https://doi.org/10.31649/1997-9266-2022-162-3-36-51

УДК 004.942

О. М. Данильчук[1]
В. В. Ковтун[2]
О. Д. Никитенко[2]
Ю. Ю. Нестюк[2]
В. В. Присяжнюк[2]

ЕЛЕМЕНТИ МЕТОДОЛОГІЇ ПРЕЦИЗІЙНОГО ФОНЕТИЧНОГО АНАЛІЗУ ФОНОГРАМ УСНОГО МОВЛЕННЯ

[1]Донецький національний університет імені Василя Стуса;
[2]Вінницький національний технічний університет

Дослідження наріжного для сучасної лінгвістики об'єкта — процесу мовленнєвої і текстової міжособистісної комунікації, зважаючи на обсяг інфосфери двадцять першого століття, є неможливим без ґрунтовного та цілеспрямованого залучення інформаційних технологій з інших галузей знань, зокрема, комп'ютерних наук. Утворена в результаті порівняно молода наука — комп'ютерна лінгвістика, ставить за мету автоматичний аналіз природних мов у всіх спектрах їх реалізацій. З довгого списку актуальних задач, активно досліджуваних у парадигмі комп'ютерної лінгвістики, згадаємо автоматизацію складання та лінгвістичної обробки мовних корпусів, автоматизовану класифікацію та реферування документів, створення точних лінгвістичних моделей природних мов, екстракцію фактографічної інформації з неформалізованих лінгвістичних даних тощо. Рушійною силою для поліпшення результатів розв'язання цих дослідницьких задач потенційно є ефективна, строго формалізована методологія обчислювального фонетичного аналізу лінгвістичної інформації, особливо мовленнєвої. Цей тезис цілком відповідає вмісту статті, що доводить актуальність поданих в ній наукових і прикладних результатів. Відповідно, в роботі подані елементи методології прецизійного фонетичного аналізу фонограм усного мовлення з урахуванням явища фонетичної фузії. Математичний апарат створених методів ґрунтується на положеннях теорії розпізнавання образів, теорії інформації і акустичної теорії мовотворення. Цей базис забезпечив основу для аналітичної формалізації проблеми багатокритеріальності процесу розпізнавання мовних одиниць мовлення людиною. В результаті, запропоновано метод для достовірної кластеризації персональних фонетичних алфавітів мовців. Також запропоновані: метод для детектування потенційно ненадійно класифікованих мовних одиниць та коригування результатів процесу автоматизованого транскрибування мовленнєвих сигналів; метод оцінювання впливу середовища поширення досліджуваних мовленнєвих сигналів на результат транскрибування.

Ключові слова: комп'ютерна лінгвістика, класифікація мовних одиниць, автоматизоване транскрибування, фонетичний аналіз мовлення.

Вступ

Обчислювальний фонетичний аналіз є фундаментальним компонентом більшості інформаційних технологій розпізнавання природної мови, когнітивного аналізу мовлення, автоматизованого транскрибування мовлення тощо. Висока достовірність фонетичного аналізу є запорукою якісного результату функціонування всіх цих типів систем. Особливо актуальним є первинний фонетико-морфологічний аналіз флективних мов і мовлення. Основним джерелом похибок при цьому є фузія [1]—[4]. Це явище характеризує високу варіативність індивідуального озвучування фонем, особливо на стику морфем. Явище фузії об'єктивно зумовлено фонологічною еволюцією природної мови і не може ігноруватися в процесі створення прецизійних технологій обчислювального фонетичного аналізу мовлення.

Отже, *об'єктом дослідження* є фузія процесу злитого мовлення.

Огляд аналогів

Задача обчислювального фонетико-морфологічного аналізу мови або мовлення об'єктивно ускладнюється, по-перше, особливостями самої мови як процесу фізіологічно-когнітивної діяльності людини, по-друге, особливостями задіяних профільних інформаційних технологій.

Відзначимо основні інтегральні фактори першого джерела ускладнень [5]—[9].

Омонімія флексій. Омонімічними можуть бути флексії, які: належать до єдиної словозмінюваної парадигми; характеризують єдину лексико-граматичну категорію, але належать до різних словозмінюваних парадигм; інколи зустрічаються в парадигмах різних частин мови. Цей фактор є джерелом неоднозначності фонетико-морфологічного аналізу. Зменшити негативний вплив цього фактору можна, застосовуючи інформаційні технології аналізу лінгвістичного контексту та обчислювального фонетичного аналізу тощо.

Внутрішня флексія. Цей тип флексій проявляється у разі використання базової колекції мовних одиниць, репрезентативність якої залежить від наповненості словоформами. Якщо колекція не використовується, то необхідно сформулювати правила лінгвістичного поліморфізму, властивого досліджуваній мові.

Складні лексеми. Лексеми, фонація (напис) яких включає специфічні техніки артикуляції (спеціальні символи), які потребують визначення відмінювання для кожного компонента у складі словоформи.

Аналітичні словоформи. Аналітичні словоформи зустрічаються в багатьох мовах і можуть викликати значні ускладнення в фонетико-морфологічному аналізі, адже компоненти словоформи можуть бути розділені і навіть знаходитися в різних позиціях у реченні.

Великий лексичний фонд мови. Незважаючи на стрімку позитивну динаміку характеристик обчислювальної потужності і обсягів швидкої пам'яті сучасної комп'ютерної техніки, робота з базовою колекцією мовних одиниць досліджуваної мови (тим більше — з базовою колекцією словоформ) в разі реалізації фонетико-морфологічного аналізу досі залишається задачею високої обчислювальної складності.

Мінливість лексичного рівня мови. Оновлення колекцій мовних одиниць для відповідного типу систем фонетико-морфологічного аналізу не встигає за поліморфізмом живої мови (особливо, якщо брати до уваги діалекти), який проявляється у буденному явищі появі нових лексем (зокрема, термінів) і словоформ. Безсловникові системи обчислювального фонетико-морфологічного аналізу менше потерпають від впливу цього ускладнюючого фактору.

Зазначено лише найпоширеніші фактори природно-мовного походження, що негативно впливають на результативність обчислювального фонетико-морфологічного аналізу мовлення. В залежності від задіяних інформаційних технологій цей перелік розширюється.

Дослідимо сучасний стан теоретично-аналітичного базису актуальних інформаційних технологій обчислювального фонетико-морфологічного аналізу. Ґрунтуючись на результатах інформаційного пошуку [10], [11], виділимо два актуальних підходи — раціоналістичний та емпіричний. Перший підхід застосовує лінгвістичні знання для аналізу і синтезу мовних одиниць. Другий підхід заснований на узагальненні емпіричних даних, наприклад, у вигляді статистичної моделі мови (мовлення) [12], [13]. Втім, в сучасній комп'ютерній лінгвістиці найпродуктивнішими показали себе технології, які у певній пропорції інтегрують обидва цих підходи. За вмістом вживаної колекції мовних одиниць, як системоутворювального елемента для здійснення обчислювального фонетико-морфологічного аналізу, технології-аналізатори можна розділити на [14]—[17]: 1) системи з колекцією фонем-морфем; 2) системи з колекцією лексем і словоформ; 3) системи без базових колекцій.

Центральним елементом систем першого типу є колекція відносно фонетично та лінгвістично сталих мовних одиниць (морфем, фонем, алофонів) досліджуваної мови. Відповідна технологія-аналізатор розкладає мовленнєвий сигнал (текст) на визначену послідовність неділимих порцій, здійснюючи щодо кожної процедуру розпізнавання. Така елементарно-комбінаторна модель найчастіше застосовується для аналізу флективних та аглютинативних мов. При цьому порядок частин лексем визначається як конкатенація відповідних класів морфем у колекції. Для визначення порядку переходу між класами морфем зазвичай використовується математичний апарат кінцевих автоматів [18]. Кількість класів морфем у колекції визначається за результатом попередньої морфологічної класифікації досліджуваної мови. Окрім декларативної інформації щодо складу мор-

фем, у колекції може зберігатися і процедурна інформація. Така інформація визначає допустимий діапазон варіації образу морфем і найчастіше оформлюється як система продукційних правил [10], [11], [14], [19]. Для розпізнавання в системах цього типу найчастіше використовують емпіричні імовірнісно-статистичні методи [4], [12], [20].

Системи другого типу орієнтовані на обчислювальний морфологічний аналіз. Відповідно, вміст колекцій еталонів мовних одиниць у таких системах утворений морфемами та короткими лексемами. Системи цього типу розглядають словоформи як послідовність таких мовних одиниць, сформовану згідно з композиційними і(або) продукційними правилами. Під час дослідження словоформи система за визначеними правилами генерує для неї лему [21], [22]. Якщо у базовій колекції така лема присутня, то словоформа вважається розпізнаною. Якщо не зважати на ресурсоємність, то ефективність таких систем визначається здебільшого репрезентативністю вмісту базової колекції. Колекції морфем або лексем використовуються у фонетико-морфологічному аналізі для нормалізації досліджуваних словоформ. За наявності колекції морфем нормалізація реалізується у формі стемінгу. За наявності колекції лексем нормалізація реалізується у формі лем. Окремо згадаємо підклас систем другого типу, у яких використовується колекція словоформ. Призначенням таких систем є граматично-морфологічний аналіз, в якому у колекції представляється множина комбінацій словоформ, якій поставлена у відповідність множина граматичних міток [11], [19], [23]. За достатньо насиченої колекції джерелом похибок аналізу в цих системах є лише омонімія повної словоформи.

