УДК 004.63

О. М. Ткаченко, Я. О. Тютюнник, П. В. Чирва, В. Л. Комаров

СИСТЕМА РОЗПІЗНАВАННЯ ЗВУКОВИХ СИГНАЛІВ НЕМОВЛЕННЄВОГО ПОХОДЖЕННЯ

Вінницький національний технічний університет

Анотація. Дана робота присвячена розробці програмного забезпечення, яке дало б змогу проводити розпізнавання звукових сигналів немовленнєвого походження. У статті проведений огляд існуючих на сьогоднішній день систем розпізнавання звуків, визначено їх переваги та недоліки. Приводиться список найбільш поширених алгоритмів, які можуть бути використанні для процесу навчання моделі та класифікації звуків. Більш детально розглядається модель гаусівського змішаного розподілу, яка і використовується для опису моделей звуків. Наводиться опис процесу розпізнавання звукових сигналів та їх подальша класифікація. Дана програма може використовуватись для аудіодетекції сигналів, наявних у базі даних. Продукт може використовуватись як самостійно, так і входити до складу програмно-апаратних комплаксів відповідного призначення. Використання засобів мови програмування С++ дозволило зробити його ефективним та зберегти швидкодію. Також наводиться статистика результатів розпізнавання та робиться висновок щодо ефективності системи.

Ключові слова: розпізнавання звуків немовленнєвого походження, Гаусівська змішана модель, класифікація, технології машинного навчання, системи розпізнавання звуків.

Аннотация. Данная работа посвящена разработке программного обеспечения, которое позволило бы проводить распознавания звуковых сигналов неречевых происхождения. В статье проведен обзор существующих на сегодняшний день систем распознавания звуков выделены их преимущества и недостатки. Приводится список наиболее распространенных алгоритмов, которые могут быть использованы для процесса обучения модели и классификации звуков. Более подробно рассматривается модель гауссовского смешанного распределения, которая и используется для описания моделей звуков. Приводится описание процесса распознавания звуковых сигналов и их дальнейшая классификация. Данная программа может использоваться для аудиодетекции сигналов имеющиеся в базе данных. Продукт может использоваться как самостоятельно, так и входить в состав программно-аппаратных комплексов соответствующего назначения. Использование средств языка программирования С ++ позволило сделать его эффективным и сохранить быстродействие. Также приводится статистика результатов распознавания и делается вывод об эффективности системы.

Ключевые слова: распознавания звуков неречевых происхождения, гауссовской смешанная модель, классификация, технологии машинного обучения, системы распознавания звуков.

Abstract. This work is devoted to the development of software that would allow the recognition of sound signals of non-speech origin. The article reviews the current sound recognition systems, identifies their advantages and disadvantages. The list of the most widespread algorithms which can be used for process of training of model and classification of sounds is resulted. The Gaussian mixed distribution model, which is used to describe the program's sound models, is considered in more detail. An overview of the process of recognition of sound signals and their further classification is given. This program can be used for audio detection of signals that are available in the database. The product can be used on its own or included in other more advanced recognition systems. Using the tools of the C ++ programming language allowed to make it efficient and maintain speed. Statistics of recognition results are also given and a conclusion on the efficiency of the system is made.

Keywords: recognition of sounds of non-speech origin, Gaussian mixed model, classification, machine learning technologies, sound recognition systems.

DOI: https://doi.org/10.31649/1999-9941-2020-49-3-30-36.

Вступ

Передумовою стрімкого розвитку голосових технологій є значне збільшення обчислювальних можливостей при значному зменшенні габаритів електронних обчислювальних пристроїв. Також варто відзначити розвиток математичних методів, що дозволяють виконати необхідну обробку аудіосигналу шляхом виділення з нього інформативних ознак.

Тема ідентифікації звукових сигналів є популярною на сьогоднішній день. Багато задач виникає при бажанні використати голосові команди для взаємодії з мобільними телефонами та іншою електронікою. Для прикладу, введення голосових команд для отримання інформації з мережі Інтернету, прокладання маршруту руху транспортного засобу, надиктовка тексту повідомлення голосом замість того щоб набирати його на клавіатурі. З розвитком функцій розумного будинку, з'явилася можливість управління домашньою, офісною технікою за допомогою електронних пристроїв голосовими командами. Все це знаходить широкого застосування в сучасному світі.

