IMARE JAKO EFEKTYWNE NARZĘDZIE DO TRANSKRYPCJI MUZYKI

Piotr Różański, Marcin Radoszewski, Mariusz Tycz, Maciej Włoch, Bartosz Zasada

Nazwa instytucji: Uniwersytet Mikołaja Kopernika w Toruniu, Wydział Matematyki i Informatyki
Opiekun naukowy: mgr Michał Dudkiewicz

Streszczenie:

Przedstawiona została efektywna metoda transkrypcji muzyki, opracowana w ramach zespołowego projektu studenckiego na Uniwersytecie Mikołaja Kopernika. Szczegółowo omówione zostały kolejne etapy przetwarzania, zastosowane metody numeryczne oraz szczegóły implementacji. Nakreślony został także końcowy efekt projektu (program komputerowy IMARE) i otrzymane przy jego pomocy wyniki na tle innych, często komercyjnych rozwiązań.

Słowa kluczowe: przetwarzanie sygnałów, transkrypcja dźwięku, analiza widmowa, zapis nutowy

1. Wstęp

Transkrypcja muzyki jest zagadnieniem bardzo skomplikowanym, o czym świadczyć może to, że nawet ludzki zmysł słuchu, wykształcony przez miliony lat ewolucji w celu rozpoznawania zagrożeń i korzyści pochodzących ze środowiska, bez odpowiedniego kształcenia muzycznego bardzo słabo radzi sobie z tym zagadnieniem. Dopiero po długotrwałym treningu (choć także nie wszyscy) jesteśmy w stanie dokonać w miarę dokładnej transkrypcji, czyli przetworzyć muzykę, którą słyszymy, na czytelny dla wszystkich zapis nutowy.

Problem automatycznej transkrypcji muzyki podjęty został w ramach projektu studenckiego z Programowania Zespołowego na Uniwersytecie Mikołaja Kopernika w roku akademickim 2009/2010. W skład zespołu wchodzili: Mariusz Tycz (kierownik), Bartosz Zasada (sekretarz), Marcin Radoszewski, Piotr Różański i Maciej Włoch (programiści). Opiekunem naukowym zespołu był mgr Michał Dudkiewicz.

2. Materiał i metody

Materiał wykorzystywany do testów miał postać cyfrowych plików dźwiękowych, zarówno o wysokiej jakości (muzyczne płyty kompaktowe), jak i o jakości znacznie gorszej (utwory po kompresji MP3). W przypadku plików stereofonicznych, pierwszym etapem przetwarzania była redukcja do sygnału monofonicznego, więc opracowana heurystyka dotyczy *de facto* muzyki monofonicznej.

Sygnały dźwiękowe w postaci cyfrowej (PCM), będące podmiotem badań, mają postać ciągu liczb, przedstawiających wartości chwilowe sygnału w dobrze określonych odstępach czasu. Dokładność przedstawienia kolejnych takich liczb określa tzw. kwantyzacja sygnału, wynosząca najczęściej 8, 16 (dla jakości CD) lub 24 bity. Kwantyzacja jest wprost związana z jakością odwzorowania dźwięku analogowego. Drugim parametrem sygnału jest częstotliwość próbkowania (44100 Hz dla jakości CD), odwrotnie proporcjonalna do odstępu czasowego pomiędzy dwoma sąsiednimi próbkami. Ten parametr z kolei związany jest z możliwym do odwzorowania zakresem częstotliwości. Połowa częstotliwości próbkowania, zwana częstotliwością Nyquista, określa maksymalną częstotliwość składową sygnału analogowego, która może być odwzorowana w sygnale cyfrowym.

a) Wyznaczenie widma dźwięku

Podstawowym narzędziem wykorzystywanym do analizy dźwięku jest zwykle Transformacja Fouriera. Przekształca ona czasowy przebieg sygnału (X_n) w funkcję określoną na częstotliwościach, wyrażającą natężenia i fazy poszczególnych składowych (T_k). Funkcję taką określa się mianem *widma* lub *transformaty* dźwięku. Transformacja taka ma bardzo dobrze określoną postać matematyczną i może zostać obliczona jako:

$$TF_k = \frac{1}{N} \sum_{n=0}^{N-1} X_n e^{-2\pi i \frac{kn}{N}}.$$

Oczywiście, obliczanie transformaty Fouriera przy pomocy powyższego wzoru jest nieefektywne (wymaga N^2 operacji), zatem zazwyczaj wykorzystuje się algorytm Szybkiej Transformacji Fouriera, który do wyznaczenia tego samego wyniku potrzebuje tylko $N \log N$ operacji (z dokładnością do stałego czynnika).

