ПРИКЛАДНА ГЕОМЕТРІЯ, ІНЖЕНЕРНА ГРАФІКА ТА ЕРГОНОМІКА

УДК 004.6
DOI https://doi.org/10.32838/2663-5941/2021.4/01

Кірей К.О.
Чорноморський національний університет імені Петра Могили

ПРОБЛЕМИ ВИКОРИСТАННЯ АКУСТИЧНОГО ВІДБИТКА ЩОДО ІДЕНТИФІКАЦІЇ МЕДІАКОНТЕНТУ

У статті розглядається дослідження алгоритмів аудіофінгерпринтингу щодо розв'язку проблеми ідентифікації невідомого аудіо за допомогою відчинених і вільно доступних програмних компонентів, вебсервісів і баз медіаданих. Збільшення інформації, зокрема медіаінформації, приводить до потреби в пошуку нових і вдосконаленню наявних засобів ідентифікації медіаконтенту. Нині спостерігається розвиток відповідних технологій, проте проблема ідентифікації медіаконтенту, зокрема музичного контенту, ще не набула остаточного розв'язання. Для ідентифікації музичного контенту застосовується концепція акустичного фінгерпринту (акустичного відбитка). Надійний алгоритм акустичного відбитка має враховувати перцептивні характеристики аудіо, бути стійким до деградації якості аудіо (радіоперешкоди, артефакти, шуми тощо). Також він має брати до уваги особливості різних форматів кодування аудіо. Зокрема, це сімейство lossy-форматів, які вносять значні зміни до цифрового кодування звукового файлу для максимального зменшення його розміру, водночас без значного впливу на те, як закодовані звуки сприйматиме людина. Надалі автоматизація ідентифікації аудіо розвивається в напрямі автоматизації аналізу спектрограм, що показують інтенсивність деяких частот упродовж часу. З першої половини 2000-х років дослідники почали застосовувати до цих зображень техніки комп'ютерного бачення. У 2010 році словацький програміст Л. Лалинский, базуючись на численних академічних дослідженнях, почав розроблення нового алгоритму аудіофінгерпринтингу – Chromaprint. Ключовою відмінністю алгоритму стала подальша обробка спектрограми для визначення деяких музичних нот. Це дало змогу зберігати дані зі спектрограми більш компактно, а також зробити їх стійкішими до пошкоджень аудіосигналу, спричиненого lossy-кодеками. Отже, головна проблема індексації музичного контенту полягає в одержанні розумного балансу між якістю ідентифікації, швидкістю цього процесу й обсягом сгенерованої вихідної інформації. Ми вважаємо, що зараз найкраще себе показав алгоритм фінгерпринтингу Chromaprint, стійкий до зазначених проблем ідентифікації аудіо, який зберігає результуючі акустичні відбитки в компактному форматі, водночас підтримуючи високу швидкість індексації та пошуку.

Ключові слова: *ідентифікація аудіо, формат аудіофайлу, спектрограма, хромаграма, алгоритми аудіофінгерпринтингу.*

Постановка проблеми. Бурхливий розвиток технологій упродовж останніх двох століть призвів до того, що кількість інформації, доступної кожній людині зростає експоненційно. Від численних цифрованих копій книг, газет, журналів, кінофільмів, музичних та художніх творів, до створеного вже в цифровому форматі різноманітного контенту – до всього цього можна отримати доступ завдяки інтернет технологіям. Така загальнодоступність інформації дала змогу ще більше пришвидшити розвиток освіти та технологій, а також дати змогу невідомим раніше авторам здобути заслужену аудиторію, не обмежену бажаннями та цілями наукових журналів, літературних видавництв чи музичних студій. І хоча беззаперечно, це є черговим кроком людства в напрямку суспільного розвитку, проте таке збільшення інформації спричиняє певні проблеми. Коли інформації забагато, знайти потрібну часто нетривіальна, а інколи й майже неможлива задача.

