УДК 621.39

О. М. ТКАЧЕНКО, О. Ф. ГРІЙО ТУКАЛО

Вінницький національний технічний університет, м. Вінниця

ПІДХІД ДО ОЦІНЮВАННЯ ТРИВАЛОСТІ ФРАГМЕНТА ДЛЯ ПОШУКУ МУЗИЧНОГО ТВОРУ ЗА ЗАДАНИМ ШАБЛОНОМ

Анотація. Стаття належить до області інформаційних технологій, зокрема ідентифікації музичного твору на основі аудіо контенту. Теоретично обґрунтовано можливість ідентифікації музичного твору за його фрагментом. Застосовано кластерний аналіз під час формування шаблонів музичних творів в БД, що дозволяє зменшити обсяги пам'яті для їх зберігання. Запропоновано критерій порівняння фрагменту музичного твору з шаблонами БД. Визначено мінімальну тривалість фрагменту, що дозволяє суттєво зменшити складність обчислень в процесі ідентифікації музичного твору. Експериментальні результати підтвердили коректність теоретичних положень.

Ключові слова: ідентифікація за фрагментом аудіо запису, параметризація, мел-частотні кепстральні коефіцієнти, кластерний аналіз, Евклідова відстань.

Аннотация. Статья относится к области информационных технологий, в частности идентификации музыкального произведения по аудио контенту. Теоретически обоснована возможность идентификации музыкального произведения по его фрагменту. Применен кластерный анализ при формировании шаблонов музыкальных произведений в БД, что позволяет уменьшить объемы памяти для их хранения. Предложен критерий сравнения фрагмента музыкального произведения с шаблонами БД. Определена минимальная продолжительность фрагмента, что позволяет существенно уменьшить сложность вычислений в процессе идентификации музыкального произведения. Экспериментальные результаты подтвердили корректность теоретических положений.

Ключевые слова: идентификация по фрагменту аудио записи, параметризация, мел-частотные кепстральные коэффициенты, кластерный анализ, Евклидовое расстояние.

Abstract. Article relates to the field of information technology, in particular the content-based identification of songs. It is theoretically justified the ability to identify a piece of music by its fragment. Cluster analysis is applied when forming templates of music in the database, which reduces the amount of memory to store them. We propose the criterion for comparing the song fragment with templates database. The minimal length of the fragment is determined, which significantly reduces the computational complexity in the process of the song identifying. The experimental results confirm the correctness of theoretical positions.

Keywords: audio identification by fragment, parameterization, mel-frequency cepstral coefficients, cluster analysis, Euclidean distance.

Вступ

Дана стаття належить до області інформаційних технологій, зокрема автоматичної ідентифікації музичного твору за фрагментом аудіо запису.

В сучасних комп’ютерних мережах основний об’єм трафіку припадає на мультимедійну, зокрема, аудіо інформацію. Зростання обсягу мультимедійної інформації, що передається і обробляється в комп’ютерних системах, зумовила необхідність автоматизації процесів аналізу і пошуку даних. Більшість сучасних систем пошуку аудіо інформації використовує метадані (текстові описи про виконавця, назву музичного твору, рік запису тощо) та текстові анотації аудіо контенту [1]. Недоліком пошуку виключно на основі метаданих є те, що користувач пошукової системи повинен мати досить чітке уявлення про зміст того, що він шукає. Тому метадані доповнюють анотаціями (тегами) змісту аудіо запису [2]. Генерація таких анотацій аудіо контенту є трудомістким і тривалим процесом [1, 3]. Таким чином, в сучасних системах обробки аудіо інформації виникає необхідність автоматичного швидкого пошуку музичних творів на основі аудіо контенту у віддалених базах даних (БД) великого розміру на сервері. Виходячи з вищесказаного, суть пошуку музичного твору на основі аудіо контенту полягає в тому, щоб автоматично отримувати файли аудіо записів музичних творів, подібних до заданого аудіо запису під час запиту. При цьому, враховуючи великі обсяги аудіо інформації в БД, велике значення має швидкість пошуку. В зв’язку з цим в даній статті йде мова про можливість ідентифікації музичного твору саме за коротким фрагментом. Важливо, щоб тривалість фрагменту була якомога меншою, оскільки це дозволить: 1. збільшити швидкість пошуку; 2. зменшити час завантаження та мережевий трафік. Разом з тим зменшення тривалості фрагменту може зумовити зростання ймовірності помилки під час ідентифікації музичного твору. Тому головне питання, на яке потрібно дати відповідь: яка мінімальна тривалість фрагменту є достатньою, щоб характеризувати весь музичний твір.

Мета та задачі статті

Метою даної статті є теоретичне обґрунтування можливості ідентифікації музичного твору за його фрагментом та мінімальної тривалості фрагменту, що дозволяє зменшити складність обчислень в процесі автоматизованої ідентифікації музичного твору.

Для реалізації системи ідентифікації музичного твору необхідно розв’язати такі задачі: 1) обрати параметри, які дозволили б однозначно та компактно описати музичний твір; 2) створити швидкий та надійний метод порівняння за обраними параметрами вхідного фрагменту музичного твору та попередньо створених еталонів (шаблонів), що зберігаються в базі даних; 3) обґрунтувати мінімальну тривалість фрагменту, який дозволяє ідентифікувати музичний твір; 4) провести експериментальну

перевірку запропонованих теоретичних положень. Результатом розпізнавання буде шаблон БД з мінімальним розходженням відносно вхідного аудіо запису.

Далі в роботі вважається, що аудіо запис, який треба ідентифікувати, точно збігається з одним із записів, що містяться в БД. Проте всі зроблені надалі висновки залишаються справедливими і в тому випадку, коли вхідний аудіо запис та відповідний йому шаблон БД не є ідентичними в силу таких факторів як наявність шумів, зміна темпу, частоти дискретизації тощо.