Warto w tym momencie zauważyć, że w zagadnieniu analizy muzyki potrzebna jest nie tyle informacja o widmie uzyskanym z całego przebiegu sygnału, lecz raczej o zmienności widma w czasie. Należy zatem wyznaczyć wiele transformat Fouriera w określonych odstępach czasu. Odstępy te mogą być zgodne z częstotliwością próbkowania, ale z uwagi na efektywność zazwyczaj wybiera się odstęp czasowy wielokrotnie większy. Należy także ustalić szerokość okna (N), czyli ilość próbek branych za każdym razem do obliczenia widma. Formalnie, transformację tą określa się, mnożąc wartości sygnału przez odpowiednio przesuniętą tzw. *funkcję okna* $W(x)$, przyjmującą niezerowe wartości tylko w żądanym przedziale w okolicy zera. W zależności od doboru funkcji okna, można znacząco zredukować niepożądane efekty brzegowe (tzw. aliasing). Otrzymana transformata określana jest

angielską nazwą *Short-Time Fourier Transform* i wyraża się następująco:

$$STFT_k(n_0) = \frac{1}{N} \sum_{n=0}^{N-1} W(n-n_0)\, X_n e^{-2\pi i \frac{kn}{N}} .$$

Niestety, w przypadku, gdy interesuje nas zapis nutowy, klasyczna transformacja Fouriera nie jest odpowiednim rozwiązaniem. Dzieje się tak, ponieważ skala muzyczna, na której odwzorować chcemy sygnał, charakteryzuje się tym, że częstotliwości kolejnych dźwięków skali wzrastają nie liniowo, a wykładniczo. W związku z tym, w przypadku użycia transformaty Fouriera, duża część skali muzycznej, odpowiadająca niższym częstotliwościom, odwzorowana zostałaby na bardzo wąskim zakresie widma, co katastrofalnie wpłynęłoby na jakość odwzorowania niskich częstotliwości.

Zamiast transformacji Fouriera zastosowano zatem spokrewnioną z nią transformację, określaną mianem Constant Q [Brown 1991]. W tym przypadku, otrzymywana skala częstotliwości okazuje się być skalą liniową. Co więcej, zakładana jest stała jakość odwzorowania dla wszystkich częstotliwości z badanego przedziału. Naturalną konsekwencją tego założenia jest to, że dla różnych częstotliwości różna będzie także szerokość okna (*N*). Tytułowy parametr tej transformaty, czyli *Q*, określa jakość odwzorowania i jest związany z rozdzielczością spektralną otrzymanego widma. W celu transkrypcji muzyki europejskiej, opartej o system dwunastopółtonowy, parametr *Q* musi wynosić co najmniej 17.

Transformacja Constant Q w postaci klasycznej jest, podobnie jak klasyczna Transformacja Fouriera, bardzo kosztowna obliczeniowo. Na szczęście, jak pokazano w pracy [Brown, Puckette 1992], można zaimplementować transformację Constant Q przy użyciu Szybkiej Transformacji Fouriera, wykorzystując obliczanie splotu transformaty z odpowiednio dobranymi funkcjami, tzw. *kernel functions*. Implementacja tej metody pozwoliła na znaczące skrócenie czasu obliczeń.

b) Analiza widmowa

Otrzymana transformata sygnału, wyznaczona dla wielu momentów czasu, może zostać poddana dalszej analizie, w celu określenia zawartych w sygnale dźwięków. Pierwszym krokiem tego etapu analizy jest oszacowanie, dla każdej transformaty, poziomu szumu (mediany wszystkich wartości z widma), na którego podstawie wyznaczony zostanie próg detekcji. Odczytuje się następnie wszystkie maksima lokalne widma, które przekraczają próg detekcji, a zarazem znajdują się na liście maksimów widma sąsiedniego (w czasie). Tylko maksima spełniające zarazem dwa powyższe warunku przekazywane są do dalszej analizy. W ten sposób eliminowane są artefakty numeryczne lub będące wynikiem szumu w sygnale wejściowym.