Недоліком всіх систем фонетико-морфологічного аналізу першого і другого типів є використання колекцій мовних одиниць великого об’сягу. Втім, за цим критерієм системи, орієнтовані на використання фонетико-морфологічних колекцій, виглядають виграшніше за умови прийнятної ефективності процесу розпізнавання.

Системи третього типу здійснюють фонетико-морфологічний аналіз виключно на основі математичних методів машинного навчання (машини опорних векторів, ЕМ-метод, генетичні алгоритми, мережі Кохонена тощо) [24]—[34]. Прийнятними є будь-які методи, здатні до графематичного аналізу, результатом якого є автоматичне або автоматизоване формування фонетико-морфологічних колекцій. Перевагою систем третього типу є методично обумовлена висока евристичність та адаптивність, що потенційно дозволяє розпізнавати мовні одиниці в мовному матеріалі з явною вираженою невизначеністю. Недоліком таких систем є складність навчання згаданих псевдоінтелектуальних методів, а також необхідність наявності начальних даних і обчислювальних ресурсів, обсяги яких перевищують необхідні для систем першого і другого типу не в рази, а на порядки.

Зважаючи на вищенаведені переваги і недоліки систем фонетико-лінгвістичного аналізу, сформулюємо *мету дослідження* як аналітичну формалізацію статично адекватної концепції фонетичного аналізу мовлення, варіативність якого буде врахована у парадигмі теорії інформації. *Предметом дослідження* будуть методи теорії імовірності і математичної статистики, теорії інформації, теорії розпізнавання образів і акустичної теорії мовотворення.

Моделі і методи

1. Постановка задачі дослідження

Функційне призначення типової сучасної інформаційної технології обчислювального аналізу мовленнєвих образів реалізується шляхом зіставлення параметризованного представлення досліджуваної мовної одиниці і відповідного їй еталону у визначеному параметричному просторі. При цьому основним джерелом невизначеності процесу зіставлення є біологічне походження мовленнєвого сигналу та його спотворення під час передавання і оброблення. Втім, акустична варіативність фонації мовленнєвих одиниць (перш за все, фонем), зумовлена існуванням діалектів, носить порівняно усталений характер. Враховуючи цей факт, припустимо, що реалізується одночасне порівняння досліджуваного образу фонограми з виголошеною фонемою x з кожним елементом $x_{r,j}$ множини еталонів $X_r = \{x_{r,j}\}$, де $j = \overline{1, J_r}$ — номер еталону, який характеризує відповідний діалект виголошення фонеми $r = \overline{1, R}$, де R — потужність фонетичного алфавіту, J_r — потужність множини визнаних діалектів для фонеми r. Тоді, якщо відстань $\rho\left(x / x_{r,j}\right)$, $r = \overline{1, J_r}$, між досліджуваним образом x і хоча б одним з елементів $x_{r,j}$ кластера *r*-ї фонеми не перевищує зада-

не порогове значення ρ_0:

$$\frac{1}{J_r}\sum_{j=1}^{J_r}\rho\left(\frac{x}{x_{r,j}}\right)\le\rho_0, \tag{1}$$

то можна розпізнати образ x як фонему $r\in X_r$. Такий процес розпізнавання мовних одиниць буде настільки об'єктивним (зокрема, нечутливім до діалектів фонації мовних одиниць), наскільки репрезентативно визначені кластери $\{x_{r,j}\}$ для фонетичного алфавіту X_r. Залежно від значення порогу ρ_0, результатом аналізу досліджуваного образу x за правилом (1) буде:

– його розпізнавання як однієї з фонем: $x=r$;

– його ототожнення з кількома фонемами: $x=\{r_i\}$, $r_i\in X_r$, $i\le J_r$;

– визнання його маргінальним відносно досліджуваного фонетичного алфавіту: $x\ne\forall r\in X_r$.

Для спрощення обчислень перетворимо правило (1) до вигляду

$$\rho_r(x)=x_r^*=x_{r,v}:\frac{1}{J_r}\sum_{j=1}^{J_r}\rho\left(\frac{x_{r,j}}{x_{r,v}}\right)=\min_{i\le J_r}\frac{1}{J_r}\sum_{j=1}^{J_r}\rho\left(\frac{x_{r,j}}{x_{r,i}}\right)\triangleq\rho_r^*\le\rho_0, \tag{2}$$

де в процесі розпізнавання образу x в межах кластера X_r розраховується одна відстань $\rho_r(x)\triangleq\rho\left(x/x_r^*\right)$ від нього до центру кластера x_r^*, координати якого визначають усереднений за діалектами еталон фонеми $r\in X_r$. Враховуючи правило (2), визначимо процедуру обчислювального фонетичного аналізу мовлення як порівняння емпіричної (виголошеної мовцем) $\{x_v^*\}$ і еталонної $\{x_r^*\}$ множин однакової потужності, попарно взяті елементи яких узагальнено характеризують відповідні фонеми досліджуваної мови як на боці мовця $v\in V$, так і на боці еталонного фонетичного корпусу $r\in R$. В цьому контексті *задачами дослідження* є:

– створити інформаційну технологію процесу обчислювального фонетичного аналізу мовлення з урахуванням діалектів та привнесеної мовцем специфіки фонації;

– сформулювати критерій оцінювання якості мовлення на основі запропонованої моделі з урахуванням збурювального впливу каналу поширення мовленнєвих сигналів в процесі фонації;

– довести адекватність і функціональність отриманих теоретичних результатів.

2. Метод обчислювального фонетичного аналізу мовлення з врахуванням діалекту та індивідуальності фонації

Зважаючи на положення теорії інформації, аргументуємо розв'язувальне правило (2) в контексті функціоналу відносної ентропії [35]—[37]

$$\rho(x)\triangleq\int\ldots\int\ln\frac{dP(x)}{dP_r(x)}P(dx), \tag{3}$$

де $P(x)$ — вибірковий розподіл імовірності досліджуваного (емпіричного) мовленнєвого сигналу x відносно еталонного розподілу імовірності $P_r(x)$, $r=\overline{1,R}$. Вважатимемо закон розподілу $P(x)$ нормальним: $P(x)=N(K_X)$, де K_X — вибіркова матриця автокореляції мовленнєвого сигналу x розмірністю $n\times n$. Врахуємо це у виразі (3): $\rho_r(x)=\frac{1}{2}\left(\mathrm{tr}\left(\frac{K_x}{K_r}\right)-\ln\left(\frac{K_x}{K_r}\right)-n\right)$, де $\mathrm{tr}(A)$ — операція знаходження сліду матриці A. Якщо ж передбачити, що досліджуваний мовленнєвий сигнал нормується на його ентропію, то останній вираз можна ще спростити до вигляду

$$\rho_r(x)=\frac{1}{2}\left(\mathrm{tr}\left(\frac{K_x}{K_r}\right)-n\right).$$

Представимо функціонал (3) у частотному просторі як оптимальну розв'язувальну статистику [35]. Для одного відліку досліджуваного мовленнєвого сигналу отримаємо

$$\rho_r(x) = \frac{1}{F}\left|\frac{1-\sum\limits_{m=1}^{p} a_r(m)e^{-j\pi m\frac{f}{F}}}{1-\sum\limits_{m=1}^{p} a_x(m)e^{-j\pi m\frac{f}{F}}}\right|^2, \quad (4)$$

де f — дискретне значення частоти для аналізованого відліку мовленнєвого сигналу, F — верхнє граничне значення частоти мовленнєвого сигналу, рівне половині частоти його дискретизації, $\{a_r(m)\}$ і $\{a_x(m)\}$ — вектори коефіцієнтів лінійної авторегресії порядку p, розраховані для еталонного сигналу x_r^* і емпіричного сигналу x, відповідно. Вираз у чисельнику (4) є амплітудно-частотною характеристикою вибілюючого фільтра, налаштованого на виділення ознак r-ї фонеми x_r^*, $r=\overline{1,R}$.

Вирази (2) і (4) дозволяють розрахувати кількісні характеристики, на основі яких можна обґрунтовано приймати рішення щодо належності досліджуваного образу x до кластера x_r^* відповідної фонеми $r \in X_r$. При цьому варіювати похибками такого процесу розпізнавання можна, змінюючи значення порогу ρ_0. За умови гаусівської апроксимації мовленнєвого сигналу, ймовірність похибки першого роду α для процесу розпізнавання фонем з урахуванням діалектів досліджуваної мови пропонується визначити в термінах χ^2- критерію з M ступенями свободи

$$\alpha \triangleq P\left\{\rho_r(x) \geq \rho_0\big|_{x\in X_r}\right\} = P\left\{\chi^2_M > M(1+\rho_0)\right\}, \quad (5)$$

де $P\{.\}$ — імовірність випадкової події, $M = \text{const}$. В загальному випадку значення константи M розраховується за виразом $M \approx L - p$, де p — порядок вибілюючого фільтра, $L = 2F\tau$ — параметр, значення якого залежить від кількості інтервалів стаціонарності τ, виділених у досліджуваному мовленнєвому сигналі x. Визначене за виразом (5) значення похибки α обернено пропорційне значенню порогу ρ_0. Наприклад, для заданого значення $\alpha = 0{,}1$, коли $\tau = 5$ мс, $F = 8$ кГц, $p = 20$, отримаємо $L = 80$ і, відповідно, $M = 60$. Користуючись таблицями χ^2-розподілу для рівня значимості $\beta = 1-\alpha = 1-0{,}99 = 0{,}01$ знайдемо значення квантилю $\chi^2_{M;\beta} = \chi^2_{60;0{,}01} = 88{,}38$, використовуючи яке розрахуємо значення порогу ρ_0: $\rho_0 = \chi^2_{M;\beta}/M - 1 = 0{,}473$.

