Krzysztof Marasek, dr hab.
Polsko-Japońska Wyższa Szkoła
Technik Komputerowych
ul. Koszykowa 86
02-008 Warszawa

Warszawa, 23 grudzień 2010 r.

Recenzja pracy doktorskiej mgr Elżbiety Kubery
pt. „*Rola atrybutów temporalnych w rozpoznawaniu instrumentów muzycznych w nagraniu wielobrzmieniowym*"

1. Uwagi wstępne

Rozprawa doktorska p. Kubery dotyczy w głównej mierze wykorzystania autorskich metod opisu zmian dźwięku muzycznego w automatycznym wyszukiwaniu instrumentów w dźwiękach polifonicznych. Ten główny wątek pracy dotyczy szybko rozwijającej się dziedziny wyszukiwania informacji muzycznej (MIR – *music information retrieval*), której zastosowania dotyczą multimedialnych baz danych, analizy i percepcji muzyki, interakcji człowiek-komputer, ale także formalnych metod opisu dźwięku muzycznego oraz jego automatycznej klasyfikacji. Z tego powodu uważam ogólny temat rozprawy za dobrze sformułowany i sytuujący rozprawę w głównym nurcie prac dotyczących MIR. Warto też zauważyć, że Autorka postanowiła się zmierzyć z niebanalnym problemem wyszukiwania poszczególnych instrumentów w nagraniu polifonicznym i wielobrzmieniowym, zagadnieniem, które pomimo wielu prób nie znalazło do tej pory satysfakcjonującego rozwiązania.

2. Temat i zakres rozprawy

Głównym celem pracy było „zaprezentowanie nowych cech opisujących zmiany charakterystyk dźwięku w czasie i ocena przydatności tych cech w budowie systemów klasyfikacji instrumentów muzycznych". Doktorantka postawiła w rozprawie trzy następujące tezy:

„1. Analiza zmian charakterystyk dźwięku w czasie (czyli analiza temporalna) umożliwia dokładniejszy opis barwy dźwięku instrumentów muzycznych niż inne analizy nie uwzględniające informacji o sposobie zmienności cech dźwięku.
2. Lepsze wyniki klasyfikacji instrumentów w wielobrzmieniowych fragmentach muzyki uzyskuje się, gdy do zbioru dźwięków wykorzystywanych w procesie uczenia klasyfikatorów dodamy dźwięki wielobrzmieniowe, np. sztucznie utworzone miksy pojedynczych dźwięków.
3. Klasyfikacja hierarchiczna daje lepsze rezultaty w zadaniu rozpoznawania wielu instrumentów w nagraniach polifonicznych i wielobrzmieniowych niż klasyczna klasyfikacja bez wykorzystania podziału systematycznego instrumentów."

Można zauważyć pewien rozziew pomiędzy celem pracy a jej tezami, które są dość ostrożnie sformułowane i skoncentrowane są raczej na metodach analizy (eksploracji) danych i konstrukcji klasyfikatorów niż na aspektach muzycznych i tworzeniu cech opisujących brzmienie instrumentów.

3. Ogólna charakterystyka pracy

Praca o objętości 108 stron składa się z 6 podstawowych rozdziałów, wstępu oraz płyty CD, na której zawarto tylko wersję elektroniczną rozprawy (pdf) bez żadnych danych eksperymentalnych. Bibliografia pracy liczy 97 pozycji i jest starannie zredagowana. W pracy nie zamieszczono niestety spisu tabel i ilustracji ani indeksu pojęć lub oznaczeń.
Praca ma postać klasyczną: rozdziały początkowe (1-4) opisują światowy stan badań, a następnie przedstawiono wkład własny Autorki.