Аналіз останніх досліджень і публікацій. Інформаційні технології пошуку інформації, зокрема текстової, набули значного розвитку. Про це свідчить чимало праць вітчизняних і зарубіжних вчених у сфері інформаційних технологій (А. Бродера, Е. Большакової,

А. Вавіленкової, О. Іванова, Дж. Лакоффа, Д. Ланде, Д. Люгера, Д. Маккарти, А. Ньюелла, У. Питтса, Е. Попова, Д. Поспєлова, Л. Шеченко, Дж. Гопкрофта та інших) [1–5]. Такі провідні ІТ-компанії, як Microsoft, Google, Amazon, Facebook та інші успішно використовують технології аналізу текстів із метою подальшої автоматизації процесів обробки даних. Зокрема, індексація гіпертексту займає значно менше обчислювальних ресурсів і місця в пам'яті ЕОМ ніж індексація інших типів інформації. Тому, наприклад, донедавна пошук додаткової інформації про медіафайли в Інтернеті, залежав від наявності пов'язаної з ними текстової інформації. Проте така інформація не завжди є доступною й це перетворює пошук метаданих цих медіафайлів у довгий і складний процес. Особливо пошук ускладняється у випадку наявності достатньо великої медіатеки, де ручний пошук часто не є здійсненним. Отже, процес пошуку медіаконтенту потребує автоматизованого вирішення саме за вмістом аудіо. Розвиток систем штучного інтелекту, комп'ютерного бачення та слухання дав змогу машинам сприймати медіаінформацію схожим чином до того, як її сприймають люди. І, завдяки набагато кращим (та швидшим) здібностям до пошуку інформації, ніж у людей, комп'ютери, нарешті, можуть розв'язати проблему ідентифікації майже будь-якої інформації та пошуку її метаданих.

Дослідження засноване на працях видатних авторів в області розробки програмного забезпечення, таких як Х. Грегорі, Т. Калкер, Ю. Ке, Л. Лалінський, А. Лерч, А. Марк, Р. Суктханкар, Дж. Хайтсм, Д. Хойем, Т. Цай, Д. Чан та ін. [7; 8; 10; 11; 13–15]. Принцип роботи бази даних акустичних відбитків описано в роботах [11; 12].

Постановка завдання. Проте виявилося, що проблема ідентифікації медіаконтенту, зокрема, музичного контенту ще не набула остаточного вирішення. Отже, метою цієї статті є дослідження алгоритмів аудіофінгерпринтингу щодо розв'язання проблеми ідентифікації невідомого аудіо з допомогою відкритих та вільно доступних програмних компонентів, вебсервісів і баз медіаданих.

Виклад основного матеріалу дослідження. Акустичний фінгерпринт або акустичний відбиток (англ. Acoustic Fingerprint) – це представлення аудіосигналу у вигляді набору значень, що описують його фізичні властивості [6]. Практичне використання акустичного фінгерпринтингу полягає в ідентифікації пісень, мелодій чи певних звуків. Ідентифікація застосовується до аудіо з радіо трансляцій, CD-дисків, сервісів потокової передачі медіа чи мереж типу peer-to-peer. Метою ідентифікації аудіо може бути як моніторинг забезпечення збереження авторських прав, ліцензування чи інших схем монетизації, так і для особистого використання, наприклад, для визначення автора та назви невідомої композиції.

Надійний алгоритм акустичного відбитка має враховувати перцептивні характеристики аудіо [7, с. 7]. Тобто, якщо два звукових файли звучать однаково для людського вуха, їхні акустичні відбитки мають збігатись, навіть якщо їхні цифрові представлення є досить різними. Акустичні відбитки не є хеш-функціями, які мають бути чутливими до найменших змін вхідних даних. Вони дійсно схожі на людські відбитки пальців, для яких мінімальні зміни не є важливими. Відомо багато випадків, коли неякісний відбиток людського пальця був правильно зіставлений із відбитком у базі даних. Акустичні відбитки мають працювати схожим чином.