Похибка другого роду β в контексті задачі обчислювального фонетичного аналізу мовлення з урахуванням діалектів репрезентує імовірність сплутування фонем r і v, $r, v \in X_r$, центри кластерів x_r^* і x_v^* яких достатньо близько розташовані в параметричному просторі: $\rho_{rv} \triangleq \rho_r(x)\big|_{x=x_v^*}$. Отже, значення похибки β обернено пропорційне значенню відстані ρ_{rv}. Аналіз результатів статистично репрезентативної кількості експериментів показав, що мінімальне значення ρ_{rv} для фонетичних алфавітів англійської мови $\{x_r^*\}$ знаходиться в діапазоні $[0{,}2;\, 0{,}3]$. Відповідно, за аналогією з (5), формалізуємо вираз для розрахунку похибки другого роду β процесу розпізнавання фонем з урахуванням діалектів досліджуваної мови:

$$\beta \triangleq P\left\{\rho_r(x) \geq \rho_0\big|_{x\in X_v}\right\} = P\left\{\chi^2_M < \frac{M(1+\rho_0)}{1+\rho_{rv}}\right\}. \quad (6)$$

Узагальнюючи втілені у вирази (5) і (6) міркування, далі для практичного використання обиратимемо значення порогу ρ_0 в розв'язувальному правилі (2), виходячи з виразу

$$p_0 = (1,\ldots,2)\min_{r,v}\rho_{rv}. \quad (7)$$

Значення порогу ρ_0, обчислене за виразом (7), забезпечує баланс між значеннями похибок першого і другого роду процесу розпізнавання фонем з фонетичного алфавіту X_r з урахуванням діалектів досліджуваної мови та варіативності процесу фонації. Втім, питання впливу індивідуальних особливостей артикуляції мовців на результат фонетичного аналізу мовлення потребує детальнішої аналітичної формалізації.

В контексті положень теорії інформації розглядатимемо мовця як джерело дискретних повідомлень *X*, визначених на множині еталонів мовних одиниць $\left\{x_r^*\right\}$. Вичерпно охарактеризувати таке джерело може кількість інформації, яка припадає на мовну одиницю, ним згенеровану. Якщо знехтувати впливом індивідуальних особливостей артикуляційного апарату мовця на процес фонації та вважати, що мовленнєве повідомлення передається в умовах відсутності шумів акустичного оточення, то шукану кількість інформації визначимо як ентропію Шеннона для дискретного джерела повідомлень [35]

$$H(X) \triangleq -\sum_{r=1}^{R} P\left(X = x_r^*\right)\log P\left(X = x_r^*\right) = -\sum_{r=1}^{R} p_r \log p_r \,. \tag{8}$$

Якщо згадати про нормування $\sum_{r=1}^{R} p_r = 1$, то, за умови рівноімовірної появи мовних одиниць $\forall r \le R:\ p_r = 1/R$, отримаємо спрощену форму виразу (8): $H(X) = \log R$. Втім, в реальних умовах нехтувати артикуляційно обумовленою варіативністю фонації неможна. Мовленнєвий сигнал на виході артикуляційного тракту мовця X' може суттєво відрізнятися від еталонного *X*: $X' \ne X$. Ця аксіома правильна навіть для окремих фонем, не кажучи про масивніші мовні одиниці. За таких умов адекватну математичну модель дискретного джерела мовних повідомлень слід створювати на основі визначення фонем за виразом (5), чітко кластеризованих у параметричному просторі: $q_r \triangleq P\left(X' \ne x_r^*\right)$, $r = \overline{1,R}$, і враховуючи імовірність появи абстрактної, $R+1$-ї, мовної ознаки, до якої відноситимемо випадки ненадійного розпізнавання сигналу X': $q_{R+1} \triangleq P\left(X' \ne x_r^*, \forall r \le R\right)$. Узагальнимо ці міркування для розв'язувального правила (2):

$$\begin{aligned} q_r &= \sum_{v=1}^{R} q_{rv} = \sum_{v=1}^{R} P\left(X' = x_r^*; X = x_v^*\right) = \sum_{v=1}^{R} P\left(X = x_v^*\right) P\left(X' = x_r^* \middle| X = x_v^*\right) = \\ &= P\left(X = x_r^*\right) P\left(X' = x_r^* \middle| X = x_r^*\right) = (1-\alpha)\, p_r; \\ q_{R+1} &= \sum_{v=1}^{R} P\left(X' = x_v^*; X = x_v^*\right) = \sum_{v=1}^{R} P\left(X = x_v^*\right) P\left(X' \ne x_v^* \middle| X = x_v^*\right) = \sum_{v=1}^{R} \alpha p_v = \alpha; \\ \sum_{r=1}^{R+1} q_r &= (1-\alpha) \sum_{r=1}^{R} p_r + \alpha \equiv 1, \end{aligned} \tag{9}$$

де $P\left(X' = x_r^* \middle| X = x_r^*\right) = 1-\alpha$ — умовна імовірність розпізнавання *r*-ї фонеми за умови нехтування варіативністю її фонації, привнесеною мовцем.

Зауважимо, що вираз (8) характеризує дискретне джерело мовленнєвих повідомлень, не враховуючи збурювальний вплив каналу їхнього поширення на кінцевий результат фонації. Врахуємо цю інформацію, використавши як базовий вираз [35]

$$I(X, X') \triangleq H(X) - H\left(X \middle| X'\right), \tag{10}$$

де X — екземпляр фонації еталону x_r^* фонеми $r \in X_r$, X' — екземпляр фонації цієї фонеми мовцем (емпіричний екземпляр), $H\left(X \middle| X'\right)$ — апостеріорна ентропія, яка характеризує розсіювання корисної інформації процесу фонації в наслідок збурювального впливу каналу її поширення. Врахувавши вирази (10) і (9), сформулюємо еквівалентне представлення виразу (10):

$$I(X,X') = H(X) + H(X') - H(XX') = H(X) - \sum_{r=1}^{R+1} q_r \log q_r + \sum_{v=1}^{R} \sum_{r=1}^{R+1} q_{rv} \log q_{rv} =$$

$$H(X) - (1-\alpha) \sum_{r=1}^{R} p_r \log\left(p_r (1-\alpha)\right) - \alpha \log \alpha + \sum_{r=1}^{R} q_{rr} \log q_{rr} + \alpha \sum_{v=1}^{R} p_v \log\left(p_v \alpha\right) =$$

$$H(X) + (1-\alpha) H(X) - (1-\alpha)\left((1-\alpha) - \alpha \log \alpha + (1-\alpha) \sum_{r=1}^{R} p_r \log\left(p_r (1-\alpha)\right)\right) - \tag{11}$$

$$- \alpha H(X) + \alpha \log \alpha = (1-\alpha) H(X).$$

Зважаючи на вираз (11), можна стверджувати, що апостеріорна ентропія розсіювання інформації при фонації мовленнєвого повідомлення $H(X \mid X')$ знаходиться в прямій пропорційній залежності з ентропією дискретного джерела мовних повідомлень (8)

$$H(X \mid X') = \alpha H(X). \tag{12}$$

Зважаючи на вираз (12), можна стверджувати, що за рівноімовірного розподілу фонем у фонетичному алфавіті мовця верхню межу розсіювання корисної інформації процесу фонації можна описати виразом

$$\sup H(X \mid X') = \alpha \log R. \tag{13}$$

Отриманий результат корелює з відомою нерівністю Фано [38] для довільних розв'язувальних правил

$$H(X \mid X') \le -\alpha \log \alpha - \beta \log \beta + \alpha \log(R-1). \tag{14}$$

Довести останнє твердження можна емпірично, порівнявши розраховані значення правих частин виразів (13), (14) для експериментальних даних, коли $0 \le \alpha \le 1$ і $1 < R < \infty$.

Отже, розв'язувальне правило (2), розв'язувальна статистика (4) і вирази (7)—(9) сукупно утворюють шукану інформаційну технологію процесу обчислювального фонетичного аналізу мовлення з урахуванням діалектів та внесеної мовцем специфіки фонації. Центральним елементом концепції є матриця інформаційного розузгодження $\|\rho_{r,v}\|$ розмірністю $R \times R$. Дані з матриці $\|\rho_{r,v}\|$ є основною для розрахунку порогу ρ_0 за виразом (7). За відомого значення ρ_0 на основі виразів (2) і (5) реалізується процедура сегментації фонетичного алфавіту $X_r = \{x_{r,j}\}$ на множину фонем, які з імовірністю $\beta = 1 - \alpha$ достовірно розпізнаються попри вищеописані збурювальні фактори, та множину фонем, які з імовірністю α розпізнаються недостатньо достовірно. Визначним фактором для такої сегментації є імовірність похибки першого роду, яка розраховується за виразом (5). Імовірність похибки другого роду (6) в цій процедурі враховується опосередковано, як обмеження при визначенні порогу ρ_0 за виразом (7). Застосування виразів (10), (9) дозволяє уточнити результат процедури сегментації, врахувавши варіативність фонації досліджуваних мовних одиниць, спричинену індивідуальними особливостями артикуляції конкретного мовця. Відмітимо, що хоча представлена інформаційна технологія формулювалася з опорою на фонеми, але положення, покладені в її основу, є несуперечними і для аналізу мовлення стосовно вмісту таких мовних одиниць як морфеми і лексеми. Засноване на запропонованій концепції (8)—(10) правило (11) дозволяє оцінити похибку першого роду (5) і персоніфіковану ентропію фонетичного словника (8) в результаті аналізу емпіричних даних, обсяг вибірки яких становить $N = 2FT$. Зазначимо, що статистично репрезентативний обсяг $N = 10^6$ у дослідженні за допомогою правила (11) фонетичного алфавіту з $R = 10^2$ елементів в результаті аналізу фонограм мовленнєвих сигналів з частотою дискретизації 16 кГц досягається вже при цензурованій тривалості останніх $T = 60$ с.