Також існують системи розпізнавання звуків, за допомогою яких робиться внесок у безпековий компонент нашого буття. Якщо покрити мережу міста мікрофонами та мати можливість розпізнавати постріли, крики про допомогу чи інші підозрілі звуки, поліція зможе вчасно реагувати на них. Точно визначивши джерело звуку можна навіть сказати поверх будинку та квартиру, де відбулася неправомірна ситуація.

Застосування таких систем суттєво поліпшує те, що прийнято називати «якістю життя». Проте в них усіх є свої недоліки, що проявляються в вузькій спеціалізації, закритості програмних продуктів та високій ціні. Таким чином можна стверджувати, що тема дослідження є на сьогоднішній день актуальною.

Мета

Метою дослідження є підвищення ймовірності коректного розпізнавання сигналів звукового походження за рахунок вибору і налаштування моделі та розробки алгоритмів і програмного забезпечення, яке за допомогою підготовленої бази звукових сигналів буде проводити їх класифікацію.

Для досягнення поставленої мети необхідно розв'язати такі задачі:

- провести аналіз існуючих на ринку систем розпізнавання звуку;
- дослідити існуючі моделі класифікації та визначити модель, яка найкраще підходить для розпізнавання звукового сигналу;
- дослідити сучасні інструменти, які дозволять реалізувати на практиці обрану модель класифікації;
- розробити програмне забезпечення;
- провести тестування розробленого програмного забезпечення.

Існуючі системи розпізнавання звуків

На сьогоднішній день є доволі великий та різноманітний асортимент програм, метою яких є розпізнавання звуків. Кожна з них має як свої недоліки, так і переваги. Слід зазначити, що усі вини є спеціалізовані під конкретний розв'язок певної задачі і не можуть бути використанні для інших цілей. Велику частину з них складають системи розпізнавання музичних файлів та системи розпізнавання мови людини. Найбільш відомими прикладами серед таких програм є Siri, Google Assistant, та Shazam. Також слід окремо згадати систему, яка дає змогу виявляти постріли у міському середовищі. Вона має назву ShotSpotter. Нижче буде детальніше розказано про кожну із систем.

ShotSpotter – це система, яка має змогу виявляти та передавати інформацію про розташування стрільби з вогнепальної зброї з використанням різних типів датчиків, таких як акустичні, оптичні та інші, а також комбінації таких датчиків. Дання система використовуються правоохоронними органами та приватними підприємствами для визначення джерела та напрямку. Також система може визначати тип пострілу.

Siri − це персональний помічник від компанії Apple. Вона є частиною операційній системі iOS, на базі якої працюють телефони, комп'ютери та планшети вищезгаданої компанії. Основне її завдання – це спілкування з людиною. Комп'ютер може давати інформацію користувачеві у подібному на людський форматі. В свою чергу користувач має змогу їй відповідати голосом та давати відповідні команди. Відбувається ніби спілкування схоже між двома людьми. Варто зазначити, що Sirі є результатом роботи досліджень, які накопичувалися більше сорока років.

Google Assistant представляє з себе хмарний сервіс який виконує функції персонального асистента. За великим рахунком він є прямим конкурентом Sirі та виконує подібні функції. Розроблений віртуальний асистент корпорацією Google. Сьогодні, продукт знаходиться на такому рівні, що може безпомилково розпізнавати мову людини та вести прямий діалог між машиною та живою людиною. Сервіс є дуже цікавим і корисним. Однак основним його завданням є розпізнавання мови і заточений він саме під ці задачі. Він не підходить для розпізнавання будь яких звуків.

