розпізнавання MFCC МС, який був вимовлений особою, послідовно порівнюється з усіма наборами ознак розпізнавання з бази даних MFCC (класифікація ознак розпізнавання) та визначається найкращий результат порівняння по одному із заданих критеріїв класифікації і видається результат ідентифікації особи функцією прийняття рішень. Класифікація ознак розпізнавання MFCC відбувається на основі сумішей Гаусових розподілів та методу опорних векторів (англ. Gaussian Mixture Model and Support Vector Machine, GMM-SVM) з використанням лінійного ядра Кампбелла та методу головних компонент з проекцією на латентні структури, що у сумі забезпечить підвищення надійності ідентифікації, що проявляється у змен-шенні помилок 1-го та 2-го роду.

Головна проблема, що виникає при розробці такого роду систем, полягає у варіативній вимові одного і того ж слова як різними людьми, так і однією і тією ж людиною в різних ситуаціях. Крім того, на вхідний МС впливають численні фактори, такі як навколишній шум, відлуння і завади в каналі зв'язку. Ускладнюється це і тим, що шум і спотворення заздалегідь невідомі, тобто система не може бути підлаштована під них до початку роботи.

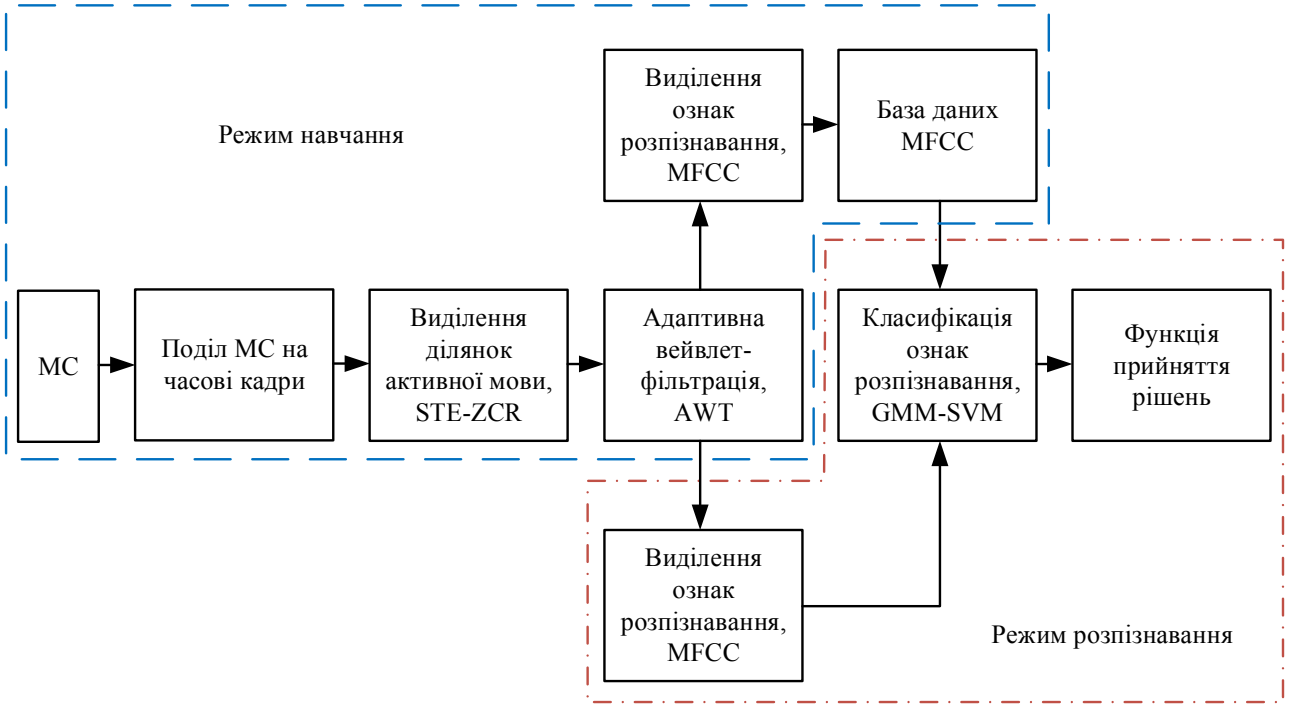


Рис. 1. Структурна схема системи голосової ідентифікації особи на основі MFCC та GMM-SVM

Визначення моментів початку та закінчення фрази за наявності шуму є першочерговою задачею при автоматичній ідентифікації особистості за голосом. Процедура виявлення моментів початку та закінчення фрази істотно зменшує число арифметичних операцій, якщо обробляти лише ті сегменти, в яких є МС. Внаслідок цього швидкість обробки суттєво збільшуватиметься. Найбільш поширеним способом стиснення мовних даних є видалення пауз між фразами, словами, окремими звуками. Як показали численні дослідження, в мові може бути до 50 % пауз, а в діалозі їх обсяг може досягати 70 %. Тому було створено різні алгоритми поділу МС на вокалізовані та невокалізовані ділянки та ділянки мовчання, які усувають надмірність мови, виділяючи лише значущі її параметри. Звуки мови, у яких є основний тон, називаються вокалізованими. При дослідженні динаміки зміни характеристик МС важливим завданням є вибір тривалості часових кадрів, на які МС повинен розбиватися.

Метод визначення активності мови працює у процесі кодування МС перед ідентифікацією особистості за голосом. Наявність пауз визначається на основі аналізу та синтезу мовних даних, що містяться у кадрах МС. Припустимо, що мова містить паузу, яку можна передбачити, тоді наявність паузи в наборі цифрових відліків МС визначається на основі порівняння сумарної

енергії кадру мовних даних з деяким граничним значенням, яке відокремлює паузу від кадру з активною мовою. У цьому випадку поріг необхідно підібрати таким чином, щоб не допустити усунення помилкових пауз, так як це може призвести до погіршення якості та втрати важливих даних МС і як наслідок зниження ефективності алгоритму ідентифікації особистості за голосом. Зазвичай для визначення пауз застосовується складний алгоритм, який враховує як часову енергію так і енергію спектральної складової кадру МС.

Тривалість кадру МС має бути досить малою, щоб послідовність кадрів точніше відображала короткочасну динаміку зміни МС, і досить великий, щоб послідовність кадрів точніше відображала довготривалу динаміку МС. Згідно з умовами вибору тривалості кадру МС, зазначеними в табл. 1, тривалість його кадру повинна бути не меншою за період основного тону $T_{от} = 1/f_{от} = 10$ мс, де $f_{от} \geq 100$ Гц – частота основного тону.

Таблиця 1

Умови вибору тривалості кадру МС

Кількість відліків	Тривалість кадру, мс	Властивості тривалості кадру МС
32	32/8 = 4	Відображає короткочасну динаміку МС і не відображає його періодичний характер
64	64/8 = 8	Відображає короткочасну динаміку МС і не повністю відображає його періодичний характер
128	128/8 = 16	Не повністю відображає короткочасну та довготривалу динаміку МС, повністю відображає його періодичний характер
256	256/8 = 32	Не відображає короткочасну динаміку МС та відображає довготривалу динаміку МС, повністю відображає його періодичний характер

На рис. 2 представлена блок-схема алгоритму поділу МС на вокалізовані та невокалізовані сегменти та сегменти мовчання (паузи). Цей алгоритм ґрунтується на припущенні, що МС – це нестаціонарний процес зі значними змінами короткочасної енергії та числа перетинів нуля між суміжними кадрами (англ. Short-Time Energy and Zero-Crossing Rate, STE-ZCR) [10].

Алгоритм містить 7 блоків.

Блок 1. Вхідний МС $x(m)$, $m = \overline{0, N-1}$.

Блок 2. Розділення МС на кадри тривалістю 16 мс.