3. Метод детектування і виправлення помилок обчислювального фонетичного аналізу мовлення

Нехай $X_r = \{x_{r,j}\}$, $r = \overline{1,R}$, $j = \overline{1,M}$ — набір незалежних класифікованих вибірок виду

$x_{r,j}=\left[x_{r,j^{(1)}},x_{r,j^{(2)}},\ldots,x_{r,j^{(n)}}\right]^T$ потужністю n з $R\geq 2$ гаусівських розподілів $P_r=N\left(K_r\right)$ з нульовим математичне сподіванням і невідомою автокореляційною матрицею $K_r=\mathrm{E}_X\left(x_{r,j}x_{r,j}^T\right)$ розмірністю $n\times n$, де j — ідентифікатор циклу спостережень за *r*-м розподілом, T — операція транспонування, E_X — математичне сподівання набору вибірок X. Позначимо як X_0 вибірку виду X_r потужністю M_0 для досліджуваного сигналу з невідомим розподілом $P\left(X\right)\subset\left\{P_r\right\}$. Задача розпізнавання сигналу X_0 передбачає *R*-альтернативну перевірку статистичних гіпотез W_r щодо закону розподілу цього сигналу

$$W_r:P\left(X\right)=P_r,\ r=\overline{1,R}. \tag{15}$$

Нехай $R=2$, тобто перевіряються дві конкуруючі гіпотези $W_1:P\left(X\right)=P_1$ і $W_2:P\left(X\right)=P_2$ за невідомих апріорі автокореляційних матриць K_1 і K_2. Перевірку здійснюватимемо за допомогою асимптотичного мінімаксного критерію відношення правдоподібності [35]—[37] на основі даних з вибірки $X\left\{X_i\right\}$, $i=\overline{0,2}$. За таких умов гіпотеза W_1 буде визнана правильною, якщо виконається умова

$$W_1:\lambda_1\left(X\right)\triangleq\frac{\sup\limits_{K_1}\sup\limits_{K_2}\left(p\left(X\middle|W_1\right)\right)}{\sup\limits_{K_1}\sup\limits_{K_2}\left(p\left(X\middle|W_2\right)\right)}\equiv\frac{\sup\limits_{K_1}\left(p\left(X_0\middle|W_1\right)p\left(X_1\right)\right)\sup\limits_{K_2}\left(p\left(X_2\right)\right)}{\sup\limits_{K_1}\left(p\left(X_0\middle|W_2\right)p\left(X_1\right)\right)\sup\limits_{K_2}\left(p\left(X_1\right)\right)}>1, \tag{16}$$

де $p\left(X_0\middle|W_r\right)$ — функція правдоподібності сигналу X_0 за умови підтвердження гіпотези W_r; $p\left(X_r\right)$ — функція правдоподібності сигналу X_r. Користуючись відомим обчислювальним алгоритмом [38] за умови незалежності спостережень $X_r=\left\{x_{r,j}\right\}$ запишемо систему рівнянь

$$\begin{cases}\ln\left(p\left(X_0\middle|W_r\right)\right)=-\dfrac{M_0}{2}\left(\ln\left|K_r\right|+\mathrm{tr}\left(\dfrac{S_0}{K_r}\right)+n\ln\left(2\pi\right)\right),\\ \ln\left(p\left(X_r\right)\right)=-\dfrac{M_r}{2}\left(\ln\left|K_r\right|+\mathrm{tr}\left(\dfrac{S_r}{K_r}\right)+n\ln\left(2\pi\right)\right),\end{cases} \tag{17}$$

де $\left|K_r\right|$ — визначник матриці K_r; $S_r\triangleq\dfrac{1}{M_r}\sum\limits_{j=1}^{M_r}x_{r,j}x_{r,j}^T$ — оцінка максимальної правдоподібності для матриці K_r; визначена на вибірці X_r, $r=\overline{0,2}$. Опишемо на основі виразу (17) факт, що верхні межі для $\ln\left(p\left(X_r\right)\right)$ досягаються коли $K_r=S_r$

$$\sup_{K_r}\left(p\left(X_r\right)\right)=-\frac{M}{2}\left(\ln\left|S_r\right|+nc\right), \tag{18}$$

де $r=\left\{1,2\right\}$; $c=\ln\left(2\pi\right)+1$. Аналогічно отримаємо вираз для визначення верхніх границь для $\ln\left(p\left(X_0\middle|W_r\right)p\left(X_r\right)\right)$:

$$\begin{aligned}\sup_{K_r}\left(\ln p\left(X_0\middle|W_r\right)p\left(X_r\right)\right)&=-\frac{1}{2}\left(\left(M_0+M\right)\left(\ln\left|S_{0r}+n\ln\left(2\pi\right)\right|\right)+M_0\mathrm{tr}\left(\frac{S_0}{S_{0r}}\right)+M\mathrm{tr}\left(\frac{S_0}{S_{0r}}\right)\right)=\\&=-\frac{M_0+M}{2}\left(\ln\left|S_{0r}\right|+nc\right),\end{aligned} \tag{19}$$

де $r=\left\{1,2\right\}$, $S_{0r}=\dfrac{M_0}{M_0+M}\left(S_0+S_r\right)$ — оцінка максимальної правдоподібності для матриці K_r,

визначена на об'єднаній вибірці $X_{0r}+\{X_0,X_r\}$ потужністю M_0+M. Підставимо вирази (18), (19) у вираз (16) і отримаємо умову, за виконання якої гіпотеза W_1 буде визнана правильною,

$$W_1(X):\lambda_1(X)=\frac{1}{2}\left((M_0+M)\ln|S_{01}|-(M_0-M)\ln|S_{02}|-M\ln|S_1|+M\ln|S_2|\right)<0\equiv \\ \equiv M_0\gamma_{1,01}+M\gamma_{1,01}<M_0\gamma_{2,02}+M\gamma_{2,02}, \tag{20}$$

де $\gamma_{k,0r}=\frac{1}{2}\left(\mathrm{tr}\left(\frac{S_k}{S_{0r}}\right)-\ln|S_k|+\ln|S_{0r}|-n\right)\geq 0$ — значення функціоналу відносної ентропії між двома гіпотетичними розподілами імовірностей з автокореляційними матрицями S_k і S_{0r}.

Масштабуємо правило (20) для задачі розпізнавання сигналів виду (15) з довільною кількістю гіпотез $R\geq 2$

$$W_v(X):\left(M_0\gamma_{0,0r}+M\gamma_{r,0r}\right)\Big|_{r=v}=\min,\ r=\overline{1,R}. \tag{21}$$

Передбачаючи однорідність пари сигналів X_0 і X_r у складі вибірки X_{0r} та враховуючи, що $\gamma_{0,0r}\leq\gamma_{0,r}$, $\gamma_{r,0r}\leq\gamma_{r,0}$ і $M=M_0$, представимо правило (21) у вигляді

$$W_v(X):\lambda_v(X)\triangleq\left(M_0\gamma_{0,r}+M\gamma_{r,0}\right)\Big|_{r=v}\triangleq\gamma_{0,r}+\gamma_{r,0}\big|r=v=\min,\ r=\overline{1,R}, \tag{22}$$

де розв'язувальні статистики функціоналу відносної ентропії

$$\gamma_{0,r}=\frac{1}{2}\left(\mathrm{tr}\left(\frac{S_0}{S_r}\right)-\ln|S_0|+\ln|S_r|-n\right); \tag{23}$$

$$\gamma_{r,0}=\frac{1}{2}\left(\mathrm{tr}\left(\frac{S_r}{S_0}\right)-\ln|S_r|+\ln|S_0|-n\right) \tag{24}$$

визначаються на R-множині пар вибіркових розподілів $N(S_0)$, $N(S_r)$, $r=\overline{1,R}$. Альтернативою виразам (23), (24) може слугувати врахування у правилі (22) принципу мінімуму величини інформаційного неспрямованого розузгодження $J(X_0,X_r)\triangleq\frac{1}{2}\left(\gamma_{0,r}+\gamma_{r,0}\right)$ між стохастичними сигналами X_0 і X_r, $r=\overline{1,R}$:

$$\tilde{W}_v(X):\tilde{\lambda}_v(X)\triangleq\gamma_{0,r}\big|_{r=v}=\min, \tag{25}$$

де розв'язувальна статистика $\gamma_{0,r}$ визначається за виразом (23).

Вираз (25) є частинним випадком критерію (22) за умови, що за необмеженого зростання об'ємів навчальних вибірок M другий доданок у виразі (21) асимптотично прямує до нуля: $\gamma_{r,0r}\to\gamma_{r,r}=0\forall r\leq R$. Отже, перехід від правила (22) до (25) є доцільним за умови спостереження значної асиметрії значень розв'язувальних статистик (23), (24).