Математична модель аудіо сигналу (MFCC)

Одним з центральних понять інформаційного пошуку музики (MIR) є схожість/подібність музичних творів. Моделювання подібності музики – це основа програм для автоматичної організації та обробки баз даних музики. Схема визначення відповідності музичних творів на основі контенту базується на використанні аудіо файлу для побудови моделі аудіо сигналу.

Порівнювати безпосередньо звукові сигнали в часовій області — довго і не дуже ефективно, тому відліки аудіо сигналу ділять на невеликі фрагменти (фрейми), для яких характеристики сигналу залишаються відносно стійкими (стаціонарний випадковий процес), з перекриттям фреймів. Тривалість одного фрейму, як правило, лежить у межах 10 – 30 мс. Для кожного фрейму виконується спектральний аналіз, на основі якого (тим чи іншим чином) обчислюється значення вектора параметрів (параметризація). Обрані параметри повинні мати такі властивості:

1. мінімізація обсягу інформації, необхідного для опису аудіо запису (за рахунок логарифмування);
2. некорельованість параметрів (за рахунок DCT);
3. однорідність параметрів, тобто однакова дисперсія в середньому;
4. можливість застосовувати прості метрики (Евклідова метрика) для визначення близькості між наборами параметрів;

Багато різних параметрів запропоновано в літературі [4, 5]. Зазначеним вище властивостям найкраще відповідають мел-частотні кепстральні коефіцієнти (MFCC – Mel Frequency Cepstral Coefficients), які вперше було запропоновано як параметри в системах розпізнавання мовлення та диктора [6], а в подальшому отримали широке використання в процесі інформаційного пошуку музики (MIR) [7, 8], зокрема, під час класифікації за жанрами, визначенні подібності аудіо тощо.

Обравши MFCC як параметри, ми отримуємо опис музичного твору у вигляді файлу з параметрами MFCC. Таким чином, якщо тривалість музичного твору в середньому складає 3хв (180с), частота дискретизації аудіо файлу – 44,1кГц, довжина фрейму – 20мс з перекриттям 0,5 фрейма, то маємо:

44100 відліків/с * 180с = 7 938 000 відліків

44100 відліків/с * 20мс/фрейм = 882 відліків/фрейм

7 938 000 відліків / 882 відліків/фрейм = 9000 фреймів

(або 180с / 20мс/фрейм = 9000фреймів)

Отже, середньостатистичний музичний твір характеризують, виходячи з наведених розрахунків, приблизно 8 млн. відліків або з врахуванням перекриття фреймів приблизно 18тис. фреймів (9000 фреймів*2–1), кожен з яких описується вектором параметрів MFCC розмірності 13. Тобто параметризація дозволяє зменшити кількість інформації, необхідної для опису музичного твору, в десятки разів:

$$\frac{7\ 938\ 000\ відліків}{17999\ MFCC\ \cdot 13} = \frac{7\,938\,000}{233\,987} \approx 34\ рази$$

Таким чином, коефіцієнти MFCC є компактним представленням спектральної обвідної, що під час розпізнавання музичного твору дозволяє успішно замінити мільйони відліків аудіо файлу.

Вибір методу порівняння невідомого музичного твору з шаблонами БД

Після визначення параметрів для опису музичних творів, необхідно перейти до етапу пошуку музичного твору серед шаблонів БД, тобто порівняння невідомого музичного твору з шаблонами та визначення шаблону, розходженням з яким буде мінімальним.

Для того, щоб ідентифікувати невідомий аудіо запис, необхідно мати критерій порівняння. Як правило, таким критерієм є відстань D. При цьому наголошуємо, що підхід порівняння за обраними параметрами має забезпечити високий рівень розрізнення власного шаблону $\tilde{\mathbf{X}}$ та шаблонів інших творів $\tilde{\mathbf{Y}}$. Тобто в результаті порівняння невідомого музичного твору $\mathbf{X}$ з шаблоном власного твору похибка між ними має бути мінімальною, і навпаки максимальною – для шаблонів інших творів:

$$\begin{cases} D(\mathbf{X}, \tilde{\mathbf{X}}) \to \min \\ D(\mathbf{X}, \tilde{\mathbf{Y}}) \to \max \end{cases} \Rightarrow D(\mathbf{X}, \tilde{\mathbf{X}}) << D(\mathbf{X}, \tilde{\mathbf{Y}}), \qquad (1)$$

де $\tilde{\mathbf{X}}$ – множина векторів параметрів шаблону власного музичного твору; $\tilde{\mathbf{Y}}$ – множина векторів параметрів шаблону іншого музичного твору.

Загальну схему ідентифікації музичного твору наведено на рисунку 1.

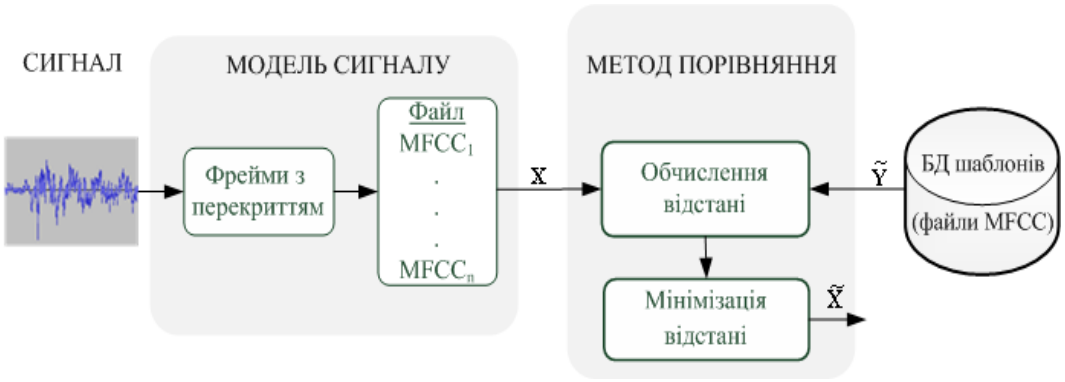

Рисунок 1 – Загальна схема ідентифікації музичного твору

Найпростішим і очевидним підходом для визначення близькості між наборами параметрів MFCC невідомого музичного твору та еталонів БД є порівняння MFCC на основі найбільш розповсюдженої Евклідової метрики, точніше квадрату Евклідової відстані $D_{Eu}{}^2$ (щоб надати велику вагу більш віддаленим об’єктам). Формула незваженої Евклідової відстані між вектором параметрів музичного твору, який треба ідентифікувати, $\mathbf{x} = (x_1, x_2, ..., x_d)$ та вектором шаблону БД $\tilde{\mathbf{y}} = (\tilde{y}_1, \tilde{y}_2, ..., \tilde{y}_d)$:

$$D_{Eu}{}^2(\mathbf{x}, \tilde{\mathbf{y}}) = \sum_{i=1}^{d} (x_i - \tilde{y}_i)^2. \tag{2}$$