Блок 3. Обчислення значень короткочасної енергії E_n (або короткочасне значення модуля енергії) та числа перетинів нуля Z_n n-го кадру. Наприклад, короткочасна енергія дорівнює

$$E_n = \sum_{m=n-N+1}^{n} x^2(m), \text{ або } E_n = \sum_{m=-\infty}^{\infty} \left[x(m)\omega(n-m)\right]^2,$$

$$E_n = \sum_{m=0}^{N-1} x^2(N-n+m), \quad \text{де } n \text{ – номер кадру;}$$

$$\omega(m) = \begin{cases} 1, m = \overline{0, N-1}, \\ 0, m \neq \overline{0, N-1} \end{cases} \text{ – віконна функція кадру;}$$

$n = \overline{0, L}$; L – кількість кадрів; $M = LN$ – число відліків МС.

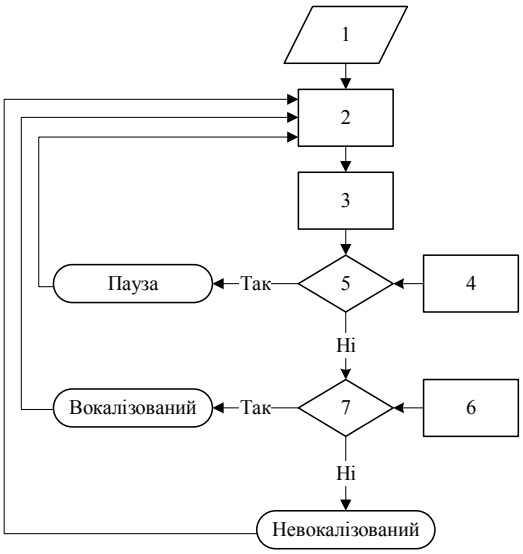


Рис. 2. Блок-схема алгоритму поділу МС на вокалізовані та невокалізовані сегменти та сегменти мовчання (паузи)

Короткочасна функція середньої кількості переходів через нуль, чи нульових перетинів, полягає в порівнянні знаків сусідніх відліків. Наприк-лад

$$z_n = \sum_{m=-\infty}^{\infty} |\operatorname{sng}(x(m)) - \operatorname{sng}(x(m-1))|\, \omega(n-m),$$

де $$W(m) = \begin{cases} \frac{1}{2}, 0 \leq m \leq N-1, \\ 0, \end{cases},$$

а $\operatorname{sgn}(X(m)) = \begin{cases} 1, & X(m) > 0, \\ -1, & X(m) < 0, \end{cases}$ – знакова функція.

Блоки 4, 6. Встановлення порогових значень $E_{пор}$ та $Z_{пор}$ для E_n та Z_n.

Блок 5. Перевірка виконання умови $E_n < E_{пор}$?: так – n-й кадр відноситься до сегменту мовчання; ні – до блоку 7.

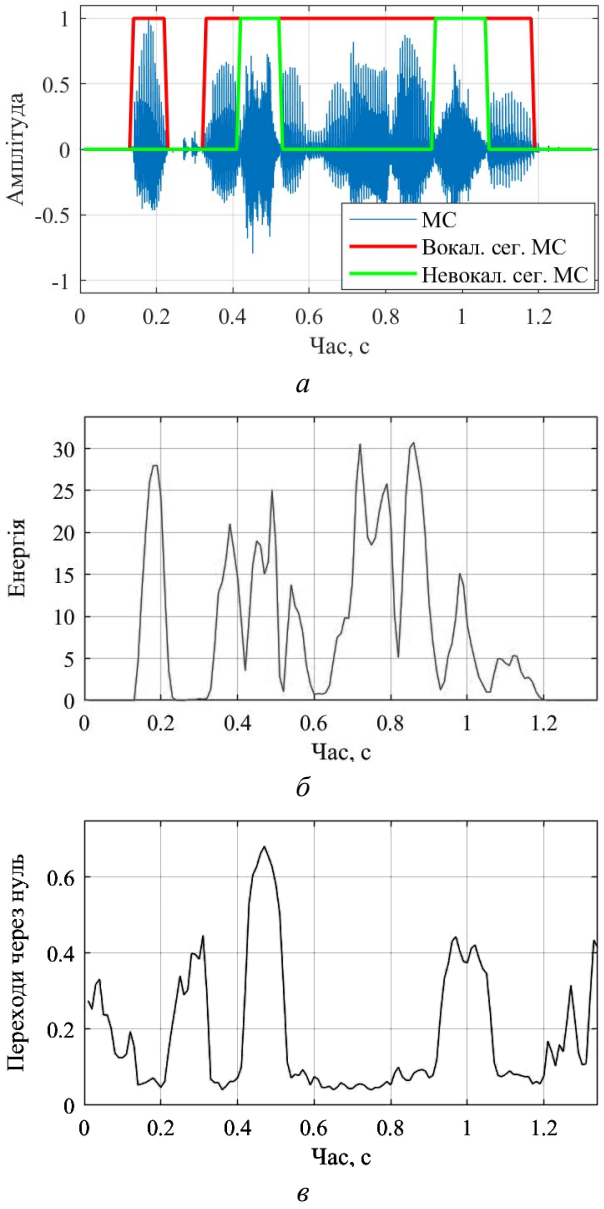
Блок 7. Перевірка виконання умови $Z_n < Z_{\text{пор}}$?: так – n-й кадр відноситься до вокалізованого сегменту; ні – n-й кадр відноситься до невокалізованого сегменту.

Недоліком даного алгоритму є висока чутливість E_n до великих значень сигналу.

Для зменшення помилок прийняття рішення щодо того, чи є ділянка вокалізованою, пропонується використати співвідношення

$$R_{rms} = \frac{E_{rms}}{Z_n},$$

де $E_{rms} = \sqrt{\overline{x^2}(m)} = \sqrt{\frac{1}{N}\sum_{m=1}^{N} x^2(m)}$ – середнє квадратичне значення МС. Отримані дані представлено на рис. 3.



а

б

в

Рис. 3. Графік визначення вокалізованих та невокалізованих сегментів *а*, енергії *б* та переходів через нуль *в* МС

Вокалізована мова характеризується великим значенням E_{rms} та малим Z_n, а невокалізована мова характеризується малим значенням E_{rms} та великим Z_n, тому справедлива умова: R_{rms} є великим для вокалізованого кадру та малим для невокалізованого кадру. В даному випадку вимоги до вибору порогового значення R_{rms} є більш простими, що зменшує можливість помилкового прийняття рішення щодо того, чи кадр вокалізований.

На наступному етапі потрібно виконати адаптивну вейвлет-фільтрацію МС (англ. Adaptive Wavelet Thresholding, AWT), для унеможливлення впливу шуму на ідентифікацію особи за голосом. Власне кажучи, порогова вейвлет-фільтрація МС аналогічна оцінюванню сигналу шляхом його усереднення за допомогою ядра, яке локально адаптоване до гладкості сигналу. Набір сполучених дзеркальних фільтрів у такому разі розкладає сигнал у дискретній області по ортогональному вейвлет-базису $\{\psi_{j,m}\}$ на кілька частотних діапазонів [11].

Шумоочищення МС виконується як повне відсікання коефіцієнтів вейвлет-перетворення виходячи з припущення, що їх значення малої амплітуди і є шум.

Таким чином, у вейвлет-базисі, де коефіцієнти з великою амплітудою відповідають різким змінам МС, така обробка зберігає лише переривчасті складові, що походять від вхідного МС без додавання інших компонентів, обумовлених шумом.

Загалом, прирівнюючи малі коефіцієнти нулю, ми виконуємо адаптивне згладжування, що залежить від гладкості вхідного МС $\acute{s}(t)$. Зберігаючи коефіцієнти великої амплітуди, ми уникаємо згладжування різких перепадів та зберігаємо локальні особливості. Проведення такої процедури на кількох масштабах веде до поступового зменшення впливу шуму як на кусочно-гладких, так і на розривних ділянках МС.

Кожен МС, представлений у дискретному вигляді, має певний відсоток значущих вейвлет-коефіцієнтів $\langle f, \psi_{j,m} \rangle$, який збільшується зі зростанням масштабу вейвлет-розкладання a^j. Цей факт пояснюється тим, що низькочастотна складова МС створює меншу кількість вейвлет-коефіцієнтів великої амплітуди, тоді як кількість вейвлет-коефіцієнтів високочастотної складової МС з малою амплітудою на рівнях збільшується. При великому значенні a^j поріг T необхідно збільшувати, щоб відфільтрувати вейвлет-

коефіцієнти малої амплітуди на всіх рівнях розкладання.