Імовірність $\alpha_{v\to r}\triangleq P\left(W_r(X)\middle|W_r\right)$ сплутування v-го і r-го сигналів, $v\neq r\leq R$, з використовуваної бази апріорних даних $\{X_r\}$ у формалізмі розв'язувального правила (22) можна описати виразом

$$\alpha_{v\to r}=P\left\{\gamma_{0,v}+\gamma_{v,0}>\gamma_{0,r}+\gamma_{r,0}\middle|W_v\right\}=P\left\{2\gamma_{v,v}>\gamma_{v,r}+\gamma_{r,v}\right\}. \tag{26}$$

Якщо врахувати, що емпіричний сигнал перед розпізнаванням нормується на величину його питомої ентропії, то виконується система асимптотичних рівнянь

$$\forall r\leq R:\frac{1}{n}\ln|S_r|=\frac{1}{n}\ln|S_0|\underset{n\to\infty}{=}\ln\sigma_0^2=\mathrm{const}.$$

Врахуємо цей факт представивши розв'язувальну статистику $\gamma_{v,r}$ у формалізмі χ^2-розподілу з

$K \le M$ ступенями свободи: $\gamma_{v,r} = \frac{1}{2} n \left(\frac{\sigma_{r,v}^2 \sigma_0^2 \chi_{r,v}^2(K)}{M} - 1 \right)$, де $\sigma_{r,v}^2 \triangleq \frac{\sigma_0^2}{n} \lim\limits_{n\to\infty} \left(M\mathrm{tr}\left(\frac{S_v}{S_r} \right) \right)$ — допоміжна змінна. Підставимо отриманий вираз для статистики $\gamma_{v,r}$ у вираз (26)

$$\alpha_{v\to r} = P\left\{ \sigma_0^2 \chi_{v,v}^2 > \frac{1}{2}\sigma_{r,v}^2 \chi_{r,v}^2 + \frac{1}{2}\sigma_{v,r}^2 \chi_{v,r}^2 \right\} = P\left\{ 2\chi_{v,v}^2 > \left(1+\rho_{r,v}\right)\chi_{r,v}^2 + \left(1+\rho_{v,r}\right)\chi_{v,r}^2 \right\}, \qquad (27)$$

де $\rho_{r,v} \triangleq \frac{\sigma_{r,v}^2}{\sigma_0^2} - 1$, $\rho_{v,r} \triangleq \frac{\sigma_{v,r}^2}{\sigma_0^2} - 1$ — питоме значення інформаційного розузгодження для досліджуваної пари розподілів $N(S_0)$ і $N(S_r)$, за умови $n \to \infty$, $\sigma_{v,r}^2 \triangleq \frac{\sigma_0^2}{n} \lim\limits_{n\to\infty} \left(M\mathrm{tr}\left(\frac{S_r}{S_v} \right) \right)$ — допоміжна змінна, однотипна з $\sigma_{r,v}^2$. Якщо припустити взаємну некорельованість трьох χ^2-розподілів у виразі (27), то вираз (26) для розрахунку імовірності сплутування $\alpha_{v\to r}$ можна представити у вигляді $\alpha_{v\to r} = P\left\{ \frac{1}{2}\left(\left(1+\rho_{r,v}\right)F_{r,v}(1,K) + \left(1+\rho_{v,r}\right)F_{v,r}(1,K) \right) < 1 \right\}$, де $F_{r,v}(1,K) = \frac{\chi_{r,v}^2}{\chi_{v,v}^2}$, $F_{v,r}(1,K) = \frac{\chi_{v,r}^2}{\chi_{v,v}^2}$ — статистики *F*-розподілу з $(1,K)$ ступенями свободи. Відповідну, верхню межу імовірності сплутування $\alpha_{v\to r}$ можна оцінити за виразом

$$\alpha_{v\to r} \le P\left\{ \frac{1}{2}\max\left[\left(1+\rho_{v.r}\right)F_{v,r}(1,K); \left(1+\rho_{r.v}\right)F_{r,v}(1,K) \right] \right\} = P\left\{ F(1,K) < \frac{2}{\max\left[\left(1+\rho_{v.r}\right); \left(1+\rho_{r.v}\right) \right]} \right\} =$$

$$= P\left\{ F(K,1) \ge \frac{1}{2}\max\left[\left(1+\rho_{v.r}\right); \left(1+\rho_{r.v}\right) \right] \right\} = 1 - \Phi_{K,1}\left\{ \max\left[\left(1+\rho_{v.r}\right); \left(1+\rho_{r.v}\right) \right] \right\}, \qquad (28)$$

де $F(1,K) = \max\left[F_{r,v}(1,K); F_{v,r}(1,K) \right]$, $F(K,1) = \frac{1}{F(1,K)}$ — статистики *F*-розподілу з $(1,K)$ і $(K,1)$ ступенями свободи, відповідно; $\Phi_{K,1}$ — інтегральна функція *F*-розподілу з $(K,1)$ ступенями свободи. З виразу (28) випливає факт існування суттєво нерівноцінних розподілів статистики $\chi_{v,v}^2$ і пари статистик $\chi_{r,v}^2$, $\chi_{v,r}^2$ за умови, що $r \ne v$. Отже, вираз (28) теоретично доводить коректність виразів (23) і (24) щодо асиметрії значення інформаційного розузгодження, що враховується у розв'язувальному правилі (22). Це означає, що при виконанні умови $\exists v, r \le R : \rho_{v,r} >> \rho_{r,v}$ доцільніше для прийняття рішень щодо розпізнавання мовних одиниць у мовленнєвому сигналі, параметризованому в парадигмі концепції (8)—(10), застосовувати розв'язувальне правило (22), а не (25). Цей тезис буде перевірено у експериментальній частині статті.

Припустимо, що під час розпізнавання досліджуваного сигналу за допомогою розв'язувального правила (25) вердикт був помилково винесений на користь гіпотези $W_\mu(X)$, а не гіпотези $W_v(X)$. Також припустимо, що під час розпізнавання цього ж сигналу за допомогою розв'язувального правила (22) вердикт був винесений на користь гіпотези $W_v(X)$. Висловлені припущення передбачають, що згідно з виразами (25), (26) одночасно виконувались нерівності $\gamma_{v,v} \ge \gamma_{v,\mu}$ і $2\gamma_{v,v} \ge \gamma_{v,\mu} + \gamma_{\mu,v}$, що можливо лише за виконання умови $\gamma_{\mu,v} >> \gamma_{v,\mu}$. Отже, аналітичною ознакою помилковості рішення, прийнятого за правилом (25) відносно аналізованої вибірки X_0, може слугувати нерівність виду $\overline{W}_\mu(X) : \gamma_{\mu,0} >> \gamma_{0,\mu}$, або:

$$\overline{W}_\mu(X) : \frac{1+\tilde{\gamma}_{\mu,0}}{1+\tilde{\gamma}_{0,\mu}} \ge c_0, \qquad (29)$$

де $\tilde{\gamma}_{0,\mu} = \frac{2\gamma_{0,\mu}}{n}$; $\tilde{\gamma}_{0,\mu} = \frac{2\gamma_{0,\mu}}{n}$ — питомі значення розв'язувальних статистик (23), (24), відповідно;

c_0 — значення порогу (мінімальне значення коефіцієнту асиметрії значень (23) і (24) у правилі (22)), встановлюване залежно від максимально допустимої похибки $\beta \triangleq P\left\{\dfrac{1+\tilde{\gamma}_{\mu,0}}{1+\tilde{\gamma}_{0,\mu}} \geq c_0 \middle| W_\mu\right\} \leq \beta_0$.

Повторюючи міркування, які супроводжували перехід від виразу (26) до виразу (28), перепишемо щойно наведений вираз для визначення імовірності β у термінах *F*-розподілу:

$$\pi_{v\to\mu} \triangleq P\left\{\frac{1+\tilde{\gamma}_{\mu,0}}{1+\tilde{\gamma}_{0,\mu}} \geq c_0 \middle| W_v\right\} = P\left\{\frac{1+\tilde{\gamma}_{v,\mu}}{1+\tilde{\gamma}_{\mu,v}} \geq \frac{1}{c_0}\right\} = P\left\{\frac{1+\tilde{\gamma}_{\mu,\mu}}{1+\tilde{\gamma}_{0,0}} \geq c_0\right\} = P\left\{\frac{\chi^2_{\mu,\mu}(K)}{\chi^2_{0,0}(K)} \geq c_0\right\} = 1-\Phi_{K,K}(c_0) \leq \beta_0. \quad (30)$$

Проаналізувавши вираз (30) отримаємо рівняння $\min c_0 = f_{K,K}(1-\beta_0)$, де $f_{K,K}(1-\beta_0)$ — квантіль *F*-розподілу з (K,K) ступенями свободи і рівнем значимості $1-\beta_0$. Наприклад, якщо $K=100$ і $\beta_0 = 0{,}01$, то з таблиць *F*-розподілу маємо $c_0 \geq f_{100,100}(0{,}99) = 1{,}59$.