Shazam – це ще один дуже корисний сервіс. Він допомагає знаходити музичні композиції по фрагменту з пісні. Якщо дати йому декілька секунд звучання будь якої частини музичної композиції, неважливо, якої саме частини та неважливо чи є в ньому спів виконавця, він знайде його у своїй базі та видасть користувачеві назву твору та усю інформацію про нього. Для цього Shazam створює спеціальну сигнатуру, що є ніби відбитком пальців, якщо проводити аналогію з людиною.

Якщо підводити підсумок то можна зазначити що існує багато систем розпізнавання звукових сигналів, проте кожна з них є вузькоспеціалізованою. Тобто кожна система може виконувати тільки задачі своєї області. Усі ці системи розроблялись приватними компаніями і відповідно усі вони є закритими технологіями. Данну характеристику можна віднести до основного недоліку наявних програм.

Вибір моделі для класифікації звуку

Алгоритми машинного навчання. Машинне навчання англійською Machine learning або ML – це окрема частина в області штучного інтелекту, а саме, набір алгоритмів та методів для побудови системи, яка буде мати змогу вчиться на власному досвіді [2]. Для навчання система обробляє великі масиви вхідних даних і має знайти у них закономірності. Саме те як система буде знаходити закономірності і описується різними методами навчання. Це сукупність алгоритмів, що дозволяють комп'ютеру робити правильні висновки на підставі даних. Варто зазначити, що у програми немає жорстко заданої логіки. Тобто машина отримує на вхід параметризованні данні та сама має знайти закономірності між ними у процесі роботи алгоритму навчання. Саме це вигідно відрізняє нейронні мережі від звичайних програм. Там де людина не зможе фізично обробити великі об'єми інформації нейромережа буде це робити. Людині не потрібно її заздалегідь програмувати. Усе необхідне вона зробить сама у процесі навчання.

Загалом процес навчання можна поділити на два типи [3]:

- навчання з учителем;

- навчання без учителя.

Відмінність полягає в тому, що на вхід системи яка навчається з учителем подаються параметризовані данні та вказується правильна відповідь. Це дозволяє системі підлаштовувати правила для класифікації таким чином, щоб отримати найкращий результат. В той же час на вхід системи яка навчається без учителя подаються тільки параметризованні данні. Такі системи часто використовуються тоді коли ми самі не знаємо правильної відповіді заздалегідь. Система розподіляє усі данні на вказане число кластерів, використовуючи принцип найбільшої подібності.

У процесі розпізнавання звуків буде використано явище класифікації. Класифікація – це процес у нейронній мережі який присвячений розв'язанню наступної задачі: у нас в наявності є множина об'єктів які деяким чином потрібно поділити на класи [4]. Також у нас в наявності є обмежена множина об'єктів, для яких відомо, до якого класу вони відносяться. Цю множину називають навчальною вибіркою. Класова ж приналежність інших об'єктів не відома. Вимагається побудувати такий алгоритм, що буде здатним класифікувати будь-який об'єкт початкової множини.

Задача класифікації відноситься до розділу навчання з учителем. Для цього у нас мають бути зразки кожного класу − об'єкти, про які заздалегідь відомо до якого класу вони належать. Для класифікації та розпізнавання звуку може бути використана більшість існуючих алгоритмів машинного навчання, які в змозі розв’язувати задачу класифікації. Основний список методів, який можна використати для задачі класифікації:

- метод опорних векторів;
- логістична регресія;
- метод k-найближчих сусідів;
- GMM класифікатор;
- наївний баєсівський класифікатор;
- дерева рішень.

На основі проведеного аналізу та результатів попередніх досліджень для навчання системи та подальшого розпізнавання було обрано класифікатор на основі Гаусовських змішаних моделей.

GMM класифікатор. Модель Гаусової суміші (англ. Gaussian Mixture Models) є імовірнісною моделлю. Вона базується на припущенні, що всі об’єкти, які належать до певних класів, утворюються з суміші скінченної кількості гаусових розподілів з невідомими параметрами [5].