Таким чином, для вирішення задачі адаптивного шумоочищення необхідно провести адаптивну генерацію мікролокальних порогів, що дозволить зменшити вплив адитивного шуму на чисту форму МС, і зберегти значущі вейвлет-коефіцієнти великої амплітуди, які характеризують локальні особливості МС.

Представимо модель МС $\acute{r}(t)$, (рис. 4 (а)), спотвореного адитивним шумом, як

$$X(t) = \acute{r}(t) + \eta(t). \quad (1)$$

Тоді при розкладанні такого сигналу набором сполучених дзеркальних фільтрів по деякому дискретному ортогональному базису $\{\psi_m\}$ дає:

$$WX[m] = \langle X, \psi_m \rangle;$$

$$W\acute{r}[m] = \langle \acute{r}, \psi_m \rangle;$$

$$W\eta[m] = \langle \eta, \psi_m \rangle.$$

Скалярний добуток (1) з ψ_m дає

$$WX[m] = Wf[m] + W\eta[m].$$

Це означає, що модель шуму не залежить від базису розкладання і залишається в ньому такою ж, як у вхідному МС.

Введемо лінійний оператор D, що оцінює $Wf[m]$ по $WX[m]$ за допомогою функції $d_m(x)$. Результуюча оцінка є

$$\widetilde{F} = DX = \sum_{m=0}^{N-1} d_m(WX[m])\psi_m.$$

Коли $d_m(x)$ – порогова функція, ризик цієї оцінки може бути зведений до мінімуму.

Порогова фільтрація виконується за допомогою

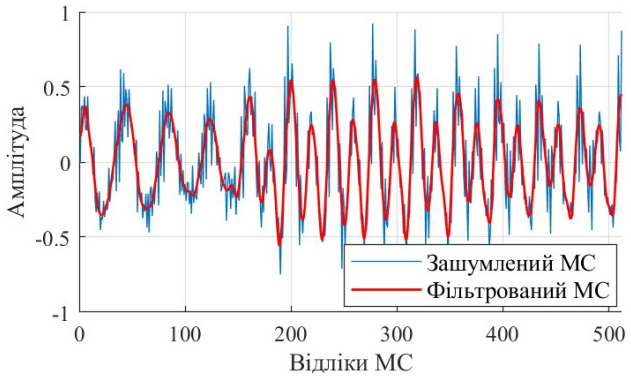
$$d_m(x) = \rho_T(x) = \begin{cases} x, & \text{якщо } |x| > T; \\ 0, & \text{якщо } |x| \le T, \end{cases}$$

та видаляє всі коефіцієнти, амплітуда яких нижче встановленого порогового значення T (рис. 4, *б*).

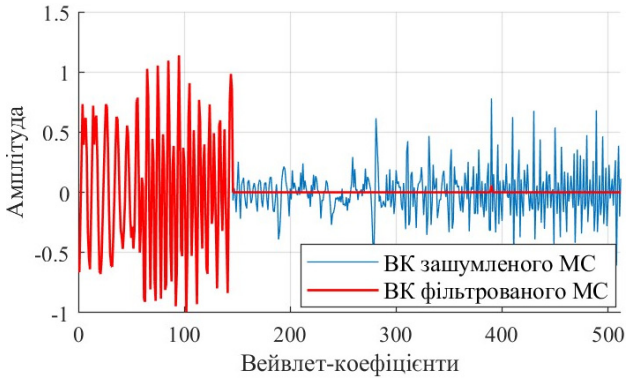
У вейвлет-базисах коефіцієнти великої амплітуди відповідають розривам МС та його різким змінам. Це означає, що оцінка зберігає у розкладанні лише нерегулярні компоненти, які походять від вхідного МС, без додавання паразитних сплесків, породжених шумом.

Поріг повинен вибиратися адаптивно і бути за величиною дещо більшим, ніж максимальний рівень шуму. Тобто значення $|W\eta[m]|$ повинні бути з великою ймовірністю менші за T. Так досягається мінімальний рівень ризику в пороговій вейвлет-фільтрації МС.

Нехай $r_t(x,T)$ ризик порогової оцінки, обчисленої з порогом T. Тоді оцінка $\tilde{r}_t(x,T)$ ризику $r_t(x,T)$ має обчислюватися за МС $X(t)$, спотвореним шумом. Значення порога T у такому разі оптимізується мінімізацією $\tilde{r}_t(x,T)$.



а



б

Рис. 4. Вейвлет-фільтрація МС адаптивною пороговою обробкою: зашумлений та фільтрований МС *(а)*, вейвлет-коефіцієнти (ВК) зашумленого та фільтрованого МС *(б)*

Щоб знайти значення $\tilde{T}$, яке мінімізує оцінку $\tilde{r}_t(x,T)$, N коефіцієнтів даних $WX[m]$ сортують за зменшенням амплітуди. Тоді ранжовані таким чином коефіцієнти вейвлет-розкладання утворюють впорядковану множину $\{WX^r[k]\}_{1\le k\le N}$, де будь-який $WX^r[k] = WX[m_k]$ – відповідний коефіцієнт рангу k : $|WX^r[k]| \ge |WX^r[k+1]|$.

Нехай l – деякий індекс, такий що $|WX^r[l]| \le T < |WX^r[l-1]|$, тоді можемо прийняти

$$\tilde{r}_t(f,T) = \sum_{k=1}^{N} |WX^r[k]|^2 - (N-l)\sigma^2 + l(\sigma^2 + T^2), \quad (2)$$

де σ^2 – дисперсія шумової компоненти.

Тоді для мінімізації $\tilde{r}_t(x,T)$ необхідно вибрати $T = |WX^r[l]|$.

Дисперсію σ^2 шуму $\eta[n]$ можна визначити за даними (1), для чого необхідно придушити вплив $f[n]$, таку грубу оцінку можна провести за середніми значеннями вейвлет-коефіцієнтів найменшого масштабу.

Це твердження обумовлено тим, що на кожному рівні вейвлет-розкладання вхідного МС $X(t)$ довжини N множина значень $\{\langle X,\psi_m\rangle\}_{0\le m\le N/2}$ кінцева і має лише кілька коефіцієнтів великої амплітуди. Тому для більшості ділянок $\langle X,\psi_m\rangle \approx \langle \eta,\psi_m\rangle$.

Тоді, якщо M_X – медіана множини $\{|\langle X,\psi_m\rangle|\}_{0\le m\le N/2}$, то груба оцінка дисперсії σ^2 шуму η оцінюється за M_X нехтуючи впливом $f[n]$:

$$\tilde{\sigma} = \frac{M_X}{0,6745}. \qquad (3)$$

Таким чином, адаптивну процедуру шумоочищення за коефіцієнтами вейвлет-розкладання можна провести за такою схемою:

1) Обчислення оцінки $\tilde{\sigma}^2$ дисперсії шуму σ^2 за формулою медіани (3) за найменшого масштабу розкладання;

2) Обчислення порога T_j для кожного рівня декомпозиції j з мінімізацією ризику (2);

3) Порогова обробка вейвлет-коефіцієнтів розкладання отриманим порогом для кожного рівня масштабу a^j.

В якості інформативних ознак розпізнавання МС при автоматичній ідентифікації особистості за голосом використовуються мел-частотні кепстральні коефіцієнти (англ. Mel-Frequency Cepstral Coefficients, MFCC), які засновані на двох ключових поняттях: кепстр та мел-шкала [12].