Na podstawie położenia danego maksimum w widmie można, z dokładnością do rozdzielczości widma, oszacować jego częstotliwość. Dokładniejsza metoda określania rozdzielczości składowej opiera się o mechanizm znany jako „phase vocoder" [Portnoff 1976], stosowany szeroko w profesjonalnych systemach edycji dźwięku. Metoda ta opiera się na badaniu różnic fazy pomiędzy dwoma odpowiadającymi sobie maksimami w sąsiadujących widmach. Dzięki wykorzystaniu tej metody, można także odrzucić te maksima, dla których częstotliwości wyznaczona dwiema metodami nie będą zgodne. Pozwala to na jeszcze dokładniejszą redukcję szumu, przy jednoczesnym zachowaniu wyraźnych, choć słabych dźwięków.

W miarę przetwarzania kolejnych obrazów widma, znalezione w nich maksima łączone są w tony, zawierające powiązane czasowo dźwięki o zbliżonej częstotliwości. Przy dodawaniu kolejnego maksimum do przebiegu tonu, algorytm ocenia, czy stanowi on jego konfigurację, czy raczej repetycję, czyli powtórzenie dźwięku na tej samej wysokości.

c) Redukcja tonów składowych

Jak się okazuje, tony otrzymane na drodze powyższej analizy, choć rzeczywiście obecne w widmie, nie stanowią jeszcze zbioru postrzeganych dźwięków. Widma rzeczywistych instrumentów, nawet pochodzące od pojedynczych dźwięków, zawierają cały szereg częstotliwości stanowiących całkowite wielokrotności częstotliwości podstawowej. W przypadku muzyki polifonicznej problem ten staje się jeszcze bardziej złożony, ponieważ, jak się okazuje, wiele współbrzmień o różnej strukturze dźwięków daje identyczną strukturę tonów składowych, choć z różnymi natężeniami.

Dla każdego wykrytego maksimum wyznaczany jest najlepszy kandydat na jego częstotliwość podstawową. Pod uwagę brane są przy tym zarówno pokrewieństwa częstotliwości maksimów, jak również stosunki ich amplitud. Na podstawie wyznaczonych powiązań, po przeanalizowaniu całego utworu, odtwarzane są możliwe powiązania pomiędzy równoczesnymi tonami. W przypadku wielu możliwości, właściwe powiązanie wybierane jest metodą „głosowania" wszystkich maksimów wchodzących w skład tonu. Jeśli jedna z możliwości otrzyma co najmniej 50 procent głosów, tony oznaczane są jako powiązane. Spośród każdego zbioru powiązanych ze sobą tonów jako ton podstawowy wybierany jest ton o najniższej częstotliwości. Tylko takie tony zostaną uwzględnione w finalnym zapisie nutowym, a do ich natężenie dodawane będą sumy natężeń wszystkich dźwięków z nimi powiązanych.

d) Wybór tonacji muzycznej

Dla każdej z możliwych tonacji muzycznych wyznaczana jest ilość dźwięków gamowłaściwych pośród wszystkich dźwięków znalezionych w utworze. Pod uwagę brane jest 12 tonacji durowych oraz 12 tonacji molowych (moll harmoniczny). Wybierana jest tonacja, dla której ilość dźwięków gamowłaściwych jest największa. W przypadku remisu, wybierana jest tonacja o mniejszej liczbie znaków przykluczowych.