4. Szczegółowa analiza zawartości pracy

Rozdział pierwszy rozprawy przedstawia podstawowe pojęcia muzyczne, akustyki i percepcji dźwięku oraz wyszukiwania informacji muzycznej. Rozdział ten jest nieco przeładowany informacjami, choć z drugiej strony zdaję sobie sprawę z trudności związanych z wyborem i mnogością elementów niezbędnych do zrozumienia wywodu pracy. Pierwszy podrozdział opisuje cechy dźwięku muzycznego. Autorka wprowadza pojęcia cech obiektywnych i subiektywnych opisując ich charakterystyki. Opis ich jest w zasadzie poprawny, ale skoncentrowany tylko na parametrach wykorzystywanych w dalszym toku pracy, co powoduje pewne skróty i pominięcia, np. brak opisu skali muzycznej wysokości dźwięku (półtony) czy skali równomiernie temperowanej. Wzmianka o prawie Webera-Fechnera dotyczy tylko głośności, dziwi brak analizy pojęcia iloczasu, a parametry głośności są nieco

pomieszane (son jest zobiektywizowaną miarą podwojenia głośności dźwięków złożonych, a fon jest psychofizyczną miarą głośności pojedynczych tonów względem tonu wzorcowego 1 kHz, patrz [1], tak więc bardziej reprezentatywny byłby rysunek obrazujący krzywe izofoniczne niż Rys.1.1). Także w opisie skali melowej wysokości dźwięku nie nadmieniono o uwzględnionym w niej (podobnie jak w skali Bark) maskowaniu dźwięków, co ma niewątpliwe reperkusje w rozdzielczości częstotliwościowej wykorzystywanej dalej parametryzacji dźwięku (MFCC). Podrozdział 1.2 dobrze opisuje główne obszary działania systemów MIR, a z kolei 1.3 to opis podstawowych elementów systemów eksploracji danych. Podany opis jakości klasyfikacji wykorzystujący dokładność (precision) i kompletność (recall) oraz miarę F warto byłoby jednak rozszerzyć o opis parametrycznych miar, takich jak F_β oraz ROC (Receiver Operation Characteristic, pozwalający ustalić punkt pracy klasyfikatora), uwzględniających preferencje użytkownika względem dokładności lub precyzji. Wreszcie podrozdział 1.4 poświęcony jest przetwarzaniu dźwięków, a szczególnie zagadnieniu wykrywania zdarzeń muzycznych, np. nowych nut w celu segmentacji utworu na jednorodne fragmenty. Opis ten uważam za wyczerpujący i przedstawiający aktualny stan badań.

Bardzo zwięzły rozdział drugi dysertacji (7 stron) opisuje aktualny stan wiedzy na temat rozpoznawania dźwięku instrumentów muzycznych, a w szczególności ich rozpoznawania w wielodźwięku. Zasygnalizowano w tym rozdziale chyba wszystkie aktualnie wykorzystywane metody analizy, aczkolwiek zabrakło mi w nim odniesienia do badań percepcyjnych: jak ludzie odróżniają dźwięki instrumentów, jak dokładnie i w jakich warunkach są w stanie je odróżnić w wielodźwiękach. Drobna uwaga redakcyjna: Eggink i Brown nie zaproponowali, a tylko wykorzystywali teorię *missing feature* (autorem jest Martin Cooke z Univ.Sheffield).

Obszerny rozdział trzeci poświęcono parametryzacji sygnału i tworzeniu jego deskryptorów na potrzeby klasyfikacji. Jest w tym rozdziale trochę nieścisłości i nieco zbytecznych moim zdaniem informacji na temat widma mocy sygnału czy podstaw analizy widmowej, które można znaleźć w wielu podstawowych podręcznikach. I tak po kolei:

- chciałbym sprostować podane na str. 39 i 40 stwierdzenia o obniżeniu jakości klasyfikacji przy zbyt wielu cechach lub cechach ze sobą skorelowanych. To zależy od stosowanego klasyfikatora i jego założeń: od czasu stosowania tzn. graphical models [2] i CRF (Conditional Random Fields) [3] twierdzenia takie nie są uprawnione, zresztą także w zagadnieniach regresji liniowej dodawanie nowych zmiennych zależnych właściwie zawsze polepsza wyniki,