Отже, правило (29) дозволяє оцінити імовірність події визнання маргінальним правильного результату процедури розпізнавання *v*-ї фонеми за допомогою розв'язувального правила (25). Стохастична оцінка такої події характеризується виразом

$$\pi_{v\to\mu} = P\left\{\frac{\chi^2_{v,\mu}(1)}{\chi^2_{\mu,v}(1)} \geq \frac{1+\rho_{v,\mu}}{c_0\left(1+\rho_{\mu,v}\right)}\right\} = 1-\Phi_{1,1}\left(\frac{1+\rho_{v,\mu}}{c_0\left(1+\rho_{\mu,v}\right)}\right) \quad (31)$$

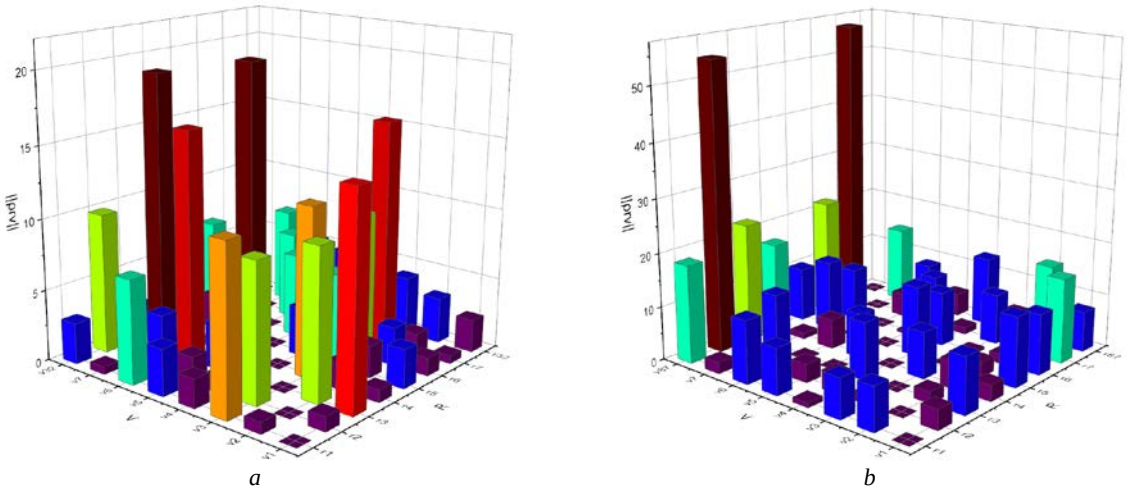
і визначається результатом порівняння елементів $\rho_{r,v}$ і $\rho_{v,r}$, що протистоять в матриці $\left\|\rho_{r,v}\right\|$.

Постановка і результати експерименту

Використаємо засноване на концепції (8)—(10) правило (11) для оцінювання фонетичної насиченості мовлення персон у складі колективу з 30 осіб. Персональний склад цього колективу сформовано збалансованим. При цьому бралися до уваги такі критерії як вік (три вікові групи: 20—29, 30—39, 40—49 років), стать (чоловіча, жіноча), освіта — вища, рідна мова — українська, рівень володіння англійською мовою згідно з CEFR – B2. Кожна з персон один раз прослухала фонограму з записом виголошеного Google перекладачем англомовного публіцистичного тексту обсягом 1800 символів. Згодом кожна персона переказала почутий текст під запис у персоніфіковану цифрову фонограму тривалістю 3 хвилини. Фонація переказу відбувалася в одному темпі і тембрі та з чіткою фіксацією мовних одиниць. Запис фонограм здійснювався за допомогою мікрофону AKG P420 без підсилювача, з'єднаного з інтегрованою в персональний комп'ютер звуковою картою Creative Audigy Rx з частотою дискретизації 16 кГц. Кожна фонограма зберігалася у форматі .wav. Для подальшого аналізу фонограми сегментувалися на відрізки тривалістю $\tau = 5$ мс ($L = 80$ відліків). На основі аналізу відповідних фонограм переказів для кожної персони сформовано індивідуальні фонетичні алфавіти $\{X_r\}$, для яких за виразом (2) визначені центри кластерів фонем $\{x_r^*\}$. Для кожної персони утворювалися два варіанти індивідуального фонетичного алфавіту — з жорсткими та м'якими умовами формування. Ці умови задавалися рівнем розузгодження $\Delta\rho = \{0{,}5; 1{,}0\}$ для однойменних фонем та їх мінімальною тривалістю $\Delta L = \tau$, $\Delta L = \{8L; 4L\}$, $\tau = \{40; 20\}$. Необхідні для розрахунку матриці інформаційного розузгодження $\left\|\rho_{r,v}\right\|$ значення коефіцієнтів авторегресії $\{a_r(m)\}$, $\{a_v(m)\}$ визначалися за допомогою рекурентної процедури Берга–Левінсона із однозначно визначеним порядком моделей $p = 20$.

На рис. 1 візуалізовані фрагменти підсумкових матриць $\left\|\rho_{r,v}\right\|$ для персони №1, розраховані з обраним жорстким (рис. 1*a*) і м'яким (рисунок 1*b*) наборами умов формування. Потужності фонетичних алфавітів склали $R^1_{hard} = 32$ та $R^1_{soft} = 87$ мовних одиниць, відповідно. Мінімальне значення інформаційного розузгодження між фонемами $\Delta\rho^{R^1_{hard}}_{rv} = 0{,}324$.

Відповідно до розв'язувального правила (2) з урахуванням виразу (7), визначено поріг $\rho_0 = 0{,}324$.



a *b*

Рис. 1. Візуалізація фрагментів матриць інформаційного розузгодження $\|\rho_{r,v}\|$ для персони № 1, розраховані з вибраними наборами умов формування: *а* — жорстким; *б* — м'яким

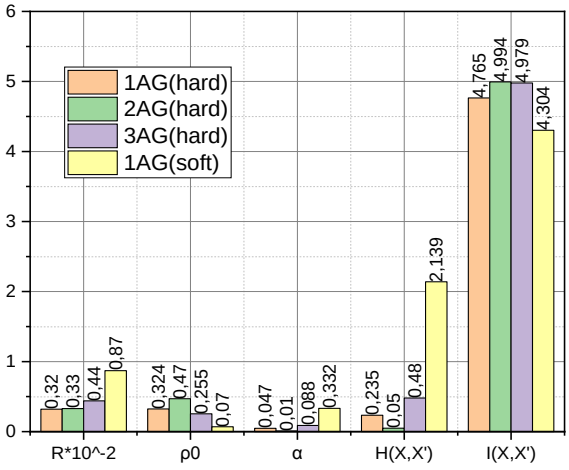


Рис. 2. Оцінювання фонетично насиченості персоніфікованого мовлення

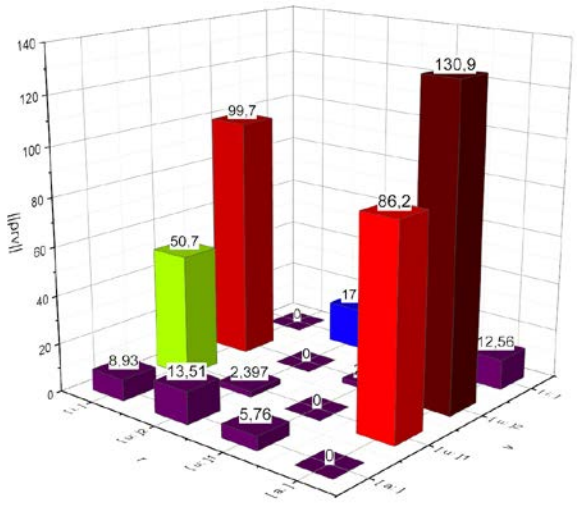


Рис. 3. Візуалізація матриці $\|\rho_{r,v}\|$, розрахованої для екземплярів фонації чотирьох фонем персоною № 1

Користуючись таблицями χ^2-розподілу для кількості ступенів свободи $M = 60$ визначено імовірність похибки першого роду $\alpha = 0,047$. Тоді, згідно з виразом (13), верхня межа розсіювання корисної інформації процесу фонації для персони № 1 дорівнює $\sup H\left(X \middle| X'\right) = \alpha \log R = 0,235$, а верхня межа фонетичної насиченості мовлення для персони № 1, згідно з виразом (11), дорівнює

$$\sup I\left(X \middle| X'\right) = (1-\alpha)\log R = 4,765.$$

Аналогічні розрахунки здійснені для решти персон з колективу дослідників. Для наочності представлення ці результати усереднені для кожної з трьох вікових груп і показані на рис. 2. Окрім того, для порівняння, для персон з першої вікової групи розраховані матриці розузгодження з вибраним м'яким набором умов формування і виконанням всіх інших вищеописаних обчислювальних операцій. Ці результати, поіменовані як «1^{AG}_{soft}», також відображені на рис. 2.

Дослідимо емпірично функціональність узагальнених розв'язувальними правилами (22) і (25) концепцій прийняття рішень в задачі обчислювального фонетичного аналізу мовлення (статистична класифікація без учителя в парадигмі концепції (8)—(10)). Емпіричним матеріалом для досліджень вибрані дві фонограми із записом єдиного за змістом мовного матеріалу від персони № 1. Фонограми представлялася вибірками X_0, X_r однакової потужності $M = 120$. Спочатку розрахували матрицю інформаційного розузгодження $\|\rho_{r,v}\|$ для чотирьох голосних фонем персони № 1. Вміст матриці наочно показаний на рис. 3. Алофони $[\mathrm{u}:]_1$ і $[\mathrm{u}:]_2$ репрезентують специфічні для персони № 1 діалекти виголошення фонеми $[\mathrm{u}:]_1$.