Найбільш часто використовуваними у сфері аудіофінгерпринтингу звуковими характеристиками є точки перетину нуля звуковою функцією, музичний темп, гармонічні ряди звуків, коефіцієнт тональності, частотний інтервал та пропускна смуга аудіо [8]. Для визначення цих характеристик більшість інструментів користується спектрограмами. Спектрограма є представленням аудіочастоти в часі. Вона дає змогу ідентифікувати частотний вміст аудіо щодо часу, упродовж якого це аудіо триває, а також наскільки тихою чи гучною є кожна частота. Проте необроблені спектрограми не є дуже корисними для визначення акустичних відбитків. По-перше, вони містять багато інформації, яка не потрібна для аудіофінгерпринтингу. По-друге, вони не є стійкими до деградації якості аудіо [8]. На рис. 2 зображено той самий аудіо файл, що й на спектрограмі рис. 1, але зіграний у шумному середовищі.

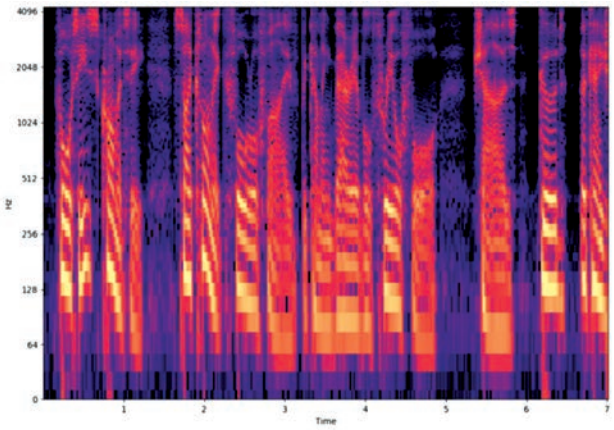

Рис. 1. Спектрограма аудіо

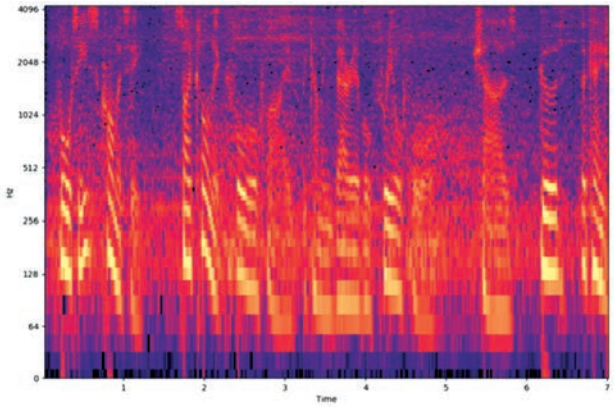

Рис. 2. Спектрограма з великою кількістю фонового шуму

Можна побачити, що фоновий шум призвів до появи зовсім іншої спектрограми. Незважаючи на це майже всі пікові рівні, аудіо залишилися легко пізнаваними. Тобто пікові рівні спектрограми є хорошим початком для генерації надійного акустичного відбитка.

Також, причиною змін спектрограми та, інколи, відчутних на слух змін звуку можуть стати формати кодування аудіо. Найпоширенішими форматами аудіофайлів є lossy-формати (MP3, AAC, Vorbis, Opus тощо). Ці формати вносять значні зміни до цифрового кодування звукового файлу для максимального зменшення його розміру, водночас без значного впливу на те, як закодовані звуки сприйматиме людина. Надійна система аудіофінгерпринтингу має ідентифікувати музичний запис, навіть після такого стиснення, якщо якість аудіо була сильно зменшена. Для використання в ідентифікації радіопередач, системи акустичного фінгерпринтингу має бути ще і стійкими до радіоперешкод та артефактів, спричинених аналоговою передачею.

Традиційно двовимірні репрезентації аудіочастот у часі, такі як спектрограми, сприймалися як зображення лише для цілей візуалізації. Але, починаючи з першої половини 2000-х років дослідники почали застосовувати техніки комп'ютерного бачення до цих зображень. Алгоритм генерації акустичних відбитків, розроблений компанією Shazam (www.shazam.com) був одним із перших [9]. Він використовує пошук пікових рівнів на спектрограмі, але водночас він об'єднує найближчі пікові точки, утворюючи своєрідну павутину із точок та ліній. Окрім того, що таке представлення займає набагато менше місця, утворена павутина з'єднань між піками надає системі додаткової стійкості щодо фонових шумів. Приблизно водночас, дослідницький відділ у сфері аудіо компанії Phillips запропонував інший алгоритм [10]. На відміну від алгоритму компанії Shazam, який вибірково запам'ятовує найважливіші частини спектрограми, алгоритм Phillips намагається якомога більше стиснути всю спектрограму. Це досягається завдяки збереженню лише змін частот у часі. Ключовою відмінністю цього алгоритму від попереднього є значно більша стійкість до різких спалахів фонового шуму, адже алгоритм зберігає всі дані, а не лише ті, що виділяються найбільше. Так само такий алгоритм має меншу стійкість до постійного фонового шуму.

У 2010 році словацький програміст Л. Лалинский, базуючись на численних академічних дослідженнях, почав розроблення нового алгоритму аудіофінгерпринтингу – Chromaprint [11]. Ключовою відмінністю алгоритму стала подальша обробка спектрограми для визначення музичних нот. Це дало змогу зберігати дані зі спектрограми більш компактно, а також зробити їх більш стійкими до пошкоджень аудіосигналу, викликаних lossy-кодеками. Оскільки алгоритм націлений на розпізнавання аудіофайлів, не потрібно вживати багатьох заходів для зменшення впливу фонових шумів як у алгоритмах, спрямованих на розпізнавання музики з навколишнього середовища. Натомість, алгоритм включає оптимізацію розміру відбитка, що дало змогу побудувати на його основі найбільшу відкриту базу даних акустичних відбитків – AcoustID [12]. База даних, на сьогодні, зберігає більше ніж 75 мільйонів акустичних відбитків і тисячі додаються туди щодня (acoustid.org/stats). Алгоритм Chromaprint є найширше використовуваним Open Source-алгоритмом акустичного фінгерпринтингу і є вільним для використання на умовах MIT-ліцензії.

Алгоритм Chromaprint базується на комп'ютерному баченні щодо ідентифікації музики, підґрунтя якого описано в роботі науковців Ян Ке, Д. Хойема й Р. Суктханкара [13]. З першого погляду, проблеми сфери аудіо мають небагато спільного з комп'ютерним баченням. Алгоритми, що працюють з аудіо, зазвичай займаються обробкою одновимірних сигналів у часі, тоді як комп'ютерне бачення найчастіше працює з інтерпретаціями одного чи більше двовимірних зображень (зазвичай знятих із тривимірної сцени). Науковці вважають, що певні проблеми у сфері аудіо легко трансформуються у форму, придатну для ефективної обробки з допомогою комп'ютерного бачення. Така думка підкріплюється тим, що науковці досить часто користуються двовимірними репрезентаціями аудіо в цілях візуалізації.

Коли люди розглядають аудіозапис, вони зазвичай дивляться на форму звукової хвилі. Таку форму показує більшість програм для роботи з аудіо, але вона не є дуже корисною для аналізу. Більш корисною репрезентацією є спектрограма, яка показує інтенсивність певних частот упродовж часу (рис. 3).

Таке зображення можна здобути розділивши оригінальне аудіо на багато кадрів, що частково перекривають один одного й застосувавши алгоритм перетворення Фур'є (а саме – віконне перетворення Фур'є). У випадку Chromaprint, віконне перетворення Фур'є виконується на вхідному аудіо, конвертованому в частоту дискретизації 11025 Гц із коефіцієнтом перекриття кадрів 2/3, кожен із яких має розмір 4096 байт.

Багато алгоритмів ідентифікації аудіо працюють із такою репрезентацією вхідного аудіо. Деякі порівнюють різниці частот упродовж часу, деякі виділяють пікові рівні в зображенні тощо. Зі свого боку Chromaprint обробляє інформацію перетворюючи частоти в музичні ноти. Алгоритм працює з нотами, а не з октавами, тому результат складається з 12 секцій, по одній для кожної ноти. Така інформація називається хромаграмою (англ. Chromagram, Chroma Features) [14]. Після незначного фільтерингу та нормалізації зображення буде виглядати так (рис. 4).

Тепер ми маємо репрезентацію аудіо, що є стійкою до змін, викликаних lossy-кодеками чи іншими схожими процесами. До того ж, такі зображення можна досить легко порівнювати між собою для перевірки того наскільки вони схожі, тобто наскільки схоже звучить оригінальне аудіо. Але для зберігання таких результатів у базі даних, з можливістю їхнього швидкого пошуку необхідна більш компактна форма. Ідея методу, який дає змогу перетворити дані в таку форму спирається на комп'ютерне бачення щодо ідентифікації музики з деякими модифікаціями, ґрунтованими на положеннях попарно прискореного акустичного відбитка [15]. Роботу результуючого алгоритму стиснення можна уявити як вікно розміром 16 на 12 пікселів, що переміщується по зображенню поступово, по пікселю за один раз. Цей процес генерує багато субзображень. До кожного з них застосовується визначений наперед набір з 16 фільтрів, що захоплює різниці інтенсивностей серед музичних нот у часі. Робота фільтрів полягає у визначенні суми специфічних частин субзображення, представлених у вигляді відтінків сірого (англ. Grayscale), і порівнянні двох отриманих сум.

Кожен фільтр має три коефіцієнти, асоційовані з ним, що показують як необхідно квантизувати дійсне число так, щоб у результаті здобути ціле число від 0 до 3. Фільтри й коефіцієнти можна обрати алгоритмом машинного навчання на основі тренувального набору аудіофайлів під час розроблення алгоритму.

Є шістнадцять фільтрів і кожен із них створює ціле число, яке може бути закодовано двома бітами (а допомогою коду Грея – системи кодування інформації, у якій два послідовні коди відрізняються значенням лише одного біта). Отже, скомбінувавши отримані результати на виході дістаємо 32-бітне ціле число. Якщо виконати вищеописані дії для кожного субзображення, генерованого вікном 16x12, що рухається по зображенню, у результаті буде отримано повний акустичний відбиток вхідного аудіо.

Далі необхідно виконати порівняння отриманих відбитків. Тут найбільш влучним способом, на нашу думку, є порівняння на основі визначення коефіцієнту бітових помилок. Отже, для порівняння отриманих відбитків нами обрано два випадкових lossless-стиснутих аудіофайлів (у форматі FLAC). Композиції з файлу 1 і файлу 2 виконуються одним і тим же співаком і містяться

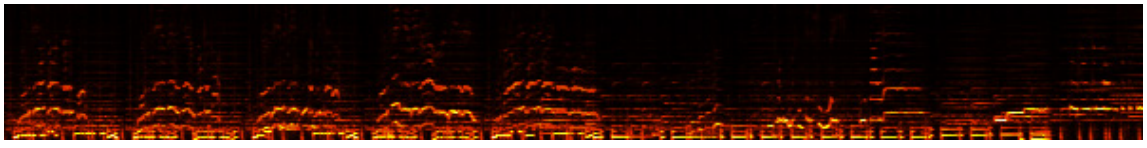

Рис. 3. Спектрограма інтенсивності частот

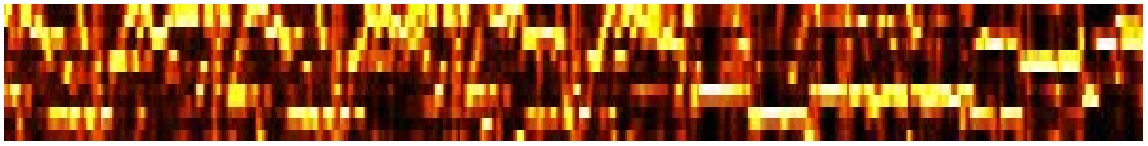

Рис. 4. Спектрограма аудіо після трансформацій, фільтерингу та нормалізації

в одному альбомі, тобто мають схожий стиль. Потім їх конвертовано в сильно стиснутий lossy-формат (MP3 32 Кбіт/сек) та згенеровано відбитки для чотирьох отриманих файлів. Сгенеровані відбитки наведено на рис. 5–8.