Назва моделі «Змішана» полягає в тому, що дані виглядають мультимодально, тобто існує більше одного «піку» в розподілі даних. Спроби встановити мультимодальний розподіл з унімодальною (однією «піковою») моделлю, як правило призведуть до поганого результату. Це відбувається через зміщення спільного центру до області де спільних точок взагалі може не бути. Це явище можна спостерігати на рис. 1 (ліворуч). Оскільки існує багато простих розподілів які є одномодальними, очевидним способом моделювання мультимодального розподілу було б припущення, що він генерується множиною одномодальних розподілів. Таким чином, моделювання мультимодальних даних як суміші багатьох одномодальних гауссових розподілів має інтуїтивний сенс. Крім того, GMM підтримують багато теоретичних та обчислювальних переваг моделей Гаусса, що робить їх практичними для ефективного моделювання дуже великих наборів даних.

Обрана модель GMM для даних класів зразків звуків дозволяє краще описати кожен звук який буде занесено в базу. В подальшому це дасть можливість краще відрізняти зразки різних звуків, що призведе до поліпшення їх розпізнавання.

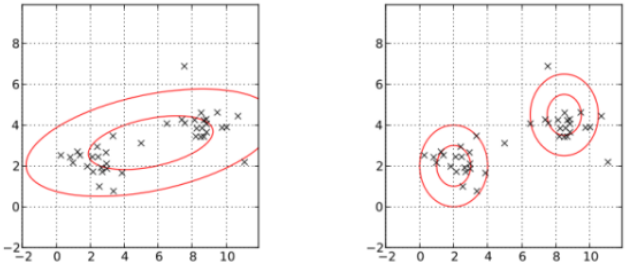

Рисунок 1 – Унімодальна та мультимодальна модель класу

Аудіодетекція звукового сигналу

Починається процес розпізнавання з перетворення акустичного сигналу в цифрову форму. Для цього використовуються методи аналогово-цифрового перетворення. На виході ми отримуємо звукові фрагменти у wav форматі. Далі, із звукового фрагменту, за допомогою спеціального математичного перетво-

рення виділяються певний набір параметрів. Тривалість кожного фрагменту складає 10 мс. Сам процес називається параметризація. На рис. 2 можна побачити загальну схему аудіодетекції звукового сигналу.

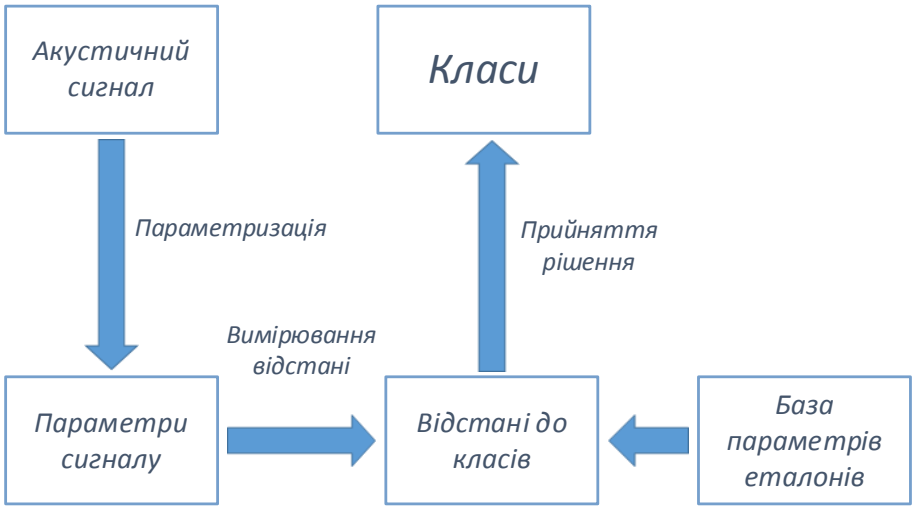

Рисунок 2 – Схема аудіодетекції звукового сигналу

На другому етапі розпізнавання відбувається вимірювання відстані до наявних шаблонних моделей, які є у базі. Так після першого етапу у нас є в наявності параметри фрейму, які ми можемо подати на вхід навченої моделі і вона нам видасть відстань до кожного класу, який наявний у базі. Данна відстань є імовірнісною характеристикою приналежності фрейму до класу.