Подальше використання даних, поданих на рис. 3, продемонструємо на прикладі. Розглянемо

дані з матриці $\|\rho_{r,v}\|$ для пари фонем $\left([\text{a}:],[\text{u}:]_1\right)$. Ці дані характеризують ситуацію, коли фонема $[\text{a}:]$ розпізнається як фонема $[\text{u}:]_1$. З рис. 3 видно, що $\rho\left([\text{a}:],[\text{u}:]_1\right)=5{,}76$. Кількість ступенів свободи для *F*-розподілу у виразі (28) беремо рівною $K=M-p=100$. Якщо рішення щодо результату розпізнавання фонеми приймається згідно з розв'язувальним правилом (25), то для $(K,1)=(100,1)$ маємо

$$\tilde{\alpha}_{v\to r}=1-\Phi_{K,1}\{1+\rho_{r,v}\}=1-\Phi_{100,1}\{5{,}76\}\approx 0{,}3.$$

Якщо ж рішення щодо результату розпізнавання фонеми приймається згідно розв'язувального правила (22), то маємо $\max\left[(1+\rho_{vr});(1+\rho_{rv})\right]=\max[87{,}2;6{,}76]=6{,}76$. Відповідно до виразу (28) маємо $\alpha_{v\to r}\le 1-\Phi_{100,1}(5{,}76)\approx 0{,}12$. Отже, імовірність помилки під час прийняття рішень щодо результату фонетичного аналізу на прикладі фонем $[\text{a}:]$ і $[\text{u}:]_1$ за допомогою розв'язувального правила (22) у порівнянні з правилом (25) є майже в три рази меншою. Розрахунки, аналогічні наведеним, здійснені для всіх пар різнойменних фонем з рис. 3. Для всіх реалізацій правило (22) дозволило отримати меншу оцінку імовірності сплутування $\alpha_{v\to r}$ порівняно з правилом (25).

Завершимо цей етап досліджень, розрахувавши за виразом (31) імовірність події визнання маргінальним правильного результату процедури розпізнавання *v*-ї фонеми за допомогою розв'язувального правила (25): $\pi_{r\to v}=1-\Phi_{1,1}\left\{\dfrac{1+86{,}2}{1{,}59(1+5{,}76)}\right\}=1-\Phi_{1,1}(8{,}11)\approx 0{,}21$.

Можна констатувати, що чим більшою є асиметричність між елементами матриці $\|\rho_{rv}\|$, що протистоять, тим більшим є значення імовірності $\pi_{r\to v}$.

Висновки

Вперше запропоновано інформаційну технологію обчислювального фонетичного аналізу мовлення. В концепції, на відміну від існуючих, проблема багатокритеріальності процесу когнітивного сприйняття мовлення людиною строго формально представлена в теоретико-аналітичному апараті теорії інформації, теорії розпізнавання образів і акустичної теорії мовотворення. Отримана інформаційна технологія дозволяє точно визначати фонетичний алфавіт персони з урахуванням властивого їй діалекту мовлення та індивідуальних особливостей фонації, а також детектувати і виправляти помилки розпізнавання мовних одиниць та надійно оцінювати фонетичну насиченість мовлення.

Запропонована інформаційна технологія представлена розв'язувальним правилом (2), розв'язувальною статистикою (4) і виразами (7)—(9). Центральним елементом концепції є матриця інформаційного розузгодження $\|\rho_{r,v}\|$ мовних одиниць персоніфікованого фонетичного алфавіту мовця. Матриця $\|\rho_{r,v}\|$ є основною для розрахунку порогу ρ_0 для реалізації обчислювального фонетичного аналізу за виразом (7). За відомого значення ρ_0 на основі виразів (2) і (5) реалізується процедура сегментації досліджуваного фонетичного алфавіту на множину фонем, які з імовірністю $\beta=1-\alpha$ достовірно розпізнаються попри збурювальні фактори, та множину фонем, які з імовірністю α розпізнаються недостатньо достовірно. Застосування виразів (10), (9) дозволяє уточнити результат процедури сегментації, врахувавши варіативність фонації досліджуваних мовних одиниць, внесену індивідуальними особливостями артикуляції конкретного мовця.

Дослідження результатів обчислювального фонетичного аналізу на основі функціоналу відносної ентропії дозволило теоретично обґрунтувати два підходи, здатних до детектування помилок процесу розпізнавання мовних одиниць (15), на основі розв'язувальних правил (22) і (25). Зазначимо формалізовану виразом (31) можливість оцінювання надійності прийнятого на основі правила (25) рішення. Якщо рішення $W_{\mu}(X)$ згідно з виразом (29) буде визнано скомпрометованим, то за допомогою обчислювальної процедури за схемою (15) можна віднайти помилково визнані ненадійними результати фонетичного аналізу і реабілітувати їх. Закладений в запропоновану інфор-

маційну технологію потенціал і подані після рис. 3 експериментальні результати доводять її перевагу над такими байєсовськими інформаційними технологіями прийняття рішень із застосуванням метрики розузгодження евклідового типу як максимуму правдоподібності та ідеального спостерігача. Проведений у заснованій на запропонованій концепції метриці $\left\{H\left(X\left|X'\right.\right);I\left(X\left|X'\right.\right)\right\}$ аналіз досліджуваного мовленнєвого сигналу дозволяє надійно встановити фонетичну насиченість мовлення, яка об'єктивно охарактеризує середовище поширення мовленнєвого сигналу і його джерело.

Подальші дослідження планується спрямовувати на аналіз функції $I\left(X,X'\right)=f\left(R,\Delta\rho,\Delta L\right)$, екстремум якої потенційно може вказати на елементи персоніфікованого фонетичного алфавіту, в яких індивідуальність та інформативність мовлення відповідної персони проявляються найбільше. Автори сподіваються, що результати такого дослідження сприятимуть зростанню практичної цінності запропонованої системи моделей для прицезійного фонетичного аналізу мовлення [39].

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

[1] A. Mandal, Kumar Prasanna, and P. K. R. Mitra, "Recent developments in spoken term detection: a survey," *Int. J. Speech Technol* 17, pp. 183-198, 2014. https://doi.org/10.1007/s10772-013-9217-1 .

[2] C. China Bhanja, M. A. Laskar, and R. H. Laskar, "Modelling multi-level prosody and spectral features using deep neural network for an automatic tonal and non-tonal pre-classification-based Indian language identification system," *Lang Resources & Evaluation*, 2021. https://doi.org/10.1007/s10579-020-09527-z .

[3] S. S. Agrawal, A. Jain, and S. Sinha, "Analysis and modeling of acoustic information for automatic dialect classification," *Int. J. Speech Technol* 19, pp. 593-609, 2016. https://doi.org/10.1007/s10772-016-9351-7 .

[4] S. Gholamdokht Firooz, S. Reza, and Y. Shekofteh, "Spoken language recognition using a new conditional cascade method to combine acoustic and phonetic results," *Int. J. Speech Technol* 21, pp. 649-657, 2018. https://doi.org/10.1007/s10772-018-9526-5 .

[5] D. Duran, et al. "A Computational Model of Unsupervised Speech Segmentation for Correspondence Learning," *Res on Lang and Comput* , no. 8, pp. 133-168, 2010. https://doi.org/10.1007/s11168-011-9075-4 .

[6] D. Mirman, "Mechanisms of Semantic Ambiguity Resolution: Insights from Speech Perception," *Res on Lang and Comput* no.6, pp. 293-309, 2008. https://doi.org/10.1007/s11168-008-9055-5 .

[7] E. M. Bender, *et al.* "Grammar Customization," *Res on Lang and Comput* no. 8, pp. 23-72, 2010. https://doi.org/10.1007/s11168-010-9070-1 .

[8] M. Dickinson, "On Morphological Analysis for Learner Language, Focusing on Russian," *Res on Lang and Comput* no. 8, pp. 273, 2010. https://doi.org/10.1007/s11168-011-9079-0 .

[9] , S. Moran, E. Grossman, and A. Verkerk, "Investigating diachronic trends in phonological inventories using BDPROTO," *Lang Resources & Evaluation* no. 55, pp. 79-103, 2021. https://doi.org/10.1007/s10579-019-09483-3 .

[10] C. van Bael, H. van den Heuvel, and H. Strik, "Validation of phonetic transcriptions in the context of automatic speech recognition," *Lang Resources & Evaluation* no. 41, pp. 129-146, 2007. https://doi.org/10.1007/s10579-007-9033-9 .

[11] N. B. Chittaragi, S. G. Koolagudi, "Automatic dialect identification system for Kannada language using single and ensemble SVM algorithms," *Lang Resources & Evaluation* no. 54, pp. 553-585, 2020. https://doi.org/10.1007/s10579-019-09481-5

[12] L. Pearl, S. Goldwater, and M. Steyvers, "Online Learning Mechanisms for Bayesian Models of Word Segmentation," *Res on Lang and Comput* no. 8, pp. 107-132, 2010. https://doi.org/10.1007/s11168-011-9074-5 .

[13] M. Kurimo, *et al.* "Modeling under-resourced languages for speech recognition," *Lang Resources & Evaluation* no. 51, pp. 961-987, 2017. https://doi.org/10.1007/s10579-016-9336-9 .

[14] A. Masmoudi, *et al.* "Automatic speech recognition system for Tunisian dialect," *Lang Resources & Evaluation* no. 52, pp. 249-267, 2018. https://doi.org/10.1007/s10579-017-9402-y .

[15] W. Elvira-García, *et al.* "A tool for automatic transcription of intonation: Eti_ToBI a ToBI transcriber for Spanish and Catalan. *Lang Resources & Evaluation*," no. 50, pp. 767-792, 2016. https://doi.org/10.1007/s10579-015-9320-9 .

[16] H. Strik, M. Hulsbosch, and C. Cucchiarini, "Analyzing and identifying multiword expressions in spoken language," *Lang Resources & Evaluation* no. 44, pp. 41-58, 2010. https://doi.org/10.1007/s10579-009-9095-y .