Кепстр (англ. Cepstrum) – результат дискретного косинусного перетворення від логарифму амплітудного спектра сигналу, що формально можна представити наступним виразом:

$$c(n) = DCT\left\{\log(|F\{f(t)\}|^2)\right\}.$$

Мел-шкала моделює частотну чутливість людського слуху. Спеціалістами з психоакустики було встановлено, що зміна частоти вдвічі в діапазоні низьких і високих частот людина сприймає по-різному. У частотній смузі до 1000 Гц суб'єктивне сприйняття подвоєння частоти збігається з реальним збільшенням частоти вдвічі, тому до 1000 Гц мел-шкала близька до лінійної. Для частот вище 1000 Гц мел-шкала є логарифмічною (рис. 5).

Перетворення із герц-шкали в мел-шкалу і навпаки відбувається за наступними формулами:

$$\hat{f}_{mel}(f_{hz}) = 1127\ln(1+\frac{f_{hz}}{700}),$$

$$\hat{f}_{hz}(f_{mel}) = 700(e^{f_{mel}/1127}-1).$$

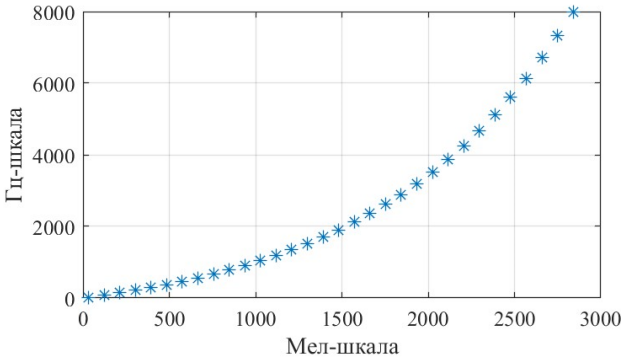


Рис. 5. Співвідношення мел-шкали з герц-шкалою

MFCC – це значення кепстра, розподілені за мел-шкалою з використанням банку фільтрів.

Алгоритм знаходження MFCC:

1) Сигнал $s[t]$, що пройшов попередню обробку, розбивається на K кадрів по N відліків, що перетинаються на 1/2 довжини кадру:

$$s[t] \to S_n[t],\ n = 1,.....,K.$$

2) У кожному кадрі виконується дискретне перетворення Фур'є (ДПФ):

$$\mathrm{Re}\,X_n[k] = \frac{2}{N}\sum_{i=1}^{N} S_n[i]\cos(2\pi k(i-1)/N),$$

$$\mathrm{Im}\,X_n[k] = -\frac{2}{N}\sum_{i=1}^{N} S_n[i]\sin(2\pi k(i-1)/N),$$

де $k = 1,\ldots\ldots,M$, $M = N/2$.

3) Знаходимо спектральну щільність потужності МС:

$$P_n[k] = A_n[k]^2,$$

$$A_n[k] = \sqrt{\mathrm{Re}\,X_n[k]^2 + \mathrm{Im}\,X_n[k]^2}.$$

4) Застосування банку фільтрів (рис. 6):

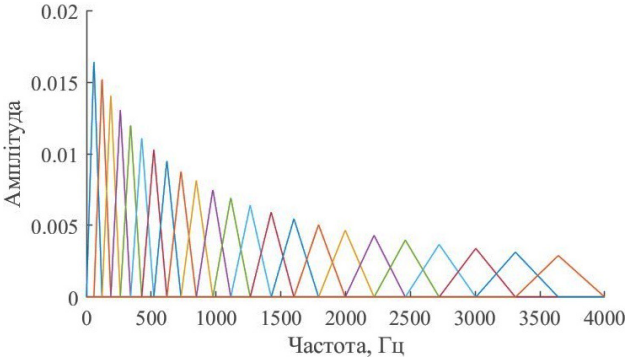


Рис. 6. Банк фільтрів

4.1) задається кількість фільтрів P, а також початкова f_l і кінцева f_h частоти (f_h не повинна перевищувати половини частоти дискретизації);

4.2) задані частоти f_l та f_h переводяться в мел-шкалу:

$$f_l^m = \hat{f}_{mel}(f_l),$$
$$f_h^m = \hat{f}_{mel}(f_h);$$

4.3) згідно мел-шкалі, відрізок $[f_l^m, f_h^m]$ розбивається на $P + 1$ рівних відрізків:

$$[f_l^m, f_{j+1}^m],\ j = 1,....,P+1,$$

які не перетинаються та мають довжину:

$$len = \frac{f_h^m - f_l^m}{P+1};$$

4.4) знаходяться центри заданих відрізків:

$$C^m[i] = f_l^m + i \cdot len,\ i = 1,...,P,$$

і переводяться у герц-шкалу:

$$C[i] = \hat{f}_{hz}(C^m[i]),\ i = 1,....,P,$$

де $C[i]$ – центральні частоти трикутних фільтрів;

4.5) центри трикутних фільтрів переводяться з герц-шкали у номери відліків масиву $P_n[k]$:

$$f_{smp}[i] = \frac{M}{Fs} C[i],\ i = 1,.....,P,$$

де Fs – частота дискретизації МС;

4.6) відліки спектральної щільності потужності МС множаться на сформований банк трикутних фільтрів:

$$X_n[i] = \sum_{k=1}^{M} P_n[k] H_i[k],\ i = 1,....,P,$$

$$H_i[k] = \begin{cases} 0, & k < f_{smp}[i-1] \\ \dfrac{(k - f_{smp}[i-1])}{f_{smp}[i] - f_{smp}[i-1]}, & f_{smp}[i-1] \le k \le f_{smp}[i] \\ \dfrac{(f_{smp}[i+1] - k)}{f_{smp}[i+1] - f_{smp}[i]}, & f_{smp}[i] \le k \le f_{smp}[i+1] \\ 0, & k > f_{smp}[i+1] \end{cases}$$

5) Взяття логарифму від спектральної щільності потужності МС після застосування банку трикутних фільтрів:

$$X_n[i] = \ln(X_n[i]),\ i = 1,.....,P.$$

6) Виконаємо дискретне косинусне перетворення до логарифмічної енергії спектру МС:

$$C_n[j] = \sum_{k=1}^{P} X_n[k] \cos\left(j\left(k - \frac{1}{2}\right)\frac{\pi}{P}\right),$$
$$i = 1,.....,P,\ j = 1,.....,J,$$

де $C_n[j]$ – масив MFCC, J – бажана кількість коефіцієнтів ($J < P$).

Отриману матрицю кепстральних коефіцієнтів можна для наочності представити у вигляді бітової карти. На рис. 7. наведено приклад такої карти. Шкала праворуч показує відповідність між відтінками кольорів (потужність коефіцієнтів) та значеннями кепстральних коефіцієнтів.

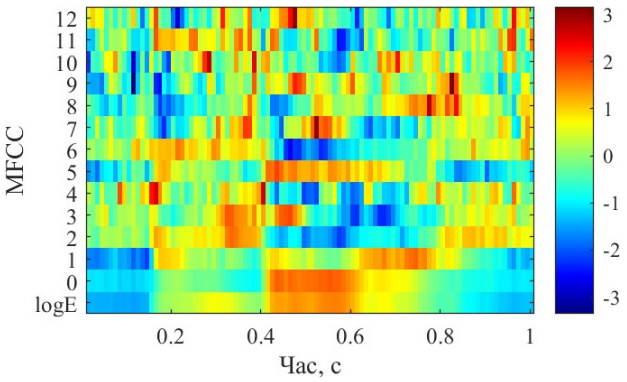


Рис. 7. Бітова карта MFCC для слова «вправо»

На наступному етапі потрібно провести класифікацію ознак МС на основі сумішей Гаусових розподілів (англ. Gaussian Mixture Model, GMM) та методу опорних векторів (англ. Support Vector Machine, SVM) [13].

GMM диктора забезпечує ймовірнісну модель основних звуків мови. Лінійна комбінація Гаусових функцій є базис, який здатний до представлення великого класу експериментальних розподілів. Перевагою GMM є здатність формувати гладкі апроксимації експериментальних розподілів компонентів акустичного простору, які мають довільну форму.