Відповідно відстань між файлами параметрів невідомого твору та шаблону БД можна знайти за формулою:

$$D(\mathbf{X}, \tilde{\mathbf{Y}}) = \sum_{j=1}^{n} D_{Eu}{}^2(\mathbf{x}_j, \tilde{\mathbf{y}}_j) = \sum_{j=1}^{n} \sum_{i=1}^{d} (x_{ji} - \tilde{y}_{ji})^2. \tag{3}$$

В ідеальному випадку при такому підході відстань між файлами параметрів MFCC одного і того ж музичного твору буде рівна нулю, для різних творів – відмінною від нуля:

$$\begin{cases} D(\mathbf{X}, \tilde{\mathbf{X}}) = 0, \\ D(\mathbf{X}, \tilde{\mathbf{Y}}) > 0. \end{cases} \tag{4}$$

Проте варто зауважити, що навіть для одного і того ж музичного твору аудіо записи можуть відрізнятися, наприклад: на початку запису може йти тиша, мелодія іншого музичного твору тощо; записи можуть мати різний темп, тривалість. Це означає, що у разі зсуву фреймів в часі відстань до власного шаблону $D(\mathbf{X}, \tilde{\mathbf{X}}) \neq 0$, тобто умова чіткого розрізнення (1) не виконується, отже, безпосереднє порівняння файлів параметрів за Евклідовою відстанню не підходить для задачі ідентифікації власного шаблону. В цьому випадку придатним для ідентифікації музичного твору є алгоритм динамічної трансформації шкали часу (DTW - Dynamic Time Warping), який дозволяє працювати з послідовностями векторів, що мають певний зсув в часі [9]. Однак використання DTW призведе до зростання кількості операцій порівняння, що є неприйнятним.

Розрахуємо складність алгоритму DTW за кількістю операцій порівняння N_{op} фрагменту та шаблону. Якщо вважати, що тривалість шаблону T_{song} =3хв, що відповідає $\approx$18000фреймів, а тривалість фрагменту τ =5с, тобто $\approx$500фреймів, тоді:

$$N_{op} = (T_{song} - \tau) \cdot \tau;$$

$$N_{op} = (18000\ фр - 500\ фр) \cdot 500\ фр = 8\,750\,000 \approx 9 \cdot 10^6\ операцій\ .$$

Очевидно, що в більшості випадків музичний твір характеризується певною періодичністю, що полягає в наявності ідентичних або дуже схожих за текстом та характером мелодії фрагментів. Відповідно можна говорити про надлишковість даних, якими описується музичний твір, і можливість скоротити кількість параметрів для його опису. З огляду на це доцільним є застосування методів кластерного аналізу. Використання кластеризації для формування еталонів, що містяться в БД, дозволить зменшити обсяги пам'яті, необхідні для їх зберігання.

Задача кластеризації даних є важливим елементом загальної проблеми обробки даних. Кластеризацію часто використовують, зокрема, під час статистичного аналізу даних, векторної квантизації, розпізнавання образів тощо. Кластеризація — це поділ множини вхідних даних (в нашому випадку векторів параметрів MFCC) на групи (кластери) за мірою «схожості» один на одного, тобто таким чином, щоб кожен кластер містив найбільш схожі об’єкти, а об’єкти різних кластерів відрізнялися між собою. Задачу кластеризації можна сформулювати так: заданий набір з n векторів, кожен з яких має розмірність d, необхідно розбити на підмножини відповідно до заданого критерію оптимізації. Як правило, таким критерієм є мінімізація спотворення $e_i^2 \to \min$. Існують різні шляхи оцінювання спотворення, але в більшості прикладних реалізацій використовують суму середньоквадратичних Евклідових відстаней між центром кластеру (центроїдом) $\mathbf{c}_i$ і векторами параметрів, які до нього належать $\mathbf{X}_i = \{\mathbf{x}\}, \mathbf{X}_i \subset \mathbf{X}$ [10, 11], тобто:

$$e_i^2 = \{\mathbf{c}_i : \sum_{j=1}^{N_i} D_{Eu}{}^2(\mathbf{x}_j, \mathbf{c}_i) \mid \mathbf{x} \in \mathbf{X}_i \le \sum_{j=1}^{N_i} D_{Eu}{}^2(\mathbf{x}_j, \mathbf{c}) \mid \mathbf{x} \in \mathbf{X}_i\},$$

$$\mathbf{X}_i \subset \mathbf{X}, \forall \mathbf{c} \in \mathbf{X} \setminus \mathbf{X}_i, e_i^2 \to \min.$$

де N_i – кількість точок, що належать центроїду $\mathbf{c}_i$.

Таким чином, шаблони музичних творів в БД можна описати кластерами параметрів MFCC, показаних на рисунку 2 у вигляді кіл (для двомірного випадку).