- rozważania dotyczące okna czy rozdzielczości analizy i dalszych zagadnień dotyczą tylko równomiernego próbkowania sygnału,
- składowa stała nie wpływa na widmo sygnału – problemem jest trend, czyli zmiana tej składowej w czasie (str. 46), stąd potrzeba jej wyznaczania i usuwania co okienko sygnału,
- Energia nie jest miarą głośności dźwięku (patrz rozdział I),
- preemfaza wzmacnia wyższe składowe sygnału (patrz str. 48). Proszę Doktorantkę o uzupełnienie brakującego uzasadnienia dlaczego jest ona potrzebna w przypadku sygnałów muzycznych,
- nie znalazłem wzmianki, że większość z używanych parametrów opisu dźwięku wywodzi się ze standardu MPEG-7 (wzmianka dopiero na str. 50 przy opisie konkretnego parametru),opis cech MPEG-7 uważam za wystarczający,
- FFT daje takie same wyniki jak DFT, ale tylko dla tych prążków widma dla których jest wyznaczane FFT. W szczególności poprzez DFT można wyznaczyć te prążki, których nie można wyznaczyć poprzez FFT,
- cepstrum wyznacza się zazwyczaj z logarytmu dziesiętnego modułu widma, a nie jak podano we wzorze (3.13) logarytmu naturalnego,
- DCT używa się tylko dla cepstrum rzeczywistego, warto zauważyć, że użycie DCT powoduje zanik homomorficzności cepstralnego przetwarzania sygnału,
- opis wyznaczania współczynników MFCC jest niepełny – brakuje informacji o zastosowaniu (zwykle) 20 trójkątnych filtrów melowych i przekształceniu ich wyników do zwykle 12 współczynników cepstrum. Zwykle tez stosowany jest też *cepstral liftering*, czyli wzmacnianie współczynników o wyższych indeksach ($c'_n = \left(1 + \frac{L}{2} sin \frac{\pi n}{L}\right) c_n$ [4]). W ogóle w pracy przewija się liczba 13 współczynników MFCC bez wyjaśnienia dlaczego zastosowano taką a nie inną liczbę parametrów, nie podano także zakresu częstotliwości filtrów melowych (pasma dla jakiego są obliczane współczynniki)
- na str. 55 podano, że strukturę harmoniczną dźwięku można wyznaczyć poprzez model autoregresyjny – nie rozumiem tego stwierdzenia, AR pozwala wyznaczyć obwiednię widma, odpowiedź impulsową systemu, a więc coś przeciwnego.

W kontekście tematu pracy zastawiło mnie także dość mechaniczne przeniesienie współczynników MFCC do analizy dźwięku muzycznego. Oczywistą bowiem ich cechą jest zmniejszenie rozdzielczości widmowej w wyższych zakresach częstotliwości, co ma sens w przypadku analizy mowy (wynika z mechanizmów tworzenia dźwięków mowy oraz zredukowanemu wpływowi wyższych formantów na zrozumiałość mowy), ale niekoniecznie

ma to związek z muzyką (MFCC nie do końca odpowiadają charakterystyce słuchu). W rozdziale tym natomiast szczegółowo i wyczerpująco opisano parametry związane z ewolucją sygnału i jego widma w czasie, a także zastosowania tych metod w śledzeniu zmian dźwięków i rozpoznawaniu gatunku muzycznego. Ostatni podrozdział poświecono metodom selekcji cech. Jest on bardzo zwięzły, szkoda, bo temat jest ciekawy. Zabrakło w nim stwierdzenia, że metody selekcji atrybutów zależą zwykle od rodzaju automatycznego klasyfikatora (o czym już wspomniałem wcześniej – CRF) oraz przedstawienia (choćby skrótowego) metod pozwalających na ortogonalizację przestrzeni cech (np. metoda składowych głównych PCA czy też analiza LDA).

Kolejny rozdział opisuje wkład własny Autorki – nowe rodzaje parametrów temporalnych. Autorka wprowadza cechy długo- i krótkoczasowe oraz metody ich analizy: deskryptory trendu, aproksymowanie ich przebiegu funkcjami rożnych typów, analizę wierzchołków przebiegów cech. Jako miarę jakości dopasowania funkcji wybrano kwadrat współczynnika korelacji Pearsona opisujący, jak słusznie zauważono, liniową zależność. Czy nie lepsza byłaby tu inna miara, np. odległość Mahalanobisa czy Bhattacharyya? Opis poszczególnych cech temporalnych nie jest zbyt jasny: często występujący w dalszej części pracy parametr dlug3 opisuje rozrzut(rozrzuty?) tych cech w krótkich ramkach sygnału znajdujących się w obrębie i-tej długiej, ale nie jest to łatwe do zrozumienia. Przy okazji, czy przeanalizowano choćby wstępnie statystyczne rozkłady stosowanych cech? Zastanawia mnie bowiem częste użycie wartości średnich: przy rozkładach wielomodalnych zapewne lepsza byłaby mediana.