Для вхідного вектора $\vec{x}$ щільність Гаусової суміші – є зважена сума M компонент суміші, та вона задається виразом:

$$p_i\left(\vec{x} \mid \lambda\right) = \sum_{i=1}^{M} \alpha_i p_i\left(\vec{x}\right),$$

де $\vec{x}$ – N-вимірний випадковий вектор, $p_i\left(\vec{x}\right)$, $i = 1,...,M$, є компонентами суміші і α_i, $\{i = 1,...,M\}$ є вагами суміші, а кожна компонента щільності – це функція Гауса.

Ваги компонентів суміші задовольняють зв'язку:

$$\sum_{i=1}^{M} \alpha_i = 1.$$

GMM параметризується набором параметрів визначених для кожної i-ої компоненти суміші: середніми векторами $\vec{\mu}_i$, матрицями коваріації

$\sum_i$ та вагами α_i. Ці параметри всі разом представлені системою позначень:

$$\lambda = \{\alpha_i, \vec{\mu}_i, \Sigma_i\},\ i = 1,....,M\ .$$

Мета навчання моделі GMM полягає в тому, щоб отримати параметри GMM λ, які дають кращу відповідність експериментальному розподілу навчальних векторів $X = \{\vec{x}_1,...,\vec{x}_T\}$. Відомо кілька методів для знаходження оптимальних параметрів GMM. Безумовно, найпоширеніший метод, який використовується для такої оцінки – це метод максимальної правдоподібності (англ. Maximum Likelihood, ML).

Мета ML-оцінки полягає в тому, щоб знайти модельні параметри, які максимізують ймовірність GMM на навчальній вибірці даних МС. Важливий крок у застосуванні ML-методу полягає у виборі фактичної функції правдоподібності. Для навчальної послідовності з T статистично незалежних векторів $X = \{\vec{x}_1,...,\vec{x}_T\}$, ця функція $p(X \mid \lambda)$ повинна мати вигляд:

$$p(X \mid \lambda) = \prod_{t=1}^{T} p_i(\vec{x}_t \mid \lambda)\ .$$

Дуже перспективним методом для верифікації диктора є метод опорних векторів (англ. Support Vector Machine, SVM). Метод SVM є дискримінантним методом на відміну від породжувального методу GMM.

Сучасний розвиток методу SVM в задачах розпізнавання диктора виявило, що саме ефективне його застосування – це використання SVM у комбінації з методом GMM. У разі такої гібридної системи GMM-SVM, SVM діє не в просторі вхідних акустичних векторів МС, а в модельному просторі супервекторів середніх μ_i GMM.

Розглянемо задачу класифікації на два класи, що не перетинаються, в якій об'єкти описуються n-вимірними дійсними векторами: $X \in R^N$, $Y \in \{-1,+1\}$. Тоді визначимо лінійний пороговий класифікатор:

$$Y(x) = sign\left(\sum_{j=1}^{n} w_j x^j - w_0\right) = sign\left(\langle w, x\rangle - w_0\right),$$

де $x = (x_1,\ldots,x_n)$ ознакове описання об'єкта X; вектор $w = (w_1,...,w_n) \in R^n$ скалярний поріг $w_0 \in R^n$ є параметрами алгоритму. Рівняння $Y(x)$ описує гіперплощину, яка розділяє класи у просторі R^n

$$\langle w, x\rangle = w_0\ .$$

У разі лінійної нероздільності класів виконують перехід від вхідного простору ознакових описів об'єктів X до нового простору H за допомогою деякого перетворення $\psi: X \to H$. Якщо простір H має досить високу розмірність, то можна сподіватися, що в ньому вибірка виявиться лінійно розділеною. Простір H називається спрямовуючим. Якщо припустити, що ознаковими описами об'єктів є вектори $\psi(x_i)$, а не вектори x_i, то побудова SVM проводиться практично так само, як і раніше. Єдина відмінність полягає в тому, що скалярний добуток $< x, x' >$ в просторі X замінюється на скалярний добуток $< \psi(x), \psi(x') >$ у просторі H.

Це означає, що при побудові SVM скалярний добуток $< x, x' >$ можна формально замінити ядром $K(x, x')$. Оскільки ядро у загальному випадку нелінійне, така заміна призводить до суттєвого розширення класу допустимих алгоритмів $a: X \to Y$.

Так, для GMM-SVM систем класифікації МС часто використовується лінійне ядро Кампбелла:

$$K_{lin}(s^a, s^b) = \sum_{i=1}^{N}\left(\sqrt{w_i}\sum_i^{-\frac{1}{2}}\mu_i^a\right)\left(\sqrt{w_i}\sum_i^{-\frac{1}{2}}\mu_i^b\right)^t\ .$$

Результати дослідження впливу завад каналу зв'язку на голосову ідентифікацію особи

Ідентифікація особи за голосом, що проводиться в реальних умовах, зустрічається з низкою серйозних труднощів.

По-перше, можливі спотворення, пов'язані безпосередньо з диктором та зумовлені особливостями його психофізичного стану, захворюванням тощо. Ці спотворення за допомогою будь-якої автоматизованої системи обробки та класифікації виключити неможливо, можна лише зменшити їхній вплив.

По-друге, виникають апаратні спотворення на різних ділянках проходження МС при його запису, обробці та зберіганні.

По-третє, на МС неминуче накладаються зовнішні механічні шуми, які можуть суттєво спотворювати його. Найважливішим завданням систем голосової ідентифікації є зменшення негативного впливу другого та третього чинників.

Нарешті, у низці завдань доводиться проводити ідентифікацію у вельми складних умовах накладення голосів кількох дикторів, зокрема з близькими акустичними характеристиками. Зазначимо, що дослідження можливостей голосової ідентифікації для цього найскладнішого випадку практично не проводилися.

Зазвичай виділяють: спотворення МС, пов'язані з самим диктором, з шумом навколишнього середовища, зі спотворенням мікрофонної системи (у тому числі електромагнітні завади), спотворення, що виникають у каналі запису та у каналі зв'язку при передачі МС, а також, спотворення при обробці МС спеціальним програмним забезпеченням (рис. 8).

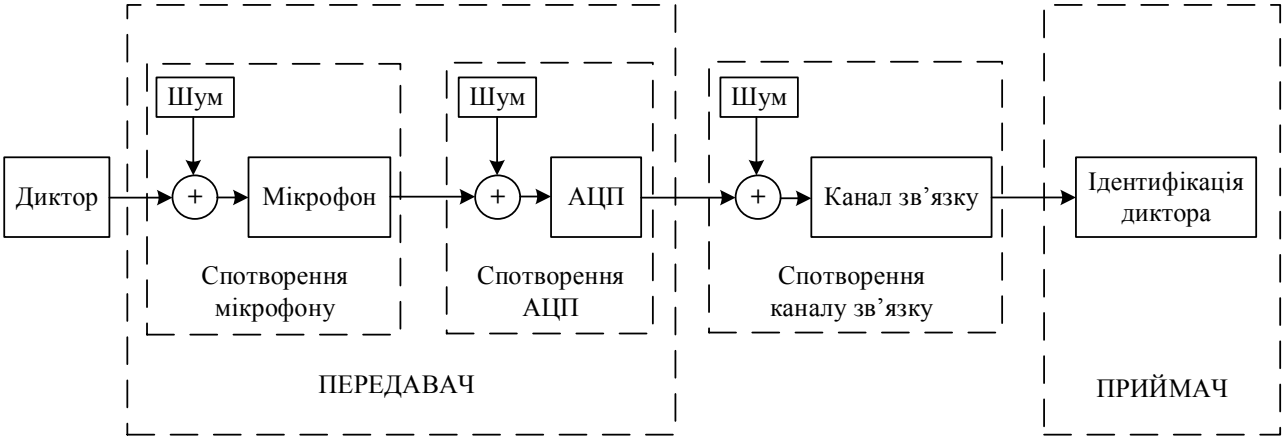


Рис. 8. Вплив шуму і завад на різні ділянки проходження МС

Далі будуть розглянуті спотворення, зумовлені зміною амплітудно-частотного спектра вхідного МС.