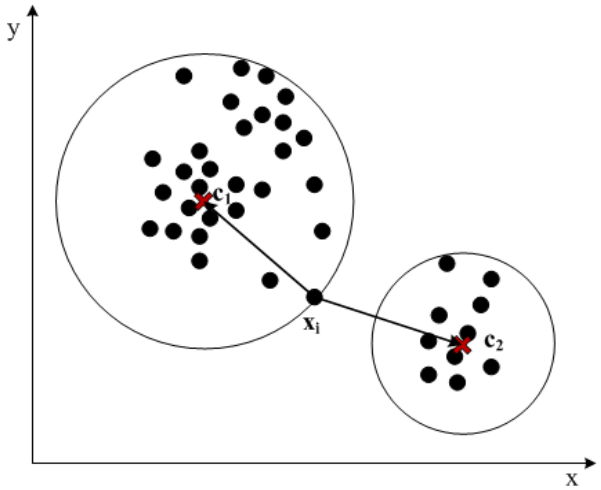

Рисунок 2 – Розбиття векторів параметрів на кластери

Кожен шаблон було представлено 1000 кластерів (замість 18 тис. векторів параметрів MFCC для твору тривалістю 3 хв.), тобто в середньому на кластер припадає близько 20 векторів. При подальшому збільшенні числа кластерів похибка кластеризації ще зменшиться, проте в такому випадку виникне ситуація, коли на кластер припадатиме менше 10 векторів, в результаті центроїди почнуть “підлаштовуватись” під конкретні вектори. Це призведе до погіршення результатів у випадку, якщо немає точного співпадіння між аудіо записами шуканого музичного твору та його власного шаблону. Зменшення кількості кластерів небажане, оскільки це призведе до зростання похибки кластеризації.

Відповідно, якщо, наприклад, шаблон представити 1000 кластерами, то обчислювальні затрати для порівняння його з фрагментом тривалістю τ=5с (500фреймів) становитимуть:

$$N_{op} = T_{song} \cdot \tau = 1000\ кл \cdot 500\ фр = 5 \cdot 10^5\ операцій\ .$$

Тобто такий підхід порівняно з DTW дозволяє зменшити кількість операцій порівняння в десятки разів.

Основні етапи порівняння файлів параметрів невідомого музичного твору з певним шаблоном БД:

1. Пошук мінімальної евклідової відстані $D\min^2$ між поточним вектором параметрів $\mathbf{x} = (x_1, x_2, ..., x_d)$ з множини параметрів $\mathbf{X} = \{\mathbf{x}\}, |\mathbf{X}| = n$ музичного твору, який треба ідентифікувати, та множиною векторів-кластерів $\widetilde{\mathbf{Y}} = \{\widetilde{\mathbf{y}}\}, \left|\widetilde{\mathbf{Y}}\right| = m$, $\widetilde{\mathbf{y}} = (\widetilde{y}_1, \widetilde{y}_2, ..., \widetilde{y}_d)$ шаблону:

$$D\min^2 = \underbrace{\min}_{j}(D_{Eu}{}^2(\mathbf{x}, \widetilde{\mathbf{y}}_j)) = \underbrace{\min}_{j}(\sum_{i=1}^{d}(x_i - \widetilde{y}_{ji})^2), j = \overline{1, m}. \tag{5}$$

2. Обчислення оцінки відстані в цілому до шаблону як суми квадратів мінімальних відстаней $D\min^2$:

$$D(\mathbf{X}, \widetilde{\mathbf{Y}}) = \sum_{l=1}^{n} D\min^2 = \sum_{l=1}^{n} \underbrace{\min}_{j}(\sum_{i=1}^{d}(x_i - \widetilde{y}_{ji})^2), j = \overline{1, m}. \tag{6}$$

Повертаючись до рисунку 2 з векторами параметрів MFCC, згрупованих в кластери, видно, що між кожним кластером та векторами, які належать до нього, є похибка e_i^2, що є сумою квадратів відстаней між ними. Звідси випливає, що відстань між файлами параметрів, що описують один і той же музичний твір до кластеризації і після (навіть якщо аудіо записи були ідентичними), буде додатною і рівною величині сумарної похибки кластеризації E^2. Таким чином, відстань до власного шаблону дорівнюватиме:

$$D(\mathbf{X}, \widetilde{\mathbf{X}}) = \sum_{l=1}^{n} D\min^2 = E^2 = \sum e_i{}^2, E^2 \to \min \tag{7}$$

Оцінювання похибки за приведеною власною відстанню

Очевидно, що оскільки твори мають різну тривалість, кожний твір характеризується власною кількістю фреймів, представлених параметрами MFCC. Під час кластеризації обчислюється однакова кількість кластерів для усіх шаблонів. Це призводить до того, що різні твори знаходяться в нерівних умовах, тобто для творів, тривалість яких більша, початкова похибка (між файлами того ж твору до і після кластеризації) теж буде більшою, оскільки в цьому випадку на кожен кластер буде припадати більше векторів параметрів. Позбутися цього можна за рахунок ділення відстані до власного шаблону БД, визначеної в формулі (7), на кількість фреймів музичного твору n. Назвемо цю величину приведеною власною відстанню ($D_{ПВ}$) музичного твору:

$$D(\mathbf{X}, \widetilde{\mathbf{X}}) = D_{ПВ} = \frac{\sum_{l=1}^{n} D\min^2}{n} = \frac{\sum_{l=1}^{n} \min(\sum_{i=1}^{d}(x_i - \widetilde{y}_{ji})^2)}{n} = \frac{E^2}{n}, j = \overline{1, m}, \tag{3.1-8}$$

Таким чином, $D_{ПВ}$ – по суті є математичним очікуванням (МО) похибки кластеризації:

$$D_{ПВ} = M_E = \frac{E^2}{n}. \tag{3.2-9}$$

Як можна побачити з формули (8) характеристика $D_{ПВ}$ не залежить від кількості фреймів (тривалості запису). Таким чином, її можна використовувати як критерій прийняття рішення як для запису в цілому, так і для його окремого фрагменту. Проте це твердження буде справедливим, тільки якщо для різних фрагментів ця характеристика буде змінюватися незначно, тобто за умови стаціонарності процесу.

Обґрунтування можливості ідентифікації музичного твору за його фрагментом

Змінення значень похибки, що виникає при порівнянні певного запису музичного твору з шаблоном, є випадковою функцією (процесом), що протікає в часі E(t). Значення похибки кожного музичного твору є окремими реалізаціями випадкової функції E(t).

Виходячи з самого принципу формування кластерів можна очікувати, що процес змінення значень похибки кластеризації в часі буде носити стаціонарний характер. Проте це припущення потребує статистичної перевірки.

Згідно з визначенням стаціонарним процесом є такий, значення якого незалежно від часу коливаються біля їх середнього значення (оцінка МО). Відповідно у разі підтвердження стаціонарності похибка повинна бути більш-менш постійною протягом усього музичного твору, це, в свою чергу, є підставою для ідентифікації музичного твору за $D_{ПВ}$, що є МО похибки кластеризації, на основі фрагменту, причому обраному незалежно від проміжку часу.

Отже, в формалізованому вигляді випадкова функція E(t) називається стаціонарною, якщо всі її ймовірнісні характеристики не залежать від часу t: математичне очікування $m_E(t)$, кореляційна функція $K_E(t, t+r)$ (включає дисперсію $D_E(t)$).

$$m_E(t) = M_E = \text{const};$$
$$K_E(t, t+r) = K_E(r), D_E(t) = K_E(t,t) = k_E(0) = \text{const}. \quad (4.1\text{-}10)$$

Перевірка (10) здійснювалась для 20 музичних творів. Результати наведені для 4 з них. Проаналізуємо отримані дані з точки зору ймовірної стаціонарності похибки кластеризації E(t). Математичне очікування $m_E(t)$ на фрагментах τ=15с відхиляється від математичного очікування, розрахованого для всього музичного твору $\tilde{M}_E$, $\tilde{m}_{E_{min}} \le \tilde{M}_E \le \tilde{m}_{E_{max}}$, незначно, наприклад: 1 – $0{,}28 \le 0{,}31 \le 0{,}32$; 2 – $0{,}36 \le 0{,}38 \le 0{,}40$; 3 – $0{,}40 \le 0{,}45 \le 0{,}48$; 4 – $037 \le 0{,}43 \le 0{,}50$. Графіки кореляційної функції (показаної на рис. 3), отримані для фрагментів з початку, середини і кінця музичного твору (а саме другого, дев'ятого і шістнадцятого фрагментів по 15с), мають подібний характер.

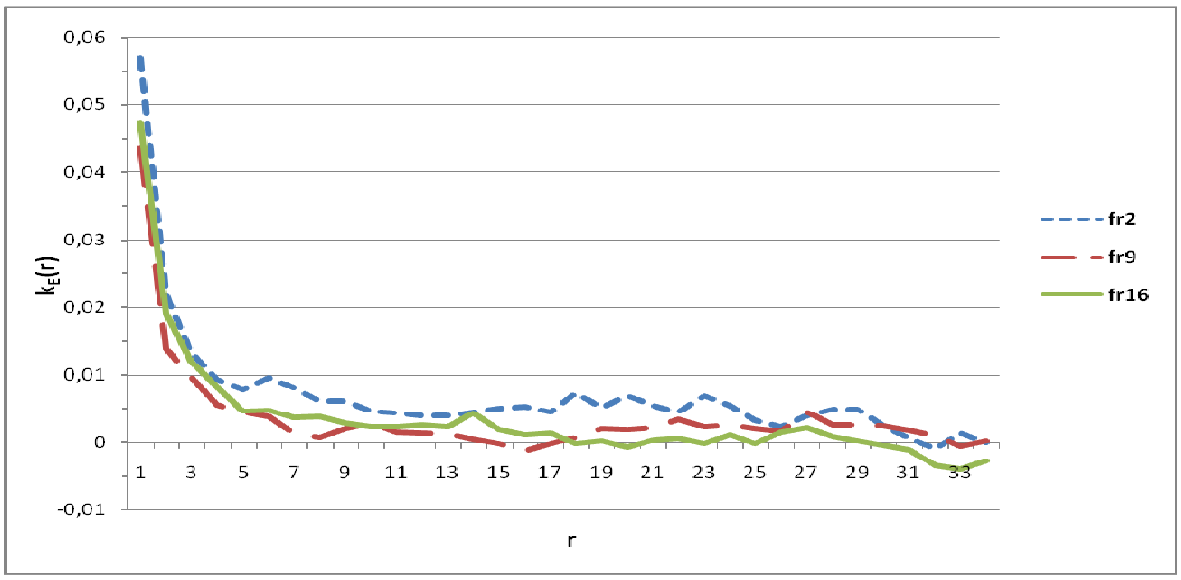

Рисунок 3 – Кореляційна функція для другого (fr2), дев'ятого (fr9) і шістнадцятого (fr16) фрагментів по 15с музичного твору

Таким чином, можна говорити про те, що припущення про стаціонарний характер процесу зміни значень похибки кластеризації E(t) підтверджено, відповідно ідентифікацію невідомого музичного твору доцільно здійснювати за його фрагментом.

Слід підкреслити, що особливо важливим є МО похибки M_E, оскільки ця характеристика є визначальною для ідентифікації музичного твору (згідно з формулами (3.1), (3.2)).

Хоча отримані статистичні дані підтвердили стаціонарність процесу зміни похибки (в результаті кластеризації), значення МО від фрагменту до фрагменту $\tilde{m}_E(t_i), i = \tau, 2\tau, \ldots, T_{song}$ все ж таки дещо коливаються відносно МО в цілому для музичного твору $\tilde{M}_E = D_{ПВ}$ в межах від $\tilde{m}_{E_{min}} = \min(\tilde{m}_E(t_i)), i = \tau, 2\tau, \ldots, T_{song}$ до $\tilde{m}_{E_{max}} = \max(\tilde{m}_E(t_i)), i = \tau, 2\tau, \ldots, T_{song}$:

$$\tilde{m}_{E_{min}} \le \tilde{M}_E \le \tilde{m}_{E_{max}}, \text{ або } \tilde{m}_{E_{min}} \le D_{ПВ} \le \tilde{m}_{E_{max}}.$$