Rozdział piąty pracy opisuje eksperymenty dotyczące rozpoznawania instrumentów muzycznych, zawężonych do grupy 10 instrumentów (klarnet w stroju B, obój, flet poprzeczny, puzon tenorowy, waltornia (róg), fortepian, skrzypce, altówka, wiolonczela, kontrabas). Niestety, Autorka nie podaje żadnych przyczyn dla takiego wyboru instrumentów. Także arbitralnie wybrano grupę klasyfikatorów używanych do wstępnego eksperymentu wyboru klasyfikatora. Uważam, że konieczne jest uzupełnienie rozprawy poprzez podanie uzasadnienia dokonanych wyborów instrumentów i grup klasyfikatorów. Proszę też o przedstawienie w trakcie obrony pracy poprawionej ilustracji 5.1 (mam nadzieję, że jest to oczywisty błąd skali pionowej na rysunku, jeśli nie, to podane wyniki de facto dyskwalifikują dalsze rozważania).

Opisane w rozdziale czwartym cechy temporalne użyto następnie w klasyfikacji jednoetykietowej oraz wieloetykietowej dla fragmentów jedno- i wielobrzmieniowych. Autorka przeprowadziła wiele eksperymentów klasyfikacyjnych wskazujących na sensowność wykorzystania zaproponowanych cech w zagadnieniu klasyfikacji instrumentów

oraz istotne polepszenie wyników, gdy w fazie uczenia wykorzysta się sztucznie zmiksowane dźwięki pojedynczych instrumentów. Nie wchodząc już drobiazgowo w poszczególne wyniki, zauważam jednak brak w tym rozdziale brak odniesienia do innych metod klasyfikacji czy parametryzacji sygnału muzycznego. Przekonywującym bowiem eksperymentem byłoby zbudowanie np. klasyfikatora wykorzystującego klasyczne cechy MFCC i ich pochodne (lub inne najczęściej stosowane w literaturze), a następnie wykazanie, że zaproponowane podejście jest istotnie lepsze. Zabrakło mi także odpowiedzi na pytanie dlaczego dodanie miksów poprawia wyniki klasyfikacji? Czy chodzi tylko o większą ilość danych uczących?
Rozdział szósty dysertacji poświęcono wykorzystaniu hierarchicznej klasyfikacji instrumentów w zagadnieniu rozpoznawania instrumentów oraz budowie systemu klasyfikacji hierarchicznej dla automatycznej taksonomii instrumentów (wynikającej z klasteryzacji) i dla ujęcia Hornbostela i Sachsa. W ostatecznych wynikach uwzględniono także kontekst muzyczny klasyfikacji. Pozwoliło to na osiągnięcie 75% poprawności rozpoznawania w utworach klasycznych dziesięciu wybranych instrumentów, co jest niewątpliwe wynikiem bardzo dobrym zważywszy złożoność zagadnienia.
Wreszcie rozdział siódmy stanowi podsumowanie i wnioski. Uważam, że jest zaskakująco krótki i niepełny. W szczególności brakuje w nim odniesienia do tez pracy.

5. Ocena edytorska pracy

Praca jest niewątpliwie starannie zredagowana i stosunkowo poprawna pod względem językowym, aczkolwiek da się zauważyć kilka słabości redakcyjnych. Przede wszystkim o czym już wspomniano brakuje indeksu pojęć: w opisie eksperymentów częste są odwołania do wprowadzonych przez Autorkę oznaczeń parametrów, odszukanie ich znaczenia nie jest łatwe. Zabrakło mi też w rozprawie choćby krótkiego przedstawienia stosowanych baz nagrań (warunków akustycznych nagrań) oraz sposobu miksowania dźwięków.
Drobne literówki i nieliczne błędy stylistyczne oraz interpunkcyjne nie wpływają w żadnej mierze na jakość ocenianej rozprawy doktorskiej. Nie mam też zastrzeżeń co do układu pracy i kolejności rozdziałów.

6. Podsumowanie

Rozprawa doktorska p. Elżbiety Kubery jest niewątpliwie interesującą próbą zmierzenia się z problemem wykrywania instrumentów w nagraniach muzycznych, co jest zagadnieniem