Завади, що виникають у апаратній частині системи ідентифікації, в кінцевому підсумку зводяться до частотних та амплітудних спотворень вхідного МС та його спектру. Це може бути викликано недоліком мікрофонних пристроїв, які мають нелінійні амплітудно-частотні характеристики, використанням різних фільтрів при записі МС, а також спотвореннями під час аналого-цифрового перетворення.

Для аналізу впливу спотворень МС на ідентифікацію диктора застосовано нижченаведений підхід.

Для математичного моделювання спотворень МС було застосовано алгоритм передискретизації, який базується на використанні дискретного перетворення Фур'є і дозволяє підвищувати частоту дискретизації сигналу у задане ціле або дробове число разів. Моделювання конкретного спотворення здійснювалося наступним чином [14].

Нехай вхідний МС характеризувався скінченною кількістю відліків $a(n)$. На першому кроці алгоритму проводилось обчислення коефіцієнтів $A(k)$ прямого перетворення Фур'є:

$$A(k)=\sum_{n=1}^{N} a(n)\cdot e^{-j2\pi\frac{k}{N}n},\ k=1,2,...,N,$$

На другому кроці в область біля відліку з номером $N/2$ спектру вставлялися нульові компоненти, кількість яких задавалась значеннями початкової кількості відліків N та кількості відліків у передискретизованому сигналі M.

Коефіцієнти $H(i)$ передискретизованого спектра у випадку непарних чисел N визначалися формулами:

$$\begin{cases} H(i)=A(i), 1\le i\le \dfrac{N+1}{2}, \\ H(i)=0, \dfrac{N+1}{2}+1\le i\le \dfrac{N+1}{2}+M-N, \\ H(i)=A(i-M+N), \dfrac{N+1}{2}+M-N\le i\le M, \end{cases}$$

у разі парних N – формулами:

$$\begin{cases} H(i)=A(i), 1\le i\le \dfrac{N}{2}, \\ H(i)=\dfrac{A(N/2+1)}{2}, i=\dfrac{N}{2}+1, \\ H(i)=0, \dfrac{N}{2}+2\le i\le \dfrac{N}{2}+M-N, \\ H(i)=\dfrac{A(N/2+1)}{2}, i=\dfrac{N}{2}+M-N+1, \\ H(i)=A(i-M+N), \dfrac{N}{2}+M-N+2\le i\le M. \end{cases}$$

На завершальному кроці алгоритму обчислювалися відліки $h(m)$ оберненого дискретного перетворення Фур'є з нормуванням:

$$h(m)=\frac{1}{M}\sum_{k=1}^{M} H(k)\cdot e^{-j2\pi\frac{k}{M}m},\ m=1,2,....,M\ .$$

На цьому формування спотвореного МС закінчувалося.

В якості величини, яка кількісно характеризує спотворення, використовується коефіцієнт нелінійних спотворень K, який вводиться як відношення середньоквадратичної суми спектральних

компонентів вихідного МС, які відсутні у спектрі вхідного МС, до середньоквадратичної суми спектральних компонентів вхідного МС:

$$K = \frac{\sqrt{\frac{1}{L}\sum_{l=1}^{L} H^2(l)}}{\sqrt{\frac{1}{N}\sum_{k=1}^{N} A^2(k)}},$$

де $H(l)$ – спектральні компоненти вихідного МС, які відсутні в спектрі вхідного МС $A(k)$, L – кількість спектральних компонентів $H(l)$.

Спотворення генерувалися таким чином, що частота дискретизації:

$$F = \frac{M}{t},$$

де t – тривалість вхідного МС, збільшувалася за рахунок зміни кількості відліків від N до M, при цьому для кожного спотвореного МС розраховувався коефіцієнт нелінійних спотворень K.

На рис. 9 для ілюстрації наведено сегмент частотного спектру вхідного МС та відповідний частотний спектр спотвореного МС ($K = 0,4$) для інтервалу частот від $f = 0$ Гц до $f = 4000$ Гц, для якого спотворення виявилися найбільш помітними (тут A – амплітуда звукових коливань).

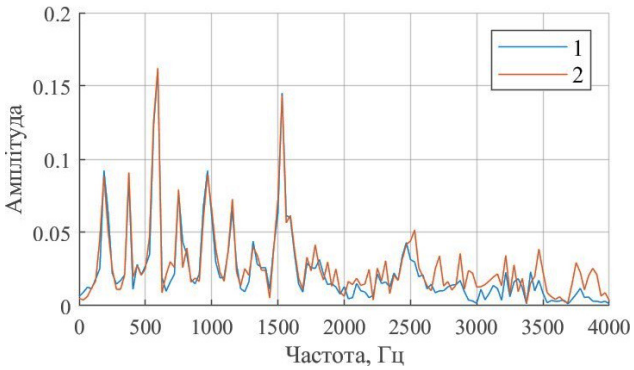


Рис. 9. Сегменти частотних спектрів МС: 1 – спектр вхідного МС; 2 – спектр спотвореного МС

Далі вхідний та спотворений МС піддавалися ідентифікації за допомогою підходу, який був описаний в попередньому підрозділі.

Отримані при розрахунках дані подавалися у вигляді графіків класифікації у просторі перших головних компонентів (ГК), що дозволяють наочно інтерпретувати результат голосової ідентифікації дикторів.

Для проведення експерименту в якості вхідних МС використовувалися записи голосу диктора, що являли собою фразу «один два три чотири п'ять шість сім вісім дев'ять», з частотою дискретизації 8 кГц і розрядністю 8 біт. Спотворення у вхідний МС вносилися описаним вище алгоритмом і полягали у додаванні до вузького сегменту спектра відліків, які мали нульову амплітуду.

В якості графічної ілюстрації на рис. 10 наведено сегменти передискретизованих спектрів, які зазнавали спотворень, що відповідали коефіцієнтам $K = 0,1;\ 0,2;\ 0,3;\ 0,4$.

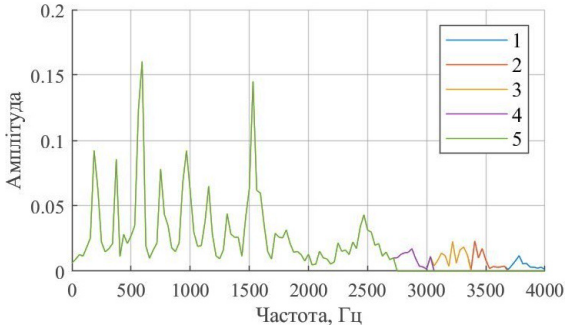


Рис. 10. Сегменти передискретизованих спектрів МС: 1 – $K = 0$; 2 – $K = 0,1$; 3 – $K = 0,2$; 4 – $K = 0,3$; 5 – $K = 0,4$

Розрахований для вхідного та спотвореного МС графік показників ідентифікації представлений на рис. 11. Тут окрема точка, як і раніше, відповідає одному голосовому запису, вхідні МС (дані одного диктора) зображені не залитими червоними точками, спотворені МС з різними значеннями коефіцієнта спотворення K – зеленими залитими точками (точка 1 – $K = 0,1$; 2 – $K = 0,2$; 3 – $K = 0,3$; 4 – $K = 0,4$).

Всі вхідні (неспотворені) МС представлені компактною областю, виділеною еліпсом, який будувався за максимальним розкидом відповідних точок. Попадання точок, що описували спотворені МС у виділену область означає правильну ідентифікацію диктора. Випадки, що відповідали точкам за межами виділеної області, означали, що диктор сприймався як «чужий», тобто спотворення МС були настільки значними, що ідентифікація не досягалася.

З рис. 11 видно, що при коефіцієнтах спотворення $K = 0,1$ і $K = 0,2$ (точки 1 і 2) диктор, незважаючи на спотворення МС, ідентифікувався правильно. При коефіцієнтах $K = 0,3$ і $K = 0,4$ (точки 3 і 4, розташовані за межами еліпсу) ідентифікація вже не досягалась.