3. Wyniki i dyskusja

Na podstawie opisanej powyżej heurystyki, stworzony został program o nazwie IMARE. Aplikacja posiada graficzny interfejs użytkownika, zaś napisana została w języku Java, można ją zatem uruchomić na każdym ze współczesnych systemów operacyjnych. Program wczytuje zarówno pliki dźwiękowe w formacie WAVE (PCM) jak i skompresowane pliki dźwiękowe w formacie MP3. Możliwe jest skorzystanie z ustawień domyślnych, lub też ustalenie wartości parametrów wykorzystywanych w procesie transkrypcji. Program generuje zapis nutowy w konwencji fortepianowej (dwie pięciolinie, klucz wiolinowy i basowy), który można wyświetlić, odsłuchać i/lub zapisać jako utwór w formacie MIDI. Aplikacja posiada także możliwość transkrypcji muzyki w czasie rzeczywistym (z wykorzystaniem zewnętrznego źródła dźwięku), przy użyciu nieco zmienionego algorytmu.

Pierwsza część analizy, czyli wyznaczanie widma dźwięku i maksimów w tym widmie, została zaprogramowana współbieżnie w sposób trywialny, polegający na rozdzieleniu całego pliku dźwiękowego na określoną ilość segmentów i przypisanie każdemu segmentowi osobnego wątku przeprowadzającego na nim analizę. Po ukończeniu pracy przez wszystkie wątki, dane są zbierane i pozostała część analizy wykonywana jest już tylko przez jeden wątek. Równoległe wykonywanie pozostałych etapów analizy nie jest konieczne, gdyż stanowią one bardzo niewielki ułamek całkowitego czasu pracy.

Przeprowadzono dużą liczbę testów na plikach muzycznych różnej jakości, a także reprezentujących różne gatunki muzyczne. Porównanie z wieloma istniejącymi (często komercyjnymi) rozwiązaniami ukazało wyraźną przewagę programu IMARE ze względu na dokładność transkrypcji, brak nadmiarowych nut (artefaktów) oraz jakość wyników dla ustawień domyślnych. Również algorytm wyboru tonacji, choć koncepcyjnie bardzo prosty, dawał niespodziewanie dobre efekty dla testowanych plików dźwiękowych.

Zaimplementowany sposób wskazywania tonów podstawowych opiera się na założeniu, że ton podstawowy jest najsilniejszym dźwiękiem spośród tonów występujących w szeregu harmonicznym. Prawdą jest, że dla niektórych instrumentów nie jest to spełnione; w rzadkich przypadkach ton podstawowy nie występuje w ogóle. Niemniej jednak, dla większości próbek dźwiękowych założenie to daje dobre rezultaty. Co więcej, można przyjąć, że zadaniem programu nie jest odtworzenie pierwotnego zapisu muzycznego, ale stworzenie takiej transkrypcji na określony instrument, która pozwoli na uzyskanie najlepszego efektu brzmieniowego. Przy takiej interpretacji, założenie o tonie podstawowym jest w pełni uzasadnione, jeżeli jest prawdziwe dla rzeczonego instrumentu.

4. Wnioski

Stworzony program nie ustępuje współczesnym rozwiązaniom komercyjnym, a niekiedy nawet je przewyższa. W dziedzinie transkrypcji muzyki, podobnie jak w dziedzinie rozpoznawania mowy, gestów, kształtów jest wciąż wiele problemów otwartych, nie czysto akademickich, lecz takich, których rozwiązanie można przekuć w gotowy, użyteczny produkt.

Opracowana aplikacja w wersji Lite wraz z dokumentacją jest dostępna bezpłatnie pod adresem www.imare.pl.

Literatura

Brown Judith C. 1991. Calculation of a constant Q spectral transform, *Journal of the Acoustical Society of America* 89(1):425-434

Brown Judith C., Puckette Miller S. 1992. An efficient algorithm for the calculation of a constant Q transform, *Journal of the Acoustical Society of America* 92(5):2698-2701

Portnoff M.R. 1976. Implementation of the digital phase vocoder using the Fast Fourier Transform, *IEEE Trans. Acoustics, Speech, and Signal Processing*, ASSP-24, nr 3

Adres do korespondencji

Piotr Różański (piotr.rozanski@onet.pl)
Wydział Matematyki i Informatyki
Uniwersytet Mikołaja Kopernika w Toruniu