Останнім етапом розпізнавання є процес прийняття рішення. Так як у нас є в наявності відстані до кожної моделі класу в базі ми можемо порівняти їх та визначити якій з них належить фрейм який проходить перевірку. Проте процес прийняття рішення зводиться до більш складної процедури ніж простий вибір найбільш імовірного значення. Так варто враховувати мінімальну кількість фреймів яка необхідна для ідентифікації певного звуку, попередні фрейми які проходили розпізнавання, можливість розмиття значення відстані, шляхом знаходження середнього серед певної кількості фреймів ті інші додаткові параметри. Після цього ми можемо віднести фрейм до певного класу, або констатувати що він не є схожим на жоден звук з бази.

Обробка аудіосигналу для отримання параметрів моделі. Параметризація аудіосигналу має на меті виділити в звуковому файлі інформацію, яка є значущою для задачі розпізнавання звуків. Данні ознаки будуть використані для створення шаблонних класів, а у подальшому для порівняння з ними вхідних фреймів. Загалом, важко сказати які з ознак краще підходять для задачі розпізнавання звуків. Вибір певного набору ознак встановлюється шляхом емпіричного підбору із подальшою перевіркою результатів та їх порівнянням [1].

Був використаний підхід за якого великий звуковий файл розбивається на маленькі частинки – фрейми. Тривалість фрейму вибирається в межах 5-20 мс. Передбачається, що параметри сигналу на таких малих проміжках змінюються не суттєво. Експериментальним шляхом було встановлено, що для даного підходу в розпізнаванні краще всього підходять кепстральні коефіцієнти.

Ми будемо використовувати підхід за якого кепстральні коефіцієнти будуються на основі лінійного передбачення. За допомогою методу лінійного передбачення можна апроксимувати поточний відлік застосовуючи лінійну комбінацією певної кількості попередніх відліків (1).

$$x_n \approx \sum_{k=1}^{p} a_k x_{n-k}, \tag{1}$$

За допомогою рекурсивного алгоритму Дарбіна ми можемо знайти коефіцієнти лінійного передбачення. В подальшому на їх основі розраховуються кепстральні коефіцієнти [4]. Варто зазначити, що в теорії таких коефіцієнтів може бути навіть більше ніж коефіцієнтів лінійного передбачення (2).

$$x_n = \begin{cases} a_n + \sum_{k=1}^{n-1} \frac{k}{n} c_k a_{n-k}, & 1 \le n \le p; \\ \sum_{k=n-p}^{n-1} \frac{k}{n} c_k a_{n-k}, & n > p. \end{cases} \quad (2)$$

У нашому випадку ми будемо використовувати 10 коефіцієнтів лінійного передбачення. На їх основі буде згенеровано 10 кепстральних коефіцієнтів.

Процес класифікації. Вектор із ознаками фрейму, у нашому випадку це набір кепстральних коефіцієнтів, який ми отримуємо після стадії параметризації, використовується для побудови шаблонних моделей та для порівняння з наявними в базі моделями. Для процесу розпізнавання ми маємо визначити спосіб за яким буде обчислюватись ступінь подібності зразка, який проходить перевірку, з одним або декількома шаблонами. Данна ступінь подібності являє собою певну метрику наближеності вхідного зразка до наявних класів. Найпоширенішими способами обчислення метрики є наступні:

- манхетенська відстань;
- евклідова відстань;
- відстань Махалонобіса;

Для обрахунку відстані в Гаусових змішаних моделях найчастіше використовують останню. Її формула наведена нижче. У цій формулі W^{-1} є матрицею коваріації.

$$d(x, y) = (x - y)^T * W^{-1} * (x - y). \quad (3)$$

Програма, яка буде порівнювати вхідний вектор із моделями які наявні в базі, для кожної моделі буде генеруватись число, яке показує наближеність фрейму до неї. Можна сказати, що чим це число є більшим – тим більшою є ймовірність того, що вектор є приналежним до даного класу.

