Waldemar CUDNY[1], **Krzysztof TYBUREK**[2], **Witold KOSIŃSKI**[1]
[1] POLSKO-JAPOŃSKA WYŻSZA SZKOŁA TECHNIK KOMPUTEROWYCH
[2] UNIWERSYTET KAZIMIERZA WIELKIEGO, INSTYTUT MECHANIKI ŚRODOWISKA I INFORMATYKI STOSOWANEJ

Automatyczna klasyfikacja instrumentów szarpanych w multimedialnych bazach danych

Dr Waldemar CUDNY

Jest pracownikiem naukowo-dydaktycznym w Instytucie Mechaniki Środowiska i Informatyki Stosowanej Uniwersytetu Kazimierza Wielkiego w Bydgoszczy oraz współpracuje z Polsko-Japońska Wyższą Szkołą Technik Komputerowych w Warszawie. Specjalizuje się w analizie jedno i dwuwymiarowych sygnałach cyfrowych.

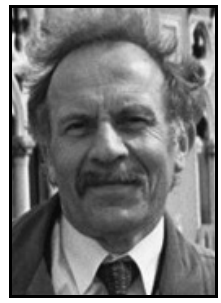
e-mail: wcudny@ukw.edu.pl

Mgr Krzysztof TYBUREK

Jest pracownikiem naukowo-dydaktycznym w Instytucie Mechaniki Środowiska i Informatyki Stosowanej Uniwersytetu Kazimierza Wielkiego w Bydgoszczy. Specjalizuje się w dziedzinach systemów baz danych i ich zastosowaniach, