УДК 004-26
DOI 10.17721/ISTS.2023.1.51-60

Сергій Зибін, orcid.org/0000-0002-2670-2823,
zysv@ukr.net
Яна Бєлозьорова, orcid.org/0000-0002-0688-3436,
bryuhanova.ya@gmail.com
Національний авіаційний університет, Київ, Україна

МЕТОД ВИЗНАЧЕННЯ ФОРМАНТНИХ ЧАСТОТ ІЗ ВИКОРИСТАННЯМ СПЕКТРАЛЬНОГО РОЗКЛАДАННЯ МОВНОГО СИГНАЛУ

Форманти є одним з основних компонентів систем ідентифікації мовця, а точність визначення формант – це основа ефективності систем ідентифікації мовця. Поліпшення існуючих систем розпізнавання мови дозволить істотно спростити взаємодію людини з комп'ютером у тому випадку, коли використання класичних інтерфейсів неможливо, а також зробити подібну роботу комфортнішою та ефективною.

Необхідність досліджень із цієї тематики пояснюється незадовільними результатами наявних систем при низькому співвідношенні сигнал/шум, залежністю результату від людини, а також невисокою швидкістю роботи подібного виду систем.

Для порівняння із запропонованим методом використовували такі чотири основні формант-трекери: PRAAT, SNACK, ASSP та DEEP. Існує багато досліджень, що стосуються порівняння формант-трекерів, однак серед них не можна виокремити такий, що має найкращу ефективність.

Виокремлення формант супроводжує цілий ряд проблем, пов'язаних з їхньою динамічної зміною у процесі мовлення. Складність також викликають проблеми, пов'язані з близьким розташуванням піків під час аналізу спектрограм і проблеми правильного визначення піків максимумів формант на спектрограмі. Розташування формант на спектрограмах мовного сигналу достатньо легко визначає людина, але автоматизація цього процесу викликає деякі труднощі.

Виокремлення формантних частот запропоновано виконувати у декілька етапів. Результатом проведеного огляду підходів до визначення формантних частот став алгоритм, що складається з дев'ятьох таких етапів. Сегментація мовного сигналу на вокалізовані фрагменти та паузи виконується методом оцінювання змін фрактальної розмірності. Отримання спектра мовного сигналу виконувалось із використанням комплексного вейвлету Морле на основі віконної функції Гаусса. Для дослідження розглядалися формант-трекери PRAAT, SNACK, ASSP і DEEP. Налаштування кожного з них здійснювали на основі набору параметрів за замовчуванням, що закладено розробниками цих трекерів. Набір налаштувань для кожного з трекерів використовували для порівняння. У дослідженні трекери самостійно виконували сегментацію на вокалізовані фрагменти і паузи, застосовуючи датасет VTR-TIMIT. Проведений порівняльний аналіз показав достатньо високу точність визначення формантних частот порівняно з існуючими формант-трекерами.

Ключові слова: *мовний сигнал (МС); формантні частоти; спектральна декомпозиція; обчислювальний алгоритм; вейвлет-аналіз.*

1. ВСТУП

У сучасному світі все більше значення приділяють інтерфейсам, які використовують мовне введення і виведення для взаємодії між користувачем і комп'ютером. Тому розробник систем мовної інформації має надавати увагу все більшій кількості налаштувань і параметрів у підсистемах, що реалізують акустичний інтерфейс.

Задача розпізнавання мови (у багатьох своїх проявах: від сегментації промови до верифікації та ідентифікації особи) нині є вкрай актуальною. Свідченням цього є зростаюча кількість публікацій і конференцій із цієї тематики [1], а також відкриття в транснаціональних корпораціях департаментів, що орієнтовані на дослідження в мовній інформації.

Поліпшення існуючих систем розпізнавання мови дозволить істотно спростити взаємодію людини з комп'ютером у тому випадку, коли використання класичних інтерфейсів неможливо (наприклад, під час керування автомобілем або для людей з обмеженими фізичними можливостями), а також зробити подібну роботу більш комфортною та ефективною.

Необхідність досліджень із цієї тематики пояснюється незадовільними результатами існуючих систем при низькому співвідношенні сигнал/шум, залежностями результату від людини, а також невисокою швидкістю роботи подібного роду систем [2].

Дослідження методів визначення формант та розроблення нових точніших методів дозволить знизити похибку визначення та підвищити точність роботи систем ідентифікації мовної інформації.

2. ДОСЛІДЖЕННЯ ХАРАКТЕРНИХ ОСОБЛИВОСТЕЙ МОВНИХ СИГНАЛІВ

У системах ідентифікації мовного сигналу отримання характерних ознак мови особи, є однією з основних запорук успіху роботи системи ідентифікації мовної інформації. У завданнях аналізу мовного сигналу з метою визначення найважливіших характеристик, як правило, використовують методи його частотно-часового або спектрального уявлення, одними з найбільш ефективних є методи вейвлет-перетворення мовного сигналу. Найважливішим параметром, що характеризує спектр (розподіл енергії або амплітуди по частотах) мовного сигналу є форманти [3], які визначають як концентрацію енергії в обмеженій частотній області. Форманта характеризується частотою, шириною частотної смуги й амплітудою. Зважаючи на складність визначення й опису формантної частотної смуги, часто в дослідженнях під частотою форманти мають на увазі частоту максимальної амплітуди в межах форманти. Тому часто можна зустріти визначення форманти, як амплітудний сплеск на графіку спектра [4], що має частоту, відповідну до частоти піку цього сплеску. Форманти прийнято позначати F0-F6. Форманта F0 називається також частотою основного тону мовного сигналу. Форманти F1-F6 визначають концентрацію енергії мовного сигналу по частоті і характеризують вокалізовані (як правило, голосні) звуки.

Існує безліч методів виявлення формант [5–7], основними з яких є кепстральний аналіз і метод лінійного передбачення. Кепстральний аналіз є математичною основою нелінійних методів виділення сигналів. Кепстр – це математичне перетворення, що полягає в тому, що спектральному перетворенню піддається спектр функції. Якщо взяти спектр від спектра, то вийде сама початкова функція. Природно, що для отримання більш гладкої функції в результаті спектрального перетворення необхідно згладити початковий спектр. Для цього найчастіше використовують логарифмування початкового спектра або його модуля. Такий варіант перетворення прийнято називати кепстр. В отриманому кепстрі виявляються гармонійні складові [8].

Метод лінійного передбачення (метод, який використовує LPC-коефіцієнти) полягає в пошуку комплексних коренів полінома з LPC-коефіцієнтами (коефіцієнти фільтра мовного тракту) і надалі їхні наступні перетворення. Вважається, що метод лінійного передбачення [9] забезпечує високу обчислювальну швидкість, незначну складність і максимальну точність оцінювання формант.

Зазвичай використовують ефективний метод обчислення коефіцієнтів лінійного передбачення, заснований на автокореляції, що застосовує алгоритм Левінсона – Дарбина. У задачі розпізнавання мови часто застосовують коефіцієнти лінійного передбачення, обчислені на основі статичного закону людського слуху [10, 11]. Відмінності від простого обчислення коефіцієнтів лінійного передбачення полягають у такому. По-перше, застосовується шкала Баркова і логарифмічна компресія амплітуди до застосування алгоритму Левінсона – Дарбина. По-друге, для відповідності законам людського слуху потужність спектральних компонент зводиться до ступеня 0,33. Ця модифікація застосовується в частотної області, що призводить до того, що автокореляційні коефіцієнти не можуть бути обчислені безпосередньо, отже, необхідним є додаткове перетворення Фур'є [12].

Описані методи використовують у різних комбінаціях в чотирьох основних формант-трекерах PRAAT [13], SNACK [14], ASSP [15] і DEEP [16].

PRAAT [13] виділяє форманти таким чином: для кожного фрагмента, що аналізується, застосовується вікно, подібне до вікна Гаусса. Обчислюються коефіцієнти LPC за алгоритмам Бурга [17]. Кількість максимумів, яку обчислює цей алгоритм, удвічі перевищує максимальну кількість формант, тому рекомендується встановлювати максимальну кількість формант кратну 0,5. Спочатку алгоритм знаходить максимальну кількість формант у всьому діапазоні від 0 Гц і вище. Тому знайдені форманти можуть мати аномально низькі або високі частоти, що пов'язано з артефактами алгоритму LPC.

SNACK [14] застосовує той самий підхід, що і PRAAT. Цей метод використовує відбір формантів за мінімальними значеннями вагових коефіцієнтів, в основі яких лежать відмінності між фрагментами, частота та ширина форманти. На відміну від PRAAT використовується лише один параметр, пов'язаний із відстеженням функції. Цей параметр дозволяє визначити значення першої форманти, вищі форманти знаходять на основі заданого значення для першої форманти F1. Таким чином метод вважає, що всі інші форманти пов'язані з F1 коефіцієнтами прямого зв'язку. Цей метод не дає можливості змінити поведінку обчислення або вагові коефіцієнти для інших формант.

У методі ASSP [15] значення формант отримують шляхом визначення резонансних частот за рахунок розв'язання коренів полінома LPC, використовуючи метод автокореляції на основі алгоритму Спліт – Левінсона (SLA) [18]. Потім виконується класифікація резонансних частот як формант із використанням частоти Писаренко [19] та встановлення границі діапазону частот формант, що визначаються з номінальної частоти першої форманти F1. У методі вказується, що форманті частоти необхідно збільшити на 12 % для точнішого обчислення формант жіночих голосів.

DEEP [16] використовує нейронні мережі для визначення формант. Вхідними даними є отримання на основі LPC-методу кепстральних коефіцієнтів, які корегуються частотою основного тону, отриманою зі спектрограми. Навчання нейронної мережі виконувалось на основі тестової частини мовного датасету VTR-TIMIT [20]. VTR-TIMIT є відкритим датасетом із 516 видами записів від 186 осіб, що є носіями англійської мови. У датасеті виконана ручна корекція та маркування формант спеціальної групою експертів, тому саме цей датасет найчастіше використовується для оцінювання точності роботи формант-трекерів.

Є ряд досліджень, що стосуються порівняння формант-трекерів, однак серед них не можна виділити такий, що має найкращу ефективність [21, 22].

На основі викладеного можна зробити висновок про деяку складність, а частково, навіть про відсутність фізичної аргументації механізмів визначення формант, а також відсутність достатньої точності формант-трекерів, що потребує побудови методу визначення формант, який має достатнє фізичне обґрунтування та достатню точність порівняно з аналогами.

3. МЕТОДОЛОГІЯ ДОСЛІДЖЕННЯ

Форманти є одними з основних елементів ідентифікації особи в мовному сигналі тому, що природа їхнього походження пов'язана з порожнинами людського мовного тракту. Зважаючи на індивідуальність подібних компонентів для кожної людини, можна дійти висновку, що визначення формантних частот є важливим компонентом побудови системи ідентифікації мовної інформації.

Дослідження формантних атрибутів найчастіше виконується шляхом:

- порівняння спектра формант однакових фонемічних елементів (ударних голосних, голосних у кінці або на початку слів тощо);
- порівняння спектрів формант для зрізів спектрограм (кожного елемента спектрограми або всередині слів, складів та ін.);
- порівняння динамічних змін частот формант уздовж усього МС чи у важливих його компонентах.

Виокремлення формант супроводжує ряд проблем, пов'язаних з їхньою динамічною зміною у процесі мовлення. Навіть однакові голосні змінюють формантний набір залежно від свого розташування у складі слів, складів тощо. Труднощі також визивають проблеми, пов'язані з близьким розташуванням піків під час аналізу спектрограм і проблемами правильного визначення піків максимумів формант на спектрограмі. Визначення розташування формант на спектрограмах МС достатньо легко виконується людиною, але автоматизація цього процесу визиває деякі труднощі. Типове представлення спектрограми з розміченими людиною розташуваннями формант представлено на рис. 1, що відповідає зонам розташування формант, причому характеризується не тільки середньою частотою форманти, а також її шириною.

Проведення подібного розмічення є досить складною задачею, зважаючи не велику кількість конкуруючих частотних піків. Трудомісткість подібного роду маркування формант досить висока, тому в експериментальних дослідженнях використовують заздалегідь розмічені данні, що можна знайти в різних типах мовних датасетів. В наступному дослідженні будемо використовувати датасет VTR-TIMIT, що має підготовлений набір розмічених даних подібного типу.

Крім того, під час проголошення окремих видів звуків на положення формант можуть впливати безліч факторів, що може приводити до коливань формантних частот, а на окремих фрагментах навіть відсутності деяких із них.

Існує декілька підходів для визначення положень формант на частотній шкалі, але всі вони базуються на аналізі та перетвореннях спектрограми мовного сигналу. При виділенні формантних частот першим етапом дослідження завжди є побудова спектрограми за визначеними дослідником критеріями. Серед них є ширина фрейму, тип вейвлет-базису, частотний діапазон та ін.

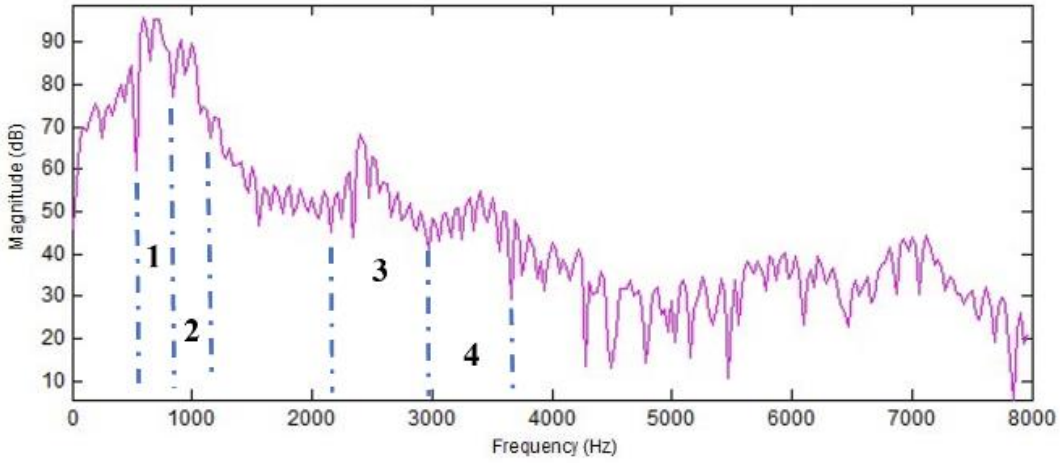

Рис. 1. Визначення формант людиною

Наступним етапом є таке представлення спектрограми, що дозволить провести сегментацію формантних частот на основі різних математичних методів, найчастіше всі вони базуються на алгоритмах кластеризації або на побудові обвідної спектра. Найвідомішими з алгоритмів, що використовують обвідну спектра, є:

- метод лінійного передбачення – коли обвідна спектра будується на алгоритмі LPC [23];
- апроксимації спектра кубічними сплайнами або іншими видами функцій [24].

Однак згідно з дослідженнями обидва алгоритми мають практично однакову точність при різній обчислювальній складності [25].

Приклад обвідної на основі алгоритму лінійного передбачення виділеного спектра для стандартного фрагмента дослідження (20 мс) зображено на рис. 2. На основі цього рисунка виділяють локальні максимуми обвідної для спектра мовного сигналу, які розглядають як центри формантних частот.

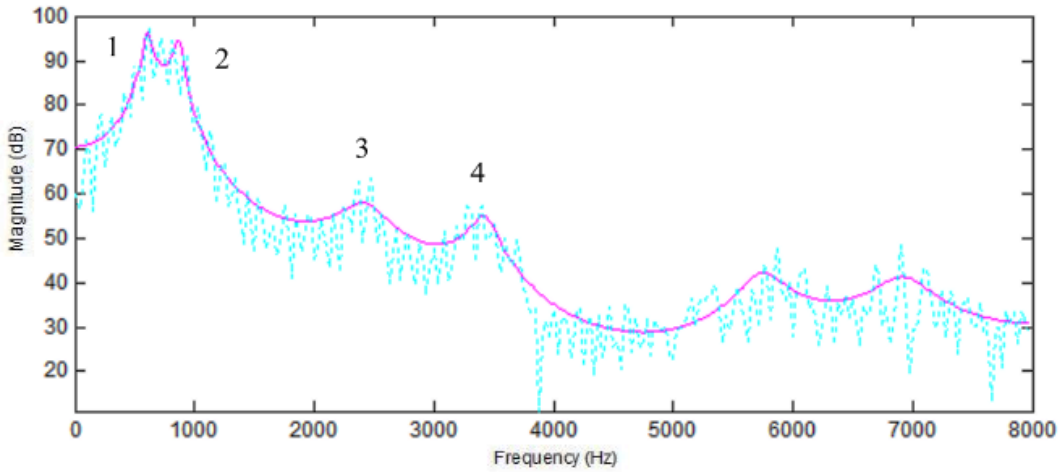

Рис. 2. Побудова обвідної для спектра мовного сигналу (числами зображено формантні діапазони 1–4)

Проведені попередні дослідження вказують на відповідність розподілу частот відносно голосних звуків, які вносять найбільшу вагу в формування формант в мовному сигналі. Установлено, що

частина голосних звуків у більшості мов розташована в частотному діапазоні 200–500 Гц, а інша частина голосних звуків у діапазоні від 500 до 1500 Гц. Зважаючи на це, раціональним є окремий розгляд цих частотних діапазонів, під час формування характерних ознак мовного сигналу, що дозволяє підвищити кількість параметрів, та набирати більшу статистику при визначенні максимумів формантних частот.

Результатом проведеного огляду підходів до визначення формантних частот став алгоритм (рис. 3), що складається з таких етапів.

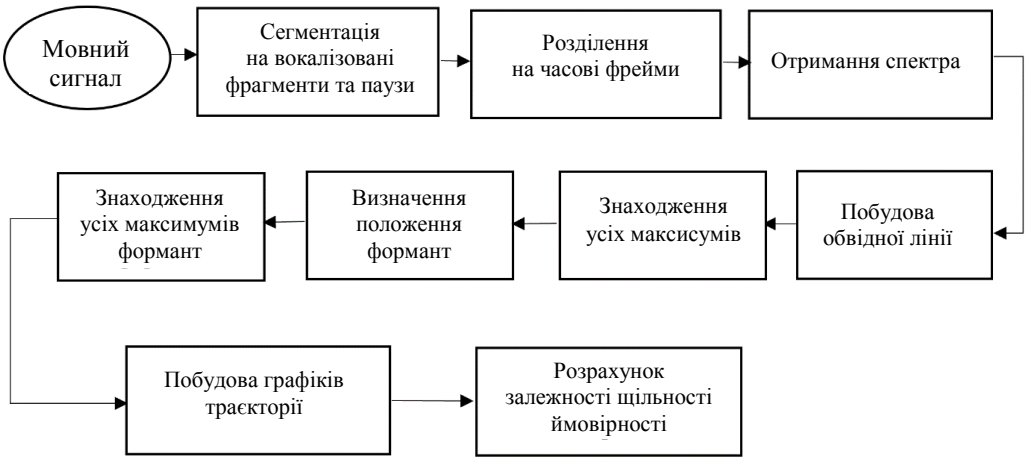

Рис. 3. Алгоритм визначення формантних частот

1. Сегментація мовного сигналу на вокалізовані фрагменти та паузи.
2. Розбиття вокалізованих фрагментів МС на часові фрейми.
3. Для кожного фрагмента отримання спектра на основі вейвлет-перетворення.
4. Побудова обвідної лінії.
5. Знаходження всіх максимумів.
6. Визначення положень формантних діапазонів.
7. Отримання максимумів формантних діапазонів.
8. Побудова графіків траєкторії положення формант (рис. 4).

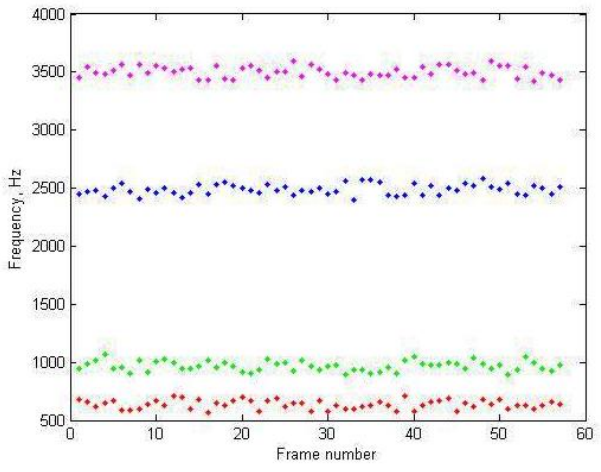

Рис. 4. Траєкторія положення формант за номерами фреймів

9. Розрахунок залежності щільності ймовірності розподілу кожної із чотирьох формантних частот (максимумів формантних частот).

Сегментація МС на вокалізовані фрагменти й паузи виконується методом оцінювання змін фрактальної розмірності [26]. У роботі визначено, що фрактальна розмірність D для невокалізованих фрагментів у 99 % випадків перебуває в межах $1{,}04 \leq D \leq 1{,}45$, а фрактальна розмірність вокалізованих фрагментів не спостерігалася менше $D = 1{,}55$ для часового вікна розміром 20 мс.

Операцію розбиття виділених вокалізованих фрагментів МС на часові фрейми виконувалась фреймами 10–20 мс для тестування робочої спроможності методу.

Отримання спектра мовного сигналу виконувалося з використанням комплексного вейвлету Морле на основі віконної функції Гаусса.

Відповідно до розглянутого підходу до визначення формантних частот необхідно побудувати обвідну спектра для кожного з фрагментів МС. Проведені в багатьох роботах дослідження при побудові обвідної пропонують використання безлічі підходів, але у разі визначення форматних частот саме кубічні сплайни дозволяють найточніше визначити вид цієї функції обвідної за заданими значеннями спектра МС. Фактично побудова функції обвідної представляє собою задачу інтерполяції. Визначимо, як ця задача розв'язується у нашому випадку.

Розглянемо спектр МС, як функцію, визначену в точках $x_0, x_1, x_2, \dots, x_N$, де відомі значення деякої функції $f(x)$, а саме $y_0, y_1, y_2, \dots, y_N$.

Інтерполюємо графік функції $f(x)$ побудовою функції $F(x)$ такої, що набуває в зазначених точках ті ж значення, тобто $F(x_0) = y_0, F(x_1) = = y_1, \dots, F(x_N) = y_N$.

З геометричної точки зору стоїть задача пошуку певного типу кривої $y = F(x)$, що проходить через набір представлених точок. Подібна задача може мати багато розв'язків чи розв'язки можуть бути відсутні. У процесі використання кубічного сплайну наша функція $F(x)$ представляє набір фрагментів, що визначені на кожному інтервалі $[x_{k-1}; x_k]$, та може мати вигляд

$$F_k(x) = a_k + b_k(x - x_k) + c_k(x - x_k)^2 + d_k(x - x_k)^3 \quad (1)$$

$F = F_1$ для $[x_0; x_1]$

…

$F = F_N$ для $[x_{N-1}; x_N]$ де N – кількість точок, за якими проводиться інтерполяція.

Звичайно, що коефіцієнти полінома a_k, b_k, c_k, d_k будуть відрізнятися для кожного з відрізків $[x_{k-1}; x_k]$.

Для визначення коефіцієнтів накладається ряд умов, серед яких такі:

Рівність других похідних функції на кінцях відрізка $[x_0; x_N]$:

$$F''(x_0) = 0, F''(x_N) = 0. \quad (2)$$

Безперервності першої і другої похідної функції $F(x)$, а також:

$$\begin{gathered} F_{k-1}(x_{k-1}) = F_k(x_{k-1}), \\ F'_{k-1}(x_{k-1}) = F'_k(x_{k-1}), \\ F''_{k-1}(x_{k-1}) = F''_k(x_{k-1}), \\ \text{при } k = 2,3, \dots, N \end{gathered} \quad (3)$$

Запишемо похідні функції F_k у вигляді

$$\begin{gathered} F'_k(x) = b_k + 2c_k(x - x_k) + 3d_k(x - x_k)^2, \\ F''_k(x) = 2c_k + 6d_k(x - x_k). \end{gathered} \quad (4)$$

Зважаючи на описані вище умови, отримуємо таку систему рівнянь:

$$\begin{gathered} a_1 - b_1h_1 + c_1h_1^2 - d_1h_1^3 = y_0, \\ a_k = y_k, k = 1,2, \dots, N, \\ a_{k-1} = a_k - b_kh_k + c_kh_k^2 - d_kh_k^3, \\ k = 2,3, \dots, N, \\ b_{k-1} = b_k - 2c_kh_k + 3d_kh_k^2, \\ k = 2,3, \dots, N, \\ c_{k-1} = c_k - 3d_kh_k, \ k = 2,3, \dots, N, \\ c_1 - 3d_1h_1 = 0, \\ c_N = 0, \end{gathered} \quad (5)$$

де $h_k = x_k - x_{k-1}, \ k = 1,2, \dots, N,$

$l_k = (y_k - y_{k-1})/h_k, \ c_0 = 0.$

Існує досить багато варіантів розв'язання представленої системи рівнянь. Для розв'язання будемо використовувати метод прогонки.

В основі методу лежить застосування коефіцієнтів, що будуть корегуватися у процесі налаштування. Визначимо коефіцієнти у вигляді

$$\begin{gathered} \delta_1 = -h_2/(2(h_1 + h_2)), \\ \lambda_1 = 3(l_2 - l_1)/(2(h_1 + h_2)), \\ \delta_{k-1} = -\frac{h_k}{2h_{k-1} + 2h_k + h_{k-1}\delta_{k-2}}, \\ k = 3,4, \dots, N, \end{gathered} \quad (6)$$

$$\lambda_{k-1} = \frac{(3l_k - 3l_{k-1} - h_{k-1}\lambda_{k-2})}{(2h_{k-1} + 2h_k + h_{k-1}\delta_{k-2})}. \quad (7)$$

На основі коефіцієнтів прогону отримуємо c_k за алгоритмом зворотного прогону:

$$\begin{gathered} c_{k-1} = \delta_{k-1}c_k + \lambda_{k-1}, \\ k = N, N-1, N-2, \dots, 2. \end{gathered} \quad (8)$$

Усе це дає можливість визначити коефіцієнти b_k і d_k за формулами

$$\begin{gathered} b_k = l_k + \frac{2c_kh_k + h_kc_{k-1}}{3}, \ k = 1,2, \dots, N \\ d_k = (c_k - c_{k-1})/(3h_k), \ k = 1,2, \dots, N. \end{gathered} \quad (9)$$

Таким чином, результатом описаного алгоритму будуть коефіцієнти b_k, c_k, d_k для кожного з розглянутих інтервалів графіка спектра МС.

Результатом розрахунку обвідної буде графік спектра МС у заданому фрагменті та обвідної спектра для цього ж інтервалу (рис. 5).

Визначаючи максимуми обвідної спектра (рис. 5), отримуємо максимуми чотирьох перших формант. Набір формант для кожного фрейму зображено на графіку відповідно до частоти форманти та номера фрейму, з якого вона була отримана (рис. 4). Аналіз цього графіка показує достатньо високу стабільність визначення формантних частот для розглянутого у попередньому дослідженні мовного сигналу.

З метою оцінювання та порівняння запропонованого методу з методами, що використовуються у відомих формант-трекерах, виконано наступне дослідження.

Для дослідження розглядались формант-трекери PRAAT, SNACK, ASSP і DEEP. Налаштування кожного з них здійснювалося на основі набору параметрів за замовчуванням, що

було закладено розробниками цих трекерів. Набір налаштувань для кожного з трекерів представлено у табл. 1. Основні параметри налаштувань відомих трекерів визначено на основі [21]. Необхідно також зазначити, що кожен із цих відомих трекерів має особливості використання відповідно до статі особи, так: PRAAT оптимізовано для мовних сигналів жінок, SNACK та ASSP оптимізовано під мовні сигнали чоловіків. DEEP, як описано у розробника, оптимізувався на наборі даних 67 жінок на 95 чоловіків, тому, можливо, він буде точніше працювати з мовними сигналами чоловіків.

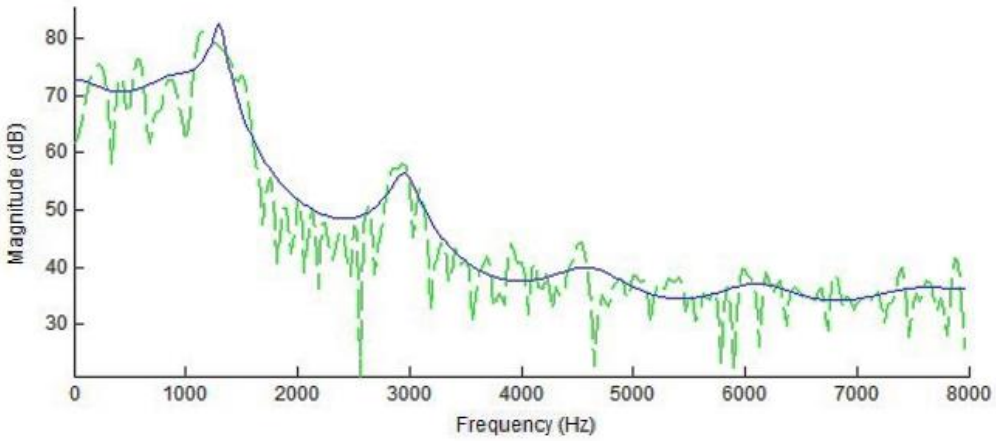

Рис. 5. Обвідна спектра МС

***Таблиця* 1**

Налаштування формант-трекерів, використаних у дослідженні

Параметр	PRAAT	SNACK	ASSP	DEEP	Пропонований метод
formants	5	4	4	4	4
LPC	10	12	18	n/a	n/a
preemph	50 Hz/oct	0,7	0.96	unk	–
window	gauss	cos4	blackman	unk	gauss
w.length, ms	25	25	25	unk	25
stepsize, ms	10	10	10	10	10

Налаштування часових інтервалів власного методу має відповідати розміру подібних інтервалів для інших формант-трекерів для коректного порівняння, тому використовують розмір вікна 25 ms із кроком 10 ms.

У дослідженні трекери самостійно виконували сегментацію на вокалізовані фрагменти і паузи, застосовуючи датасет VTR-TIMIT, у випадку некоректної сегментації (невокалізований фрагмент вважався вокалізованим) помилкові результати сегментації вилучалися з розгляду.

Як параметри порівняння використовували середньоквадратичне відхилення формант між еталонною розміткою мовних сигналів та результатами формант-трекерів (табл. 2).

***Таблиця* 2**

Середньоквадратичне відхилення визначення формант (Гц)

Formant	Female/Male	PRAAT	SNACK	ASSP	DEEP	Proposed Method
F1	f	97	104	82	61	78
	m	161	92	71	53	70
F2	f	185	197	172	99	89
	m	215	209	109	84	91
F3	f	194	183	215	156	167
	m	247	261	113	171	144

Розгляд саме трьох формант, замість чотирьох, пов'язаний із тим, що в датасеті VTR-TIMIT розмічені експертами лише три форманти. Крім того, проведений аналіз цього датасету [27] показав середнє відхилення частоти максимуму для 1–3 формант відповідно 78, 100, 111 Гц, тому отримані значення достатньою мірою відповідають раніше проведеним дослідженням.

Порівняльний аналіз показує достатньо високу точність визначення формантних частот порів-

няно з існуючими формант-трекерами. Виняток стосується DEEP, але зважаючи не те, що він натренований на саме цьому датасеті, можна вважати, що метод достатньо добре показує себе порівняно з іншими. Поряд із цим, необхідно зазначити простоту реалізації, низьку обчислювальну складність, швидкість і відповідність методу наявним фізичним процесам.

4. ВИСНОВКИ

На основі огляду існуючих підходів до алгоритму визначення формантних частот установлено наявність складності їхньої реалізації та часткову відсутність фізичного обґрунтування операцій, що входять до їхнього складу. Запропоновано метод визначення формантних частот, що як вхід використовує спектральне розкладання мовного сигналу на два діапазони, що відповідають частотним межам голосних звуків. Представлено алгоритм обчислення обвідної, а також опис методу визначення параметрів формантних частот до побудови функції щільності ймовірності, що може бути використана під час безпосереднього порівняння в задачі прийняття рішення щодо мовної ідентифікації особи. Порівняння методу з іншими подібними методами показало задовільну точність. Використання запропонованого методу разом із методологією сегментації мовного сигналу дозволять знизити похибку під час виділення формант, що приведе до підвищення точності прийняття рішення у процесі побудови систем ідентифікації мовного сигналу.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] Yegnanarayana, B., Veldhuis, R. N. J. (1998). Extraction of vocaltract system characteristics from speech signals, IEEE Trans. Speech Audio Process, 6 (4), 313–327.

[2] Kim, C., Seo, K., & Sung, W. A Robust (2006). Formant Extraction Algorithm Combining Spectral Peak Picking and Root Polishing. *EURASIP Journal on Applied Signal Processing*, 1–16.

[3] Wet, F. D., Weber, K., Boves, L., Cranen, B., Bengio, S., & Bourlard, H. (2004). Evaluation of Formant-Like Features for Automatic Speech Recognition. *Journal of the Acoustical Society of America*, 116, 1781–1791.

[4] Mallat, S. (1999.) A Wavelet Tour of Signal Processing. *Academic Press.*

[5] Yan, Q., Vaseghi, S., Zavarehei, E., Milner, B., Darch, J., White, P., & Andrianakis, I. (Jul. 2007). Formant Tracking Linear Prediction Model using HMMs and Kalman Filters for Noisy Speech Processing. *Computer Speech and Language*, vol. 21, pp. 543–561.

[6] Messaoud, Z. B., Gargouri, D., Zribi, S., & Hamida, A. B. (2009). Formant Tracking Linear Prediction Model using HMMs for Noisy Speech Processing. *International Journal of Signal Processing*, vol. 5, pp. 291–296.

[7] Cooke, M., Barker, J., Cunningham, S., & X. Shao (2006). An audio-visual corpus for speech perception and automatic speech recognition. *Journal of the Acoustical Society of America*, vol. 120.