[17] M. Aissiou, "A genetic model for acoustic and phonetic decoding of standard arabic vowels in continuous speech," *Int J Speech Technol* no. 23, pp. 425-434, 2020. https://doi.org/10.1007/s10772-020-09694-y .

[18] C. Santhosh Kumar, V. P. Mohandas, "Robust features for multilingual acoustic modeling," *Int J Speech Technol* no. 14, pp. 147-155, 2011. https://doi.org/10.1007/s10772-011-9092-6 .

[19] N. B. Chittaragi, S. G. Koolagudi, "Acoustic-phonetic feature based Kannada dialect identification from vowel sounds," *Int J Speech Technol* no. 22, pp. 1099-1113, 2019. https://doi.org/10.1007/s10772-019-09646-1 .

[20] N. T. Kleynhans, E. Barnard, "Efficient data selection for ASR," *Lang Resources & Evaluation* no. 49, pp. 327-353, 2015. https://doi.org/10.1007/s10579-014-9285-0 .

[21] C. Clavel, *et al.* "Spontaneous speech and opinion detection: mining call-centre transcripts," *Lang Resources & Evaluation* no. 47, pp. 1089-1125, 2013. https://doi.org/10.1007/s10579-013-9224-5 .

[22] F. Anitha Florence Vinola, G. Padma, "A probabilistic stochastic model for analysis on the epileptic syndrome using speech synthesis and state space representation," *Int J Speech Technol*, no. 23, pp. 35-360, 2020. https://doi.org/10.1007/s10772-020-09702-1 .

[23] M. Mehrabani, J. H. L. Hansen, "Automatic analysis of dialect/language sets," *Int J Speech Technol* no. 18, pp. 277-286, 2015. https://doi.org/10.1007/s10772-014-9268-y .

[24] X. Ma, “Evocation: analyzing and propagating a semantic link based on free word association,” *Lang Resources & Evaluation* no. 47, pp. 819-837, 2013. https://doi.org/10.1007/s10579-013-9219-2 .

[25] J. Chaki “Pattern analysis based acoustic signal processing: a survey of the state-of-art,” *Int J Speech Technol*, 2020. https://doi.org/10.1007/s10772-020-09681-3 .

[26] K. B. Bhangale, and K. Mohanaprasad, “A review on speech processing using machine learning paradigm,” *Int J Speech Technol* no. 24, pp. 367-388, 2021. https://doi.org/10.1007/s10772-021-09808-0 .

[27] P. Verma, and P. K. Das, “i-Vectors in speech processing applications: a survey,” *Int J Speech Technol*, no. 8, pp. 529-546, 2015. https://doi.org/10.1007/s10772-015-9295-3 .

[28] T. Drugman, and N. Dutoit, “The Deterministic Plus Stochastic Model of the Residual Signal and Its Applications,” *IEEE Transactions on Audio, Speech, and Language Processing*, 20, no. 3, pp. 968-981, 2012. https://doi.org/1109/TASL.2011.2169787 .

[29] X. Chen, and C. Bao, “*Phoneme-Unit-Specific Time-Delay Neural Network for Speaker Verification*,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, no. 29, pp. 1243-1255, 2021. https://doi.org/10.1109/TASLP.2021.3065202 .

[30] I. Omer, M. Zampieri, and M. Oakes, “Phonetic differences for dialect clustering,” *9th International Conference on Information and Communication Systems (ICICS)*, 2018, pp. 145-150. https://doi.org/10.1109/IACS.2018.8355457 .

[31] H. Van hamme, “Phonetic analysis of a computational model for vocabulary acquisition from auditory inputs,” *IEEE International Conference on Development and Learning (ICDL)*, 2011, pp. 1-6. https://doi.org/10.1109/DEVLRN.2011.6037365 .

[32] Z. Wang, C. Liu, H. Wang, Y. Hu, and L. Dai, “Phonetic clustering based confidence measure for embedded speech recognition,” in *7th International Symposium on Chinese Spoken Language Processing*, 2010, pp. 186-189. https://doi.org/10.1109/ISCSLP.2010.5684914 .

[33] P. Kannadaguli, and V. Bhat, “A comparison of Bayesian multivariate modeling and hidden Markov modeling (HMM) based approaches for automatic phoneme recognition in kannada,” *Recent and Emerging trends in Computer and Computational Sciences (RETCOMP)*, 2015, pp. 1-5. https://doi.org/10.1109/RETCOMP.2015.7090795 .

[34] F. A. A. Laleye, E. C. Ezin, and C. Motamed, “Automatic Text-Independent Syllable Segmentation Using Singularity Exponents and Rényi Entropy,” *J Sign Process Syst* no. 88, pp. 439-451, 2017. https://doi.org/10.1007/s11265-016-1183-9 .

[35] J. Kang, *et al.* “Lattice Based Transcription Loss for End-to-End Speech Recognition,” *J Sign Process Syst* no. 90, pp. 1013-1023, 2018. https://doi.org/10.1007/s11265-017-1292-0 .

[36] Y. Qian, *et al.* “Spoken Language Understanding of Human-Machine Conversations for Language Learning Applications,” *J Sign Process Syst* no. 92, pp. 805-817, 2020. https://doi.org/10.1007/s11265-019-01484-3 .

[37] Y. Cui, *et al.* “Simultaneous Predictive Gaussian Classifiers, *”J. Classif* no. 33**,** pp. 73-102, 2016. https://doi.org/10.1007/s00357-016-9197-3 .

[38] O. Bisikalo, O. Boivan, N. Khairova, O. Kovtun, and V. Kovtun, “Precision Automated Phonetic Analysis of Speech Signals for Information Technology of Text-dependent Authentication of a Person by Voice, ” *CEUR Workshop Proceedings*, no. 2853, pp. 276-288, 2021. urn:nbn:de:0074-2853-7 .

Рекомендована кафедрою автоматизації та інтелектуальних інформаційних технологій ВНТУ

Стаття надійшла до редакції 10.02.2022

Данильчук Оксана Миколаївна — канд. пед. наук, доцент, доцент кафедри прикладної математики, e-mail: oksanadommod@ukr.net .

Донецький національний університет імені Василя Стуса, Вінниця;

Ковтун В’ячеслав Васильович — д-р техн. наук, доцент, професор кафедри комп’ютерних систем управління, e-mail: kovtun_v_v@vntu.edu.ua ;

Никитенко Олена Дмитрівна — канд. техн. наук, доцент, доцент кафедри комп’ютерних систем управління, e-mail: lena260784@gmail.com ;

Нестюк Юлія Юріївна — студентка факультету інтелектуальних інформаційних технологій та автоматизації, e-mail: yynestiuk@gmail.com ;

Присяжнюк Василь Васильович — старший викладач кафедри метрології та промислової автоматики, e-mail: pvv_vin@ukr.net .

Вінницький національний технічний університет, Вінниця

O. M. Danylchuk[1]
V. V. Kovtun[2]
O. D. Nykytenko[2]
Yu. Yu. Nestiuk[2]
V. V. Prysiazhniuk[2]

Elements of Methodology of Precision Phonetic Analysis of Oral Phonograms

[1]Vasyl' Stus Donetsk National University;
[2]Vinnytsia National Technical University

The study of the cornerstone of modern linguistics - the process of speech and textual interpersonal communication, given the size of the infosphere of the twenty-first century, is impossible without a sound and purposeful involvement of information technology from other fields of knowledge, including computer science. The resulting relatively young science, computational linguistics, aims to automatically analyze natural languages in all spectra of their implementations. Among the long list of topical issues actively studied in the paradigm of computational linguistics, we mention the automation of compilation and linguistic processing of language corpora, automated classification and abstracting of documents, creating accurate linguistic models of natural languages, extraction of factual information from informal linguistic data. An effective, strictly formalized methodology for computational phonetic analysis of linguistic information, especially speech information, is potentially a driving force for improving the results of solving these research problems. This thesis is fully consistent with the content of the article, which proves the relevance of the presented scientific and applied results. Accordingly, the paper presents elements of the methodology of precision phonetic analysis of phonograms of oral speech, taking into account the phenomenon of phonetic fusion. The mathematical apparatus of the created methods is based on the provisions of the theory of pattern recognition, information theory and acoustic theory of language formation. This basis provided the basis for a system of analytical formalization of the problem of multicriteria of the process of recognition of language units of human speech. As a result, a method for reliable clustering of personal phonetic alphabets of speakers is presented. A method for detecting potentially unreliable classified speech units and adjusting the results of the process of automated transcription of speech signals is also presented. A method for estimating the influence of the medium of propagation of the studied speech signals on the transcription result is also proposed.

Keywords: computer linguistics, classification of language units, automated transcription, phonetic analysis of speech.

Danylchuk Oksana M. — Cand. Sc. (Educ.), Associate Professor, Associate Professor of the Chair of Applied Mathematics, e-mail: oksanadommod@ukr.net ;

Kovtun Viacheslav V. — Dr. Sc. (Eng.), Associate Professor, Professor of the Chair of Computer Control Systems, e-mail: kovtun_v_v@vntu.edu.ua ;

Nykytenko Olena D. — Cand. Sc. (Eng.), Associate Professor, Associate Professor of the Chair of Computer Control Systems, e-mail: lena260784@gmail.com ;

Nestiuk Yuliia Yu. — Student of the Department of Intelligent Information Technology and Automation, e-mail: yynestiuk@gmail.com ;

Prysiazhniuk Vasyl V. — Senior Lecturer of the Chair of Metrology and Industrial Automation, e-mail: pvv_vin@ukr.net