Для характеристики цих коливань введемо коефіцієнти $k_{\max\tau}, k_{\min\tau}$, які розраховуються для кожного з еталонів музичних творів, що містяться в БД, за формулами:

$$k_{\max\tau} = (\tilde{m}_{E_{max}} - \tilde{M}_E)/\tilde{M}_E = (\tilde{m}_{E_{max}} - D_{ПВ})/D_{ПВ},$$
$$k_{\min\tau} = (\tilde{m}_{E_{min}} - \tilde{M}_E)/\tilde{M}_E = (\tilde{m}_{E_{min}} - D_{ПВ})/D_{ПВ}. \quad (4.2\text{-}11)$$

Виходячи з цього сформулюємо умову, згідно з якою можна ідентифікувати невідомий фрагмент музичного твору, тобто визначити шаблон БД, що є його власним. Фрагмент аудіо запису можна вважати розпізнаним, якщо для відстані між фрагментом та певним шаблоном БД (що відповідно є власним шаблоном фрагмента) виконується нерівність:

$$(1 + k_{\min\tau}) \cdot D_{ПВ} \le D(\mathbf{X}, \tilde{\mathbf{X}}) \le (1 + k_{\max\tau}) \cdot D_{ПВ} \quad (4.3\text{-}12)$$

Відзначимо, що відповідно до формул (11) та (12), на основі яких має здійснюватись ідентифікація фрагмента музичного твору в БД музики, крім самих шаблонів музичних творів, для кожного музичного твору БД також мають зберігатись значення $\tilde{m}_{E_{min}}, \tilde{m}_{E_{max}}$ та $D_{ПВ}$.

Мінімальна і достатня тривалість фрагменту

В розділі 4 було доведено можливість ідентифікувати музичний твір за його фрагментом, зокрема тривалістю 15с. Зменшення тривалості фрагменту може зумовити зростання ймовірності помилки під час ідентифікації музичного твору. Однак крім надійності пошуку, велике значення має швидкість пошуку. Тому важливо, щоб тривалість фрагменту була якомога меншою. Відповідно важливою задачею є визначення мінімальної тривалості фрагменту, що дозволяє ідентифікувати музичний твір за шаблоном.

З цією метою було проведено дослідження, аналогічні описаним у розділі 4, на фрагментах тривалістю 1с та 5с.

В розділі 3 було введено поняття приведеної власної відстані музичного твору $D_{ПВ}$, на основі якої здійснюється ідентифікація фрагменту музичного твору за шаблоном згідно з формулою (12). Нагадуємо, що приведена власна відстань є МО похибки кластеризації ($D_{ПВ} = \tilde{M}_E$) відповідно до формул (8) та (9).

На рис. 4 показано процес зміни в часі значень МО похибки на фрагментах 1с, 5с та 15с. В таблиці 1 та на рис. 5 на прикладі 4 пісень показано результати для мінімального $\tilde{m}_{E_{min}}$ та максимального $\tilde{m}_{E_{max}}$ значення МО похибки на фрагментах тривалістю 1, 5, 15с та усереднене значення МО в цілому для музичного твору $\tilde{M}_E$.

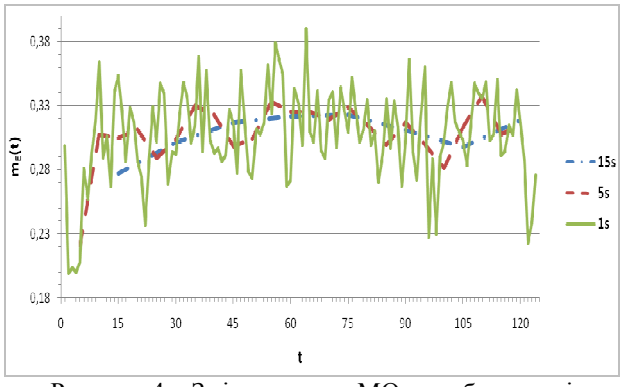

Рисунок 4 – Зміна значень МО похибки в часі для 1,5 та 15с

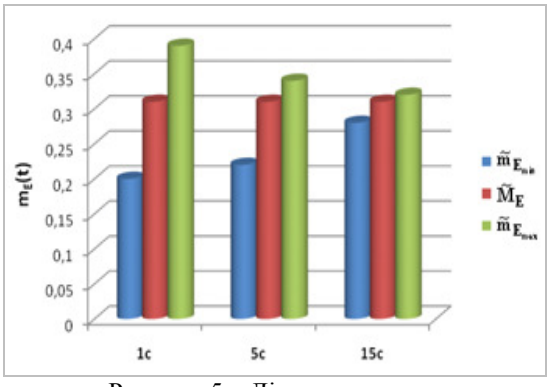

Рисунок 5 – Діапазон коливань $\tilde{m}_E(t_i), i = \tau, 2\tau, \ldots, T_{song}$ залежно від τ

Таблиця 1.Коливання МО похибки для фрагментів тривалістю 1, 5 та 15с

№	1	2	3	4
τ	1с			
$\widetilde{M}_E$	**0,31**	**0,38**	**0,44**	**0,43**
$\widetilde{m}_{E_{min}}$	0,20	0,31	0,34	0,32
$\widetilde{m}_{E_{max}}$	0,39	0,47	0,59	0,59
τ	5с			
$\widetilde{M}_E$	**0,31**	**0,38**	**0,45**	**0,43**
$\widetilde{m}_{E_{min}}$	0,22	0,34	0,37	0,35
$\widetilde{m}_{E_{max}}$	0,34	0,41	0,50	0,52
τ	15с			
$\widetilde{M}_E$	**0,31**	**0,38**	**0,45**	**0,43**
$\widetilde{m}_{E_{min}}$	0,28	0,36	0,40	0,37
$\widetilde{m}_{E_{max}}$	0,32	0,40	0,48	0,50

Отримані результати свідчать, що для усіх наведених тривалостей τ значення МО від фрагменту до фрагменту $\widetilde{m}_E(t_i), i = \tau, 2\tau, \ldots, T_{song}$ коливаються незначно відносно $\widetilde{M}_E$, тобто стаціонарність зберігається.

Однак, як правило, випадковий процес починається з нестаціонарної стадії (перехідний процес), після чого процес можна вважати стаціонарним. Виходячи з цього небажано давати для розпізнавання фрагмент на початку або в кінці музичного твору.

З таблиці 1 та рисунку 5 також можна бачити, що зі зменшенням тривалості фрагменту τ коливання $\widetilde{m}_E(t_i)$ стають більшими, тобто значення коефіцієнта k_τ зростає.

Значне зростання коефіцієнта k_τ зі зменшенням τ фрагменту, особливо за умов великої кількості музичних творів в БД, може призвести до виникнення і швидкого зростання кількості випадків неправильного прийняття рішення, коли інший шаблон музичного твору буде прийнятий за власний. Таким чином, враховуючи отримані результати та зростання рівня коливань для фрагментів тривалістю τ=1с, що виражає коефіцієнт k_τ, через недостатню кількість статистичних даних для висновків, було обрано мінімальну тривалість фрагменту – τ=5с для ідентифікації музичного твору за шаблоном.

Перевірка відповідності теоретичних припущень та експериментальних результатів

Для підтвердження теоретичних положень, наведених в попередніх розділах, було проведено експериментальні дослідження на базі 1000 музичних творів. Всі музичні твори мали формат wav (mono) з частотою дискретизації 44,1кГц. Попередньо з аудіо файлів було видалено тишу з початку та кінця записів. В процесі формування БД аудіо записи шаблонів необхідно було:

1. поділити на фрейми по 20мс з перекриттям 10мс;
2. для кожного фрейму розрахувати вектор параметрів MFCC розмірності 13.

Для отримання векторів параметрів MFCC з аудіо записів використовувався набір інструментів Матлаб – MIRtoolbox, що є відкритим ПЗ і описаний в [12].

Послідовності векторів параметрів MFCC, що описують музичні твори, було кластеризовано, використовуючи вдосконалений метод кластеризації k-середніх запропонований у одній з попередніх робіт [13]. В результаті чого кожен шаблон БД було представлено 1000 кластерів MFCC.

Одночасно в процесі кластеризаціїї для кожного з 1000 музичних творів БД було визначено значення $\widetilde{m}_{E_{min}}, \widetilde{m}_{E_{max}}$ та $D_{ПВ}$, на основі яких має здійснюватись ідентифікація фрагмента музичного твору в БД музики згідно з формулою (12).

З 1000 еталонів БД випадковим чином було обрано фрагменти 100 пісень для їх ідентифікації. Результати, отримані в процесі порівняння 1000 шаблонів музичних творів БД з 100 фрагментів, які необхідно ідентифікувати, наведено в таблиці 2.

Таблиця 2. Результати ідентифікації фрагменту за шаблоном БД

№ фрагменту	№ шаблону	Тривалість фрагменту τ								
		1с			5с			15с		
		$D(\mathbf{X},\tilde{\mathbf{X}})$	$[\tilde{m}^{(\tilde{\mathbf{X}})}_{\mathrm{Emin}}, \tilde{m}^{(\tilde{\mathbf{X}})}_{\mathrm{Emax}}]$	$D_{\min}(\mathbf{X},\tilde{\mathbf{Y}})$	$D(\mathbf{X},\tilde{\mathbf{X}})$	$[\tilde{m}^{(\tilde{\mathbf{X}})}_{\mathrm{Emin}}, \tilde{m}^{(\tilde{\mathbf{X}})}_{\mathrm{Emax}}]$	$D_{\min}(\mathbf{X},\tilde{\mathbf{Y}})$	$D(\mathbf{X},\tilde{\mathbf{X}})$	$[\tilde{m}^{(\tilde{\mathbf{X}})}_{\mathrm{Emin}}, \tilde{m}^{(\tilde{\mathbf{X}})}_{\mathrm{Emax}}]$	$D_{\min}(\mathbf{X},\tilde{\mathbf{Y}})$
1	**83**	0,408	[0,219; 0,510]	0,561	0,360	[0,255; 0,474]	0,543	0,361	[0,291; 0,437]	0,588
2	**96**	0,512	[0,285; 0,666]	1,026	0,495	[0,333; 0,618]	1,173	0,496	[0,380; 0,571]	1,215
3	**165**	0,441	[0,258; 0,601]	0,554	0,439	[0,301; 0,558]	0,638	0,422	[0,344; 0,515]	0,567
4	**200**	0,437	[0,275; 0,641]	0,854	0,411	[0,320; 0,595]	0,772	0,466	[0,366; 0,549]	0,897
5	**213**	0,400	[0,237; 0,553]	0,632	0,386	[0,276; 0,513]	0,683	0,386	[0,316; 0,474]	0,692
6	**268**	0,465	[0,269; 0,628]	0,646	0,437	[0,314; 0,583]	0,737	0,459	[0,359; 0,538]	0,894
7	**403**	0,483	[0,279; 0,651]	0,587	0,454	[0,325; 0,604]	0,692	0,450	[0,372; 0,558]	0,650
8	**408**	0,397	[0,226; 0,527]	0,571	0,372	[0,264; 0,490]	0,621	0,372	[0,301; 0,452]	0,592
9	**523**	0,325	[0,179; 0,418]	0,470	0,301	[0,209; 0,388]	0,450	0,297	[0,239; 0,358]	0,466
10	**525**	0,480	[0,252; 0,587]	0,711	0,452	[0,294; 0,545]	0,703	0,419	[0,336; 0,503]	0,743
11	**2**	0,420	[0,285; 0,666]	0,650	0,453	[0,333; 0,618]	0,765	0,492	[0,380; 0,571]	0,802
12	**64**	0,315	[0,199; 0,465]	0,475	0,324	[0,232; 0,431]	0,477	0,318	[0,266; 0,398]	0,503
13	**225**	0,446	[0,283; 0,660]	0,695	0,426	[0,330; 0,613]	0,659	0,440	[0,377; 0,566]	0,746
14	**600**	0,240	[0,144; 0,335]	0,400	0,227	[0,168; 0,311]	0,406	0,228	[0,192; 0,287]	0,422
15	**50**	0,362	[0,226; 0,527]	0,735	0,392	[0,263; 0,489]	0,709	0,364	[0,301; 0,451]	0,630