[8] Acero, A. Formant Analysis and Synthesis using Hidden Markov Models (1999). In *Proc. of the Eurospeech Conference.* Budapest.

[9] Veldhuis, R. (1997). A computationally e$cient alternative for the LF model and its perceptual evaluation. *J. Acoust. Soc.*, 103 (1), 566–571.

[10] Bazzi, I., Acero, A., & Deng, L. (2003). An expectation maximization approach for formant tracking using a parameter-free non-linear predictor. In *Proc. ICASSP*, vol. 1, 464–467.

[11] Ali, J. A. M. A., Spiegel, J. V. D., & Mueller P. (2002). Robust Auditory-based Processing using the Average Localized Synchrony Detection. In *IEEE Transaction Speech and Audio Processing*.

[12] Vakman, D. (1996). On the analytic signal, the Teager-Kaiser energy algorithm, and other methods for defining amplitude and frequency. *IEEE Trans. Signal Process*, SP-44, 791–797.

[13] Boersma, P., & D. Weenink, (2017). Praat: doing phonetics by computer [Computer program]. Version 6.0.23, retrieved 2021-05-17. http://www.praat.org/

[14] Kåre Sjölander(2020) The Snack Sound Toolkit [Computer program]. https://www.speech.kth.se/snack/

[15] Scheffer, M. (2017). Available: Advanced Speech Signal Processor (libassp), retrieved 2021-05-17. http://www.sourceforge.net/projects/libassp.

[16] Keshet, J. (2017). DeepFormant, retrieved 2021-05-25. https://github.com/MLSpeech.

[17] Gray, A., & Wong, D. (1980, Dec.). The Burg algorithm for LPC speech analysis/Synthesis. *In IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 6, pp. 609–615.

[18] Krishna, H., & Wang, Y. (1993). The Split Levinson Algorithm is Weakly Stable. *SIAM Journal on Numerical Analysis*, 30(5), 1498–1508., http://www.jstor.org/stable/2158249

[19] So, H. C., & Chan, K. W. (2004). Reformulation of Pisarenko Harmonic Decomposition Method for Single-Tone Frequency Estimation. *Signal Processing, IEEE Transactions* on. 52. 1128–1135. 10.1109/TSP.2004.823473.

[20] VTR Formants Database. http://www.ee.ucla.edu/~spapl/VTRFormants.rar

[21] Nearey, T. & Assmann, P. & Hillenbrand, J. (2002). Evaluation of a strategy for automatic formant tracking. *The Journal of the Acoustical Society of America*. 112. 2323. 10.1121/1.4779372.

[22] Schiel, Florian & Zitzelsberger. Thomas (2018). Evaluation of Automatic Formant Trackers. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation {LREC}*, Miyazaki, Japan.

[23] Markel, J. E. & Gray, A. H. (1982). *Linear Prediction of Speech.* New York, NY: Springer.

[24] Sun, Don X. (1995). Robust estimation of spectral center-of-gravity trajectories using mixture spline models. In *EUROSPEECH-1995*, 749–752.

[25] Schalk-Schupp, Ingo. (2012). Improved Noise Reduction for Hands-Free Communication in Automobile Environments. 10.13140/2.1.4068.6724.

[26] Бєлозьорова, Я. А. (2017). Ідентифікація диктора на основі кратномасштабного аналізу. Інженерія програмного забезпечення: наук. журн., 1(29). 15–25.

[27] Deng, L., Cui, X., Pruvenok, R., Huang, J., Momen, S., Chen, Y. N., & Alwan, A. (2006). A Database of Vocal Tract Resonance Trajectories for Research in Speech Processing. In Proc. of the Int. Conf. on Acoustics, Speech, and Signal Processing.

REFERENCES

[1] Yegnanarayana, B., Veldhuis, R. N. J. (1998). Extraction of vocaltract system characteristics from speech signals, IEEE Trans. Speech Audio Process, 6 (4), 313–327.

[2] Kim, C., Seo, K., & Sung, W. A Robust (2006). Formant Extraction Algorithm Combining Spectral Peak Picking and Root Polishing. *EURASIP Journal on Applied Signal Processing*, 1–16.

[3] Wet, F. D., Weber, K., Boves, L., Cranen, B., Bengio, S., & Bourlard, H. (2004). Evaluation of Formant-Like Features for Automatic Speech Recognition. *Journal of the Acoustical Society of America*, 116, 1781–1791.

[4] Mallat, S. (1999.) A Wavelet Tour of Signal Processing. *Academic Press.*

[5] Yan, Q., Vaseghi, S., Zavarehei, E., Milner, B., Darch, J., White, P., & Andrianakis, I. (Jul. 2007). Formant Tracking Linear Prediction Model using HMMs and Kalman Filters for Noisy Speech Processing. *Computer Speech and Language*, vol. 21, pp. 543–561.

[6] Messaoud, Z. B., Gargouri, D., Zribi, S., & Hamida, A. B. (2009). Formant Tracking Linear Prediction Model using HMMs for Noisy Speech Processing. *International Journal of Signal Processing*, vol. 5, pp. 291–296.

[7] Cooke, M., Barker, J., Cunningham, S., & X. Shao (2006). An audio-visual corpus for speech perception and automatic speech recognition. *Journal of the Acoustical Society of America*, vol. 120.

[8] Acero, A. Formant Analysis and Synthesis using Hidden Markov Models (1999). In *Proc. of the Eurospeech Conference.* Budapest.

[9] Veldhuis, R. (1997). A computationally e$cient alternative for the LF model and its perceptual evaluation. *J. Acoust. Soc.*, 103 (1), 566–571.

[10] Bazzi, I., Acero, A., & Deng, L. (2003). An expectation maximization approach for formant tracking using a parameter-free non-linear predictor. In *Proc. ICASSP*, vol. 1, 464–467.

[11] Ali, J. A. M. A., Spiegel, J. V. D., & Mueller P. (2002). Robust Auditory-based Processing using the Average Localized Synchrony Detection. In *IEEE Transaction Speech and Audio Processing*.