Так, на рис. 3 наводиться порівняння відстаней до наявних у базі моделей. База, з якою проводиться порівняння, налічує 8 зразків звуків. До неї входять спів пташки, звуки пострілу з автомати, пістолета, рушниці, реактивної пускової установки а також крик людини, звук гуркоту трактора та тиші. Для порівняння з шаблонними зразками подається один фрейм який є частиною звуку роботи трактора. Як бачимо система його розпізнала найкраще. Наближеність до моделі звуку трактора склала 8,67.

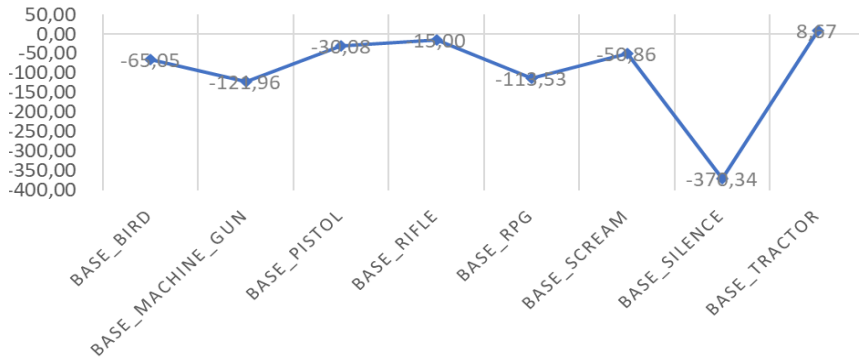

Рисунок 3 – Порівняння відстаней до наявних у базі моделей

На рисунку вище наведена ситуація за якої досить легко можна ідентифікувати приналежність вхідного вектора до певного класу. Проте існують ситуації коли результати не такі очевидні. Скажімо може виникати співпадіння одразу з декількома класами або і взагалі співпадіння з класом який не є істинним для фрейму. Тут вступає в дію розроблений алгоритм прийняття рішення. У ньому ми знаходимо середнє значення відстані із 30 послідовних фреймів, і тільки після цього приймаємо рішення про віднесення першого із них до певної моделі. Цей прийом дає змогу уникнути ситуації випадкових співпадінь.

Огляд результатів. Для тестування розробленої системи був згенерований файл тривалістю 26 секунд у якому були зібрані усі звуки які наявні у базі а також звук шуму, який там відсутній. Детальна інформація про результати розпізнавання наведена у табл. 1.

Таблиця 1 – Результати розпізнавання тестового звуку

№	Назва звуку	Усього фреймів	Розпізнано	Не розпізнано	Відсоток розпізнавання
1	Тиша	52	30	22	58%
2	Спів пташки	310	195	115	63%
3	Тиша	53	34	19	64%
4	Рушниця	157	83	74	53%
5	Тиша	38	19	19	50%
6	Пістолет	90	40	50	44%
7	Тиша	70	45	25	64%
8	Автомат	120	92	28	77%
9	Тиша	90	55	35	61%
10	Трактор	792	639	153	81%
11	Тиша	84	62	22	74%
12	Спів пташки	62	59	3	95%
13	Пістолет	82	45	37	55%
14	Тиша	131	101	30	77%
15	Шум	400	0	400	0%

Варто зазначити, що для створення тестового файлу бралися зразки звуків, які не використовувались для навчання моделей. Як видно із результатів розпізнавання, кожен звук був правильно розпізнаний та вірно віднесений до своєї моделі. Загальний відсоток розпізнавання становить 65%. Особливістю результатів є те, що при переході від одного зразка звуку до іншого втрачається точність розпізнавання. Це є очікувано, так як для розпізнавання фрейму ми накопичуємо послідовність із 30 фреймів і при граничних значеннях правильно розпізнані фрейми нівелюється дуже низькими значеннями неправильних.

Висновки

Підсумком виконання даного дослідження стала розробка програмного забезпечення, яке дає можливість розпізнавати сигнали немовленнєвого походження. Висока продуктивність та надійність системи досягається за рахунок її реалізації мовою програмування C++ в середовищі Visual Studio.