Для демонстрації можливостей алгоритму Chromaprint, порівняємо різницю відбитків оригінальних FLAC-файлів і стиснутих MP3-файлів (рис. 9-10).

Як бачимо, хоча різниця між FLAC та MP3 версіями є, вона не є значною. Тим паче, що неозброєним оком вона майже непомітна. Також нами виконано аналогічне порівняння між FLAC-версіями файлу 1 та файлу 2. Кількість шуму у разі порівнянні акустичних відбитків двох різних аудіо файлів є набагато більшою, ніж під час порівняння двох версій одного й того ж файлу, де один із них стиснуто агресивним lossy-алгоритмом. З цього можна зробити висновок, що алгоритм Chromaprint є стійким до пошкоджень файлу, викликаних lossy-кодуванням та іншими схожими процесами.

Для глибшого аналізу можливостей алгоритму Chromaprint, ми порівняли оригінальну й інструментальну версію двох інших файлів, файлу 3 та файлу 4 (рис. 11, 12).

Рис. 5. Акустичний відбиток файлу 1 у lossless-форматі

Рис. 6. Акустичний відбиток файлу 1 у lossy-форматі

Рис. 7. Акустичний відбиток файлу 2 у lossless-форматі

Рис. 8. Акустичний відбиток файлу 2 у lossy-форматі

Рис. 9. Різниця відбитків lossless- та lossy-версій файлу 1

Рис. 10. Різниця відбитків lossless- та lossy-версій файлу 2

Рис. 11. Різниця між оригінальною та інструментальною версіями файлу 3

Рис. 12. Різниця між оригінальною та інструментальною версіями файлу 4

Як бачимо, різниця залежить від конкретної композиції й може бути як досить незначною, так і вельми помітною на рівні двох різних версій однієї й тієї ж композиції. Але, навіть незначна різниця між оригінальною та інструментальною версіями композиції є помітно більшою, ніж різниця між lossy та lossless кодуваннями однієї й тієї ж композиції.

Висновки. Проведенні дослідження дали змогу виявити проблеми ідентифікації аудіа, такі як різне виконання однієї й тієї ж композиції (музичні інструменти, співаки, місце виконання тощо), деградація якості аудіо (радіоперешкоди, артефакти, шуми та ін.), особливості кодування аудіо в різних форматах. Усе це значно ускладнює автоматизацію процесу ідентифікації аудіо, де головна проблема полягає в одержанні розумного балансу між якістю ідентифікації, швидкістю цього процесу та обсягом сгенерованої вихідної інформації. Ми вважаємо, що зараз найкраще себе показав алгоритм фінгерпринтингу – Chromaprint, стійкий до зазначених проблем ідентифікації аудіо, який зберігає результуючі акустичні відбитки в компактному форматі, водночас підтримуючи високу швидкість індексації та пошуку. Це дає змогу створити велику базу даних акустичних відбитків силами однієї людини. Так, уся база даних акустичних відбитків AcoustID з відбитками та пов'язаними з ними даними про виконавців та назву композиції в стиснутому вигляді, займає менше одного гігабайта завдяки вищеописаному алгоритму стиснення.

Список літератури:

1. Вавіленкова А.І. Теоретичні основи аналізу електронних текстів : монографія. Київ : ТОВ «СІК ГРУП УКРАЇНА», 2016. 192 с.

2. Автоматическая обработка текстов на естественном языке и анализ данных : учебное пособие / Е.И. Большакова и др. Москва : НИУ ВШЭ, 2017. 269 с.

3. Іванов О.В. Класичний контент-аналіз та аналіз тексту: термінологічні та методологічні відмінності. *Вісник ХНУ ім. В.Н. Каразіна*. 2013. № 1045. Вип. 30. С. 69–74.

4. Шевченко Л.І., Дергач Д.В., Сизонов Д.Ю. Медіалінгвістика : словник термінів і понять / За ред. Л.І. Шевченко. Київ : ВПЦ Київський університет, 2014. 380 с.