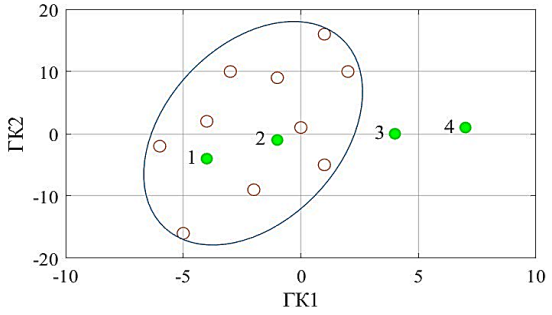


Рис. 11. Вплив спотворень на голосову ідентифікацію. Червоні не залиті точки – неспотворені дані; зелені залиті точки – дані зі спотвореннями: 1 – $K = 0,1$; 2 – $K = 0,2$; 3 – $K = 0,3$; 4 – $K = 0,4$

Таким чином, проведене математичне моделювання спотворень МС дало можливість провести кількісну оцінку величини цих спотворень, за яких можлива правильна ідентифікація особи. Це показує, що запропонований у цьому підрозділі підхід до оцінки впливів спотворень може використовуватися для аналізу надійності методів голосової ідентифікації.

Висновки

У роботі проведено аналіз ефективності системи голосової ідентифікації особи на основі MFCC і GMM-SVM в умовах впливу завад у каналі зв'язку інформаційно-телекомунікаційних мереж. Система дає змогу характеризувати індивідуальні біометричні особливості дикторів із подальшою їхньою класифікацією та ухваленням достовірного рішення щодо допуску або заборони у автентифікації особи, що перевіряється.

Запропоновану систему голосової ідентифікації особи реалізовано за допомогою таких технологій: 1) виділення ділянок активної мови зі знаходженням значень зміни короткочасної енергії та кількості перетинів нуля між суміжними кадрами МС; 2) адаптивна вейвлет-фільтрації МС для вирішення задачі шумоочищення, де необхідно провести адаптивну генерацію мікролокальних порогів, що дасть змогу зменшити вплив адитивного шуму на чисту форму МС; 3) виділення ознак розпізнавання, де як інформативні ознаки розпізнавання МС під час автоматичної ідентифікації особи за голосом використовують мел-частотні кепстральні коефіцієнти, які засновані на двох ключових поняттях – кепстр та мел-шкала; 4) класифікації ознак розпізнавання на основі сумішей Гауссових розподілів та методу опорних векторів з використанням лінійного ядра Кампбелла та методу головних компонент з проекцією на латентні структури, що у сумі забезпечить підвищення надійності ідентифікації, що проявляється у зменшенні помилок 1-го та 2-го роду.

Використання метрики Махаланобіса для побудови багатовимірних еліпсоїдів, що характеризують МС окремих дикторів, забезпечує високу наочність результатів ідентифікації особи та дає змогу розділяти голоси дикторів із близькими фізичними характеристиками. Отримані під час розрахунків дані подавалися у вигляді графіків класифікації в просторі перших голов-них компонентів, що дало змогу наочно інтерпретувати результат голосової ідентифікації дикторів.

Досліджено вплив виду та величини зовнішнього шуму, різноманітних завад та спотворень на ідентифіковані МС, що передаються каналами зв'язку інформаційно-телекомунікаційних мереж. Запропоновано методику, що дає змогу проводити класифікацію МС при накладенні шуму шляхом математичного моделювання спотворень МС через застосування алгоритму предискретизації, що ґрунтується на використанні дискретного перетворення Фур'є та дає змогу підвищувати частоту дискретизації МС у задане ціле чи дробове число разів, де як величину, яка кількісно характеризує спотворення, використовують коефіцієнт нелінійних спотворень, який вводиться як відношення середньоквадратичної суми спектральних компонентів вихідного МС, які відсутні в спектрі вхідного МС, до середньоквадратичної суми спектральних компонентів вхідного МС. Математичне моделювання спотворень МС дало змогу провести кількісну оцінку величини цих спотворень, за яких можлива правильна ідентифікація особи. Це показує, що запропонований підхід до оцінки впливів спотворень може використовуватися для аналізу надійності методів голосової ідентифікації.

Таким чином, проведене з єдиних позицій систематичне дослідження дало змогу виявити вплив зовнішнього шуму на голосову ідентифікацію, що може бути використане під час розроблення та тестування систем дистанційної голосової ідентифікації особи в інформаційно-телекомунікаційних мережах.

ЛІТЕРАТУРА

[1] S. Kinkiri and S. Keates, "Speaker Identification: Variations of a Human voice," *2020 International Conference on Advances in Computing and Communication Engineering (ICACCE)*, Las Vegas, NV, USA, 2020, pp. 1–4, doi: 10.1109/ICACCE49060.2020.9154998.

[2] M. Saleh and I. Jouny, "Multimodal Person Identification through the Fusion of Face and Voice Biometrics," *2022 17th Annual System of Systems Engineering Conference (SOSE)*, Rochester, NY, USA, 2022, pp. 164–169, doi: 10.1109/SOSE55472.2022.9812670.

[3] J. Gomes, H. Fernandes, S. Abraham and S. Chavan, "Person identification based on voice recognition," *2021 4th Biennial International Conference on Nascent Technologies in Engineering (ICNTE)*, NaviMumbai, India, 2021, pp. 1–5, doi: 10.1109/ICNTE51185.2021.9487756.

[4] O. Tymchenko, B. Havrysh, O. O. Tymchenko, O. Khamula, B. Kovalskyi and K. Havrysh, "Person Voice Recognition Methods," *2020 IEEE Third International Conference on Data Stream Mining & Processing (DSMP)*, Lviv, Ukraine, 2020, pp. 287–290, doi: 10.1109/DSMP47368.2020.9204023.

[5] V. UmaRani, M. P, S. M and S. Nischitha, "A Hybrid Mel Frequency Cepstral Coefficients and Bayesian Gaussian Mixure Model for Voice

based Authentication Websites," *2023 International Conference on Device Intelligence, Computing and Communication Technologies, (DICCT)*, Dehradun, India, 2023, pp. 367–370, doi: 10.1109/DICCT56244.2023.10110176.

[6] Q. Chen, J. Li and Y. Li, "Forensic identification for electronic disguised voice based on supervector and statistical analysis," *2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA)*, Bali, Indonesia, 2016, pp. 147–150, doi: 10.1109/ICSDA.2016.7919001.

[7] M. Nalini, R. Gayathiri, A. V, A. L. G and H. D, "Automatic Optimized Voice Based Gender Identification for Speech Recognition," *2022 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS)*, Chennai, India, 2022, pp. 1–4, doi: 10.1109/ICPECTS56089.2022.10047573.

[8] M. Aliaskar, T. Mazakov, A. Mazakova, S. Jomartova and T. Shormanov, "Human voice identification based on the detection of fundamental harmonics," *2022 IEEE 7th International Energy Conference (ENERGYCON)*, Riga, Latvia, 2022, pp. 1–4, doi: 10.1109/ENERGYCON53164.2022.9830471.

[9] B. A. Alsaify, H. S. Abu Arja, B. Y. Maayah, M. M. Al-Taweel, R. Alazrai and M. I. Daoud, "Voice-Based Human Identification using Machine Learning," *2022 13th International Conference on Information and Communication Systems (ICICS)*, Irbid, Jordan, 2022, pp. 205–208, doi: 10.1109/ICICS55353.2022.9811154.

[10] O. Lavrynenko, G. Konakhovych and D. Bakhtiiarov, "Method of voice control functions of the UAV," *2016 4th International Conference on Methods and Systems of Navigation and Motion Control (MSNMC)*, 2016, pp. 47–50, doi: 10.1109/MSNMC.2016.7783103.

[11] O. Veselska, O. Lavrynenko, R. Odarchenko, M. Zaliskyi, D. Bakhtiiarov, M. Karpinski and S. Rajba, "A Wavelet-based steganographic method for text hiding in an audio signal," *Sensors*, vol. 22, no. 15, pp. 1–25, doi: 10.3390/s22155832.