Розпізнавання звуків, так як і багатьох інших моделей розпізнавання, зводиться до вибору моделі яка буде представляти данні та алгоритмів машинного навчання і способу класифікації. За допомогою Гаусівських змішаних моделей ми описуємо шаблонні класи більш точно у мультимодальному форматі, ніж якби ми це робили за допомогою звичайних розподілів унімодально. За параметри моделі, що передаються на вхід системи розпізнавання, було використано кепстральні коефіцієнти. Їх вибір обумовлений попередніми дослідженнями.

Було проведено тестування розробленого продукту, що підтвердило ефективність розпізнавання звуків, наявних в базі. Усі звуки, які проходили перевірку були правильно віднесені до своїх шаблонних моделей. Загальний відсоток правильного розпізнавання фреймів склав 65%.

Розроблене програмне забезпечення може працювати, як самостійно на базі підготовленої бази даних звуків, так і може бути інтегроване у інші системи розпізнавання звуків немовленнєвого походження.

Список літератури

[1] Я. О. Тютюнник, Д. С. Чорний, О. М. Ткаченко, «Програмні засоби для аудіодетекції звукових сигналів немовленнєвого походження», *у Матеріали Всеукр. наук.-практ. конф. Молодь в науці, 11–30 травня. 2019 р*., Вінниця, 2019, с. 1–2.

[2] С. И. Николенко, А. А. Кадурин, Е. О. Архангельская, *Глубокое обучение. Погружение в мир нейронных сетей*. СПб, Россия: Питер, 2018.

[3] P. Domingos, P. Pazzani, «On the optimality of the simple Bayesian classifier under zeroone loss», *Machine Learning*, №29, с. 103–137, 1997.

[4] T. Hastie, R. Tibshirani, F. Jerome, *The elements of statistical learning*. Stanford, California, USA: Springer, 2008.

[5] Richard Duda, *Pattern Classification*. New York, USA: WileyInterscience, 2004.

Стаття надійшла: 19.11.2020.

References

[1] Ia. O. Tiutiunnyk, D. S. Chornyi, O. M. Tkachenko, «Prohramni zasoby dlia audiodetektsii zvukovykh syhnaliv nemovlennievoho pokhodzhennia», u *Materialy Vseukr. nauk.-prakt. konf. Molod v nautsi, 11–30 travnia. 2019 r*., Vinnytsia, 2019, s. 1–2.

[2] S. I. Nikolenko, A. A. Kadurin, E. O. Arhangel'skaja, *Glubokoe obuchenie. Pogruzhenie v mir nejronnyh setej*. SPb, Rossija: Piter, 2018.
[3] P. Domingos, P. Pazzani, «On the optimality of the simple Bayesian classifier under zeroone loss», *Machine Learning*, №29, с. 103–137, 1997.
[4] T. Hastie, R. Tibshirani, F. Jerome, *The elements of statistical learning*. Stanford, California, USA: Springer, 2008.
[5] Richard Duda, *Pattern Classification*. New York, USA: WileyInterscience, 2004.

Відомості про авторів

Ткаченко Олександр Миколайович – кандидат технічних наук, доцент, доцент кафедри обчислювальної техніки Вінницького національного технічного університету.

Тютюнник Ярослав Олександрович − магістр кафедри обчислювальної техніки Вінницького національного технічного університету.

Чирва Павло Васильович − аспірант кафедри обчислювальної техніки Вінницького національного технічного університету.

Комаров Володимир Леонідович − аспірант кафедри обчислювальної техніки Вінницького національного технічного університету.

О. М. Ткаченко, Я. А. Тютюнник, П. В. Чирва, В. Л. Комаров

СИСТЕМА РАСПОЗНАВАНИЯ ЗВУКОВЫХ СИГНАЛОВ НЕРЕЧЕВЫХ ПРОИСХОЖДЕНИЯ

Винницкий национальный технический университет, Винница

O. M. Tkachenko, Ya. O. Tiutiunny, P. V. Chyrva, V. L. Komarov

RECOGNITION SYSTEM FOR SOUND SIGNALS OF NON-SPEECH ORIGIN

Vinnytsia National Technical University, Vinnytsia