5. Гопкрофт Дж. *Вікіпедія: вільна енциклопедія*. URL: https://uk.wikipedia.org/wiki/Джон_Гопкрофт#Бібліографія (дата звернення: 11.01.2021).

6. Acoustic fingerprint. *Wikipedia: The Free Encyclopedia*. 2020. URL: https://en.wikipedia.org/wiki/Acoustic_fingerprint#cite_ref-1 (дата звернення: 11.01.2021).

7. Lerch A. An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics. The Institute of Electrical and Electronics Engineers, 2012. 248 p.

8. Audio Hashprints: Theory & Application. Electrical Engineering and Computer Sciences, University of California at Berkeley / T. Tsai. Tech. Report. UCB/EECS-2016-185, 1 Dec. 2016.

9. A. Li-Chun Wang. An Industrial-Strength Audio Search Algorithm. *Shazam Entertainment, Ltd*. 2003. URL: https://www.ee.columbia.edu/~dpwe/papers/Wang03-shazam.pdf. (дата звернення: 11.01.2021).

10. Haitsma J., Kalker T. A Highly Robust Audio Fingerprinting System. *Ismir*. 2002. Vol. 2002, Nov. P. 107–115.

11. Chromaprint 1.4 released. *Lukáš Lalinský*. URL: https://oxygene.sk/ (дата звернення: 11.01.2021).

12. Welcome to AcoustID! *AcoustID: open source audio identification*. URL: https://acoustid.org/ (дата звернення: 11.01.2021).

13. Ke Y., Hoiem D., Sukthankar R. Computer vision for music identification: Video demonstration. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2005. P. 1184.

14. Mark A., Gregory H. AudioThumbnailing of Popular Music Using Chroma-Based Representations. *IEEE Transactions On Multimedia*. 2005. Vol. 7. No. 1. P. 96–104.

15. Jang D., Yoo C., Lee S., Kim S., Kalker T. Pairwise Boosted Audio Fingerprint. *IEEE Transactions On Information Forensics And Security*. 2009. Vol. 4. No. 4. P. 995–1004.

Kirei K.O. **PROBLEMS OF USING ACOUSTIC FINGERPRINT ON MEDIA CONTENT IDENTIFICATION**

The article is considered the study of audio fingerprinting algorithms to solve the problem of identifying unknown audio using open and freely available software components, web services and media databases. The increase in information, in particular, media information leads to the need to find new and improve existing means of identifying media content. Today there is the development of appropriate technologies; however, the problem of identification of media content, including music content has not yet entered the final decision. The concept of acoustic fingerprint (acoustic imprint) is used to identify music content. A reliable acoustic imprint algorithm

must take into account the perceptual characteristics of audio, be resistant to degradation of audio quality (radio interference, artifacts, noise, etc.). It should also take into account the features of different audio encoding formats. In particular, it is a family of lossy codecs that make significant changes to the digital encoding of an audio file to minimize its size, without significantly affecting how the encoded sounds will be perceived by humans. Automation of audio identification is developing in the direction of automating the analysis of spectrograms showing the intensity of individual frequencies over time. Since the first half of the 2000s, researchers began to apply computer vision techniques to these images. In 2010, the Slovakian programmer L. Lalinski, based on numerous academic studies, began the development of a new audio fingerprinting algorithm, Chromaprint. The key difference of the algorithm was the further processing of the spectrogram to determine individual musical notes. This allowed the data from the spectrogram to be kept more compact, and also made more resistant to audio signal damage caused by lossy codecs. So the main problem of indexing music content is to achieve a reasonable balance between the quality of identification, the speed of this process and the amount of generated source information. We believe that Chromaprint, the fingerprinting algorithm, has proved to be the best tool. It is resistant to these audio identification problems and keeps the resulting acoustic prints in a compact format, while maintaining a high speed of indexing and search.

Key words: *audio identification, audio file format, spectrogram, chromagram, fingerprinting algorithms.*