[12] R. Odarchenko, O. Lavrynenko, D. Bakhtiiarov, S. Dorozhynskyi and V. A. O. Zharova, "Empirical Wavelet Transform in Speech Signal Compression Problems," *2021 IEEE 8th International Conference on Problems of Infocommunications, Science and Technology (PIC S&T)*, 2021, pp. 599–602, doi: 10.1109/PICST54195.2021.9772156.

[13] O. Lavrynenko, R. Odarchenko, G. Konakhovych, A. Taranenko, D. Bakhtiiarov and T. Dyka, "Method of Semantic Coding of Speech Signals based on Empirical Wavelet Transform," *2021 IEEE 4th International Conference on Advanced Information and Communication Technologies (AICT)*, 2021, pp. 18–22, doi: 10.1109/AICT52120.2021.9628985.

[14] O. Lavrynenko, A. Taranenko, I. Machalin, Y. Gabrousenko, I. Terentyeva and D. Bakhtiiarov, "Protected Voice Control System of UAV," *2019 IEEE 5th International Conference Actual Problems of Unmanned Aerial Vehicles Developments (APUAVD)*, 2019, pp. 295–298, doi: 10.1109/APUAVD47061.2019.8943926.

Лавриненко О. Ю., Бахтіяров Д. І., Конахович Г. Ф., Курушкін В. Є.

АНАЛІЗ ЕФЕКТИВНОСТІ СИСТЕМИ ГОЛОСОВОЇ ІДЕНТИФІКАЦІЇ НА ОСНОВІ MFCC ТА GMM-SVM ЗА УМОВ ВПЛИВУ ЗАВАД У КАНАЛІ ЗВ'ЯЗКУ

У статті розглядається проблематика голосової ідентифікації особи за умов впливу завад у каналі зв'язку інформаційно-телекомунікаційних мереж. При такій ідентифікації виникають всілякі апаратні спотворення і завади, зумовлені особливостями апаратури і пристроїв для запису, обробки і зберігання інформації, а також слід зауважити, що на мовний сигнал неминуче накладаються зовнішні акустичні шуми, які можуть істотно спотворювати індивідуальні інформативні характеристики. З огляду на це системи ідентифікації, що демонстрували досить високу ефективність у лабораторних умовах, під час аналізу мовної інформації із зовнішніми шумами можуть показати надійність, значно нижчу. Нарешті, у низці завдань доводиться проводити ідентифікацію у вельми складних умовах накладення голосів кількох дикторів, зокрема з близькими акустичними характеристиками. Зазначимо, що дослідження можливостей голосової ідентифікації для цього найскладнішого випадку практично не проводилися. Зважаючи на це, головне завдання дослідження полягає в аналізі ефективності системи голосової ідентифікації на основі MFCC та GMM-SVM за умов впливу завад у каналі зв'язку інформаційно-телекомунікаційних мереж, що дасть змогу кількісно оцінити порогові значення потужності шуму при впливі яких ідентифікація особи буде вірною, а при яких хибною. Запропоновану систему голосової ідентифікації особи реалізовано за допомогою таких технологій: 1) виділення ділянок активної мови зі знаходженням значень зміни короткочасної енергії та кількості перетинів нуля між суміжними кадрами мовного сигналу; 2) адаптивна вейвлет-фільтрації мовного сигналу для вирішення задачі шумоочищення, де необхідно провести адаптивну генерацію мікролокальних порогів, що дасть змогу зменшити вплив адитивного шуму на чисту форму мовного сигналу; 3) виділення ознак розпізнавання, де як інформативні ознаки розпізнавання мовних сигналів під час автоматичної ідентифікації особи за голосом використовують

мел-частотні кепстральні коефіцієнти, які засновані на двох ключових поняттях – кепстр та мел-шкала; 4) класифікації ознак розпізнавання мовних сигналів на основі сумішей Гауссових розподілів та методу опорних векторів з використанням лінійного ядра Кампбелла та методу головних компонент з проекцією на латентні структури, що у сумі забезпечить підвищення надійності ідентифікації, що проявляється у зменшенні помилок 1-го та 2-го роду. Запропоновано методику, що дає змогу проводити класифікацію мовних сигналів при накладенні шуму шляхом математичного моделювання спотворень через застосування алгоритму предискретизації, що ґрунтується на використанні дискретного перетворення Фур'є та дає змогу підвищувати частоту дискретизації у задане ціле чи дробове число разів, де як величину, яка кількісно характеризує спотворення, використовують коефіцієнт нелінійних спотворень, який вводиться як відношення середньоквадратичної суми спектральних компонентів вихідного мовного сигналу до середньоквадратичної суми спектральних компонентів вхідного мовного сигналу. Математичне моделювання спотворень мовних сигналів дало змогу провести кількісну оцінку величини цих спотворень, за яких можлива правильна ідентифікація особи. Це показує, що запропонований підхід до оцінки впливів спотворень може використовуватися для аналізу надійності методів голосової ідентифікації.

Ключові слова: мовний сигнал; голосова ідентифікація; короткочасна енергія; кількість перетинів нуля; адаптивна вейвлет-фільтрація; мел-частотні кепстральні коефіцієнти; суміши Гаусових розподілів; метод опорних векторів.

Lavrynenko O., Bakhtiiarov D., Konakhovych G., Kurushkin V.

ANALYSIS OF THE EFFICIENCY OF THE VOICE IDENTIFICATION SYSTEM BASED ON MFCC AND GMM-SVM UNDER THE INFLUENCE OF INTERFERENCE IN THE COMMUNICATION CHANNEL

The article deals with the issue of voice identification of a person under the influence of interference in the communication channel of information and telecommunication networks. Such identification is subject to all kinds of hardware distortions and interference due to the peculiarities of equipment and devices for recording, processing and storing information, and it should also be noted that external acoustic noise inevitably superimposes on the speech signal, which can significantly distort individual informative characteristics. For this reason, identification systems that have demonstrated fairly high efficiency in laboratory conditions may show much lower reliability when analyzing speech information with external noise. Finally, in a number of tasks, identification has to be performed in very difficult conditions of overlapping voices of several speakers, in particular, with similar acoustic characteristics. It should be noted that there has been virtually no research on voice identification capabilities for this most difficult case. In view of this, the main objective of the study is to analyze the effectiveness of a voice identification system based on MFCC and GMM-SVM under the influence of interference in the communication channel of information and telecommunication networks, which will allow us to quantify the threshold values of noise power under the influence of which the identification of a person will be correct and at which it will be false. The proposed voice identification system is implemented using the following technologies: 1) selection of active speech areas with finding the values of the change in short-term energy and the number of zero crossings between adjacent frames of the speech signal; 2) adaptive wavelet filtering of the speech signal to solve the problem of noise removal, where it is necessary to conduct adaptive generation of micro-local thresholds, which will reduce the effect of additive noise on the pure form of the speech signal; 3) identification of recognition features, where mel-frequency cepstral coefficients based on two key concepts - cepstrum and mel-scale are used as informative features of speech signal recognition in automatic voice identification; 4) classification of speech signal recognition features based on mixtures of Gaussian distributions and the support vector method using the linear Campbell kernel and the principal component method with a projection on latent structures, which in total will increase the reliability of identification, which is manifested in the reduction of errors of the 1st and 2nd kind. A methodology is proposed that allows classifying speech signals with noise by mathematical modeling of distortions through the application of a resampling algorithm based on the use of a discrete Fourier transform and allowing to increase the sampling rate by a given integer or fractional number of times, where the nonlinear distortion coefficient is used as a value that quantitatively characterizes the distortion, which is introduced as the ratio of the root mean square sum of the spectral components of the output speech signal to the root mean square sum of the spectral components of the input speech signal. Mathematical modeling of speech signal distortions made it possible to quantify the magnitude of these distortions, which can be used for correct identification of a person. This shows that the proposed approach to assessing the effects of distortion can be used to analyze the reliability of voice identification methods.

Keywords: speech signal; voice identification; short-time energy; zero-crossing rate; adaptive wavelet thresholding; mel-frequency cepstral coefficients; Gaussian mixture model; support vector machine.

Стаття надійшла до редакції 13.08.2023 р.
Прийнято до друку 11.10.2023 р.