[12] Vakman, D. (1996). On the analytic signal, the Teager-Kaiser energy algorithm, and other methods for defining amplitude and frequency. *IEEE Trans. Signal Process*, SP-44, 791–797.

[13] Boersma, P. & Weenink, D. (2017). Praat: doing phonetics by computer [Computer program]. [Online]. Available: Version 6.0.23, retrieved 2022-05-17 from http://www.praat.org/

[14] Kåre Sjölander (2020) The Snack Sound Toolkit [Computer program]. [Online]. Available: https://www.speech.kth.se/snack/

[15] Scheffer, M. (2017). [Online]. Available: Advanced Speech Signal Processor (libassp), retrieved 2022-05-17 from http://www.sourceforge.net/projects/libassp.

[16] Keshet, J. (2017). [Online]. Available: DeepFormant, retrieved 2022-05-25 from https://github.com/MLSpeech.

[17] Gray, A., & Wong, D. (1980, Dec.). The Burg algorithm for LPC speech analysis/Synthesis. *In IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 6, pp. 609–615.

[18] Krishna, H., & Wang, Y. (1993). The Split Levinson Algorithm is Weakly Stable. SIAM Journal on Numerical Analysis, 30(5), 1498–1508. [Online]. Available: Retrieved Juny 11, 2021, from http://www.jstor.org/stable/2158249

[19] So, H.C. & Chan, K.W. (2004). Reformulation of Pisarenko Harmonic Decomposition Method for Single-Tone Frequency Estimation. Signal Processing, IEEE Transactions on. 52. 1128–1135. 10.1109/TSP.2004.823473.

[20] VTR Formants Database. [Online]. Available: http://www.ee.ucla.edu/~spapl/VTRFormants.rar

[21] Nearey, Terrance & Assmann, Peter & Hillenbrand, James. (2002). Evaluation of a strategy for automatic formant tracking. The Journal of the Acoustical Society of America. 112. 2323. 10.1121/1.4779372.

[22] Schiel, Florian and Zitzelsberger, Thomas. "Evaluation of Automatic Formant Trackers", Proceedings of the Eleventh International Conference on Language Resources and Evaluation {LREC} 2018, Miyazaki, Japan.

[23] Markel, J.E. & Gray, A.H. (1982). Linear Prediction of Speech. New York, NY: Springer.

[24] Sun, Don X. (1995): "Robust estimation of spectral center-of-gravity trajectories using mixture spline models", In EUROSPEECH-1995, 749–752.

[25] Schalk-Schupp, Ingo. (2012). Improved Noise Reduction for Hands-Free Communication in Automobile Environments. 10.13140/2.1.4068.6724.

[26] Belozyorova Y. A. (2017). Speaker identification based on multiple-scale analysis. Scientific journal. Software engineering, 1(29), 15–25 [in Ukrainian].

[27] Deng, L., Cui, X., Pruvenok, R., Huang, J., Momen, S., Chen, Y. N. & Alwan, A. (2006). A Database of Vocal Tract Resonance Trajectories for Research in Speech Processing. In Proc. of the Int. Conf. on Acoustics, Speech, and Signal Processing.

Стаття надійшла до редколегії 21.03.2023

Method of determining formant frequencies using spectral decomposition of speech signal

Formants are one of the main components of speaker identification systems and the accuracy of formant determination is the basis for the efficiency of speaker identification systems. Improving existing speech recognition systems will significantly simplify human-computer interaction when the use of classic interfaces is not possible, as well as make such work more comfortable and efficient.

The necessity for research on this topic is due to unsatisfactory results of existing systems with low signal-to-noise ratio, the dependence of the result on humans, as well as low speed of such systems.

The following four main formant trackers were used for comparison with the proposed method: PRAAT, SNACK, ASSP and DEEP. There are a number of studies concerning the comparison of formant trackers, but among them it is impossible to single out the one that has the best efficiency.

The selection of formants is accompanied by a number of problems associated with their dynamic change in the language process. The complexity is also caused by a number of problems related to the close location of the peaks in the analysis of spectrograms and the problems of correctly determining the peaks of the formant maxima on the spectrogram. Determining the location of the formant on the spectrograms of the vocal signal is quite easy to perform by man, but the automation of this process causes some difficulties.

The selection of frequency formants was proposed to be performed in several stages. The result of the review of approaches to the determination of formant frequencies has been the algorithm consisting of the following nine stages. The segmentation of vocal signal into vocalized fragments and pauses is performed by estimating changes in fractal dimension. Obtaining the spectrum of the vocal signal has been performed using a complex Morlet wavelet based on the Gaussian window function. PRAAT, SNACK, ASSP and DEEP formant trackers have been considered for the study. Each of them has been configured on the basis of a set of default parameters set by the developers of these trackers. A set of settings for each of the trackers has been used for comparison. In the study, trackers independently have been performed segmentation into vocalized fragments and pauses using the VTR-TIMIT dataset.

The comparative analysis has been showed a fairly high accuracy in determining the formant frequencies in comparison with existing formant trackers.

Keywords: *speech signal; formant frequencies; spectral decomposition; computational algorithm; wavelet analysis.*

Сергій Зибін,
д-р техн. наук, проф.
Завідувач кафедри інженерії програмного забезпечення Національного авіаційного університету. Київ, Україна.

Serhii Zybin,
Dr. Sci. (Engin.), Prof.
Head of the Software Engineering Department of the National Aviation University. Kyiv, Ukraine.

Яна Бєлозьорова,
канд. техн. наук, доц.
Доцент кафедри інженерії програмного забезпечення Національного авіаційного університету. Київ, Україна.

Yana Belozyorova,
PhD (Engin.), Associate Prof.
Associate Professor of the Department of Software Engineering of the National Aviation University. Kyiv, Ukraine.