В таблиці показано результати ідентифікації за шаблоном фрагментів 15 музичних творів для тривалостей 1, 5 та 15с. Зокрема в таблиці наведено відстань до власного шаблону $D(\mathbf{X},\tilde{\mathbf{X}})$ та відповідні значення $\tilde{m}^{\tilde{\mathbf{X}}}_{E_{\min}}, \tilde{m}^{\tilde{\mathbf{X}}}_{E_{\max}}$, а також відстань до найближчого шаблону іншого музичного твору $D_{\min}(\mathbf{X},\tilde{\mathbf{Y}})$.

Фрагмент вважається розпізнаним, тобто визначено власний шаблон $\tilde{\mathbf{X}}$ фрагменту, якщо для певного шаблону БД $\tilde{\mathbf{Y}}$ виконується $D(\mathbf{X},\tilde{\mathbf{Y}}) \in \left[\tilde{m}^{(\tilde{\mathbf{Y}})}_{E_{\min}};\ \tilde{m}^{(\tilde{\mathbf{Y}})}_{E_{\max}}\right]$ відповідно до формули (12). Однак у випадку з фрагментом №3 таблиці тривалістю 1с було виявлено декілька шаблонів БД, для яких виконується (12): №159 – $0{,}554 \in [0{,}254;\ 0{,}592]$, №165 – $0{,}441 \in [0{,}258;\ 0{,}601]$. Це означає потенційну можливість неправильного прийняття рішення під час ідентифікації фрагменту тривалістю 1с, коли інший шаблон БД приймається за власний. Очевидно, що розходження з власним шаблоном повинно бути мінімальним, тобто власним все ж таки є шаблон під номером №165 з відстанню $D(\mathbf{X},\tilde{\mathbf{X}}) = 0{,}441$, тому помилки можна уникнути.

Висновки

В роботі теоретично обґрунтовано можливість ідентифікації музичного твору за його фрагментом на основі приведеної власної відстані (МО похибки кластеризації), значення якої не залежить від кількості фреймів (тривалості запису). Запропоновано аналітичний вираз для визначення власного шаблону музичного твору на основі його фрагменту. Визначено мінімальну тривалість фрагменту (5с), що дозволяє зменшити складність обчислень в десятки разів в процесі автоматизованої ідентифікації музичного твору. Проведені експериментальні дослідження підтвердили справедливість наведених теоретичних положень.

Літературні джерела

1. Cano P. A review of audio fingerprinting / P. Cano, E. Batlle, T. Kalker, and J. Haitsma // Journal of VLSI Signal Processing. – 2005. – no. 41. – pp. 271–284.
2. Downie J.S. Music information retrieval / J.S. Downie // Annual Review of Information Science and Technology. – 2003. – no. 37. – pp. 295–340.
3. Hoashi K. Personalization of user profiles for content-based music retrieval based on relevance Feedback / K. Hoashi, K. Matsumoto, and N. Inoue // Proceedings of the ACM International Conference on Multimedia. – 2003. – pp. 110–119.

4. Wang Y. Multimedia content analysis using both audio and visual cues / Y. Wang, Z. Liu, and J. C. Huang // IEEE signal processing magazine. – 2000. – no. 17. – pp. 12–36.

5. Grosche P. Audio content-based music retrieval / P. Grosche, M. Müller, J. Serrà // Dagstuhl Follow-Ups Multimodal Music Processing. – V. 3. – Dagstuhl, Germany. – 2012. – pp. 157–175.

6. Ganchev T. Comparative evaluation of various mfcc implementations on the speaker verification task / T. Ganchev, N. Fakotakis, and G. Kokkinakis // Proceedings of 9th International Conference on Speech and Computer, SPECOM'05. – 2005. – pp. 191–194.

7. Logan B. A music similarity function based on signal analysis / B. Logan and A. Salomon // Proc. IEEE Int. Conf. Multimedia Expo. – 2001. – pp. 745–748.

8. Tzanetakis G. Musical genre classification of audio signals / G. Tzanetakis and P. Cook // IEEE Trans. Speech Audio Process. – No. 5. – V. 10. – 2002. – pp. 293–301.

9. Senin P. Dynamic time warping algorithm review / P. Senin – Honolulu, USA. – 2008.

10. Gersho A. Vector Quantization and Signal Compression. / A. Gersho, R. M. Gray. – Boston: Kluwer Academic. – 1992. – 760 p.

11. Jain A. K. Algorithms for Clustering Data / A. K. Jain, R. C. Dubes. – Englewood Cliffs, N.J.: Prentice Hall. – 1988. – 334 p.

12. Ткаченко О.М. Метод кластеризації на основі послідовного запуску k-середніх з удосконаленим вибором кандидата на нову позицію вставки / О. М. Ткаченко, О. Ф. Грійо Тукало, О. В. Дзісь, С. М. Лаховець // Електронний журнал «Наукові праці ВНТУ». – №2. – В.2. – Вінниця: ВНТУ. – 2012.

13. Lartillot O. A Matlab Toolbox for Musical Feature Extraction From Audio / O. Lartillot, P. Toiviainen // International Conference on Digital Audio Effects. – Bordeaux, France. – 2007.

Інформація про авторів

Ткаченко Олександр Миколайович – к.т.н., доцент кафедри обчислювальної техніки, Вінницький національний технічний університет, Хмельницьке шосе, 95, м. Вінниця, alextk1960@gmail.com.

Грійо Тукало Оксана Франсисківна – аспірант кафедри обчислювальної техніки, Вінницький національний технічний університет, Хмельницьке шосе, 95, м. Вінниця, xxmargoxx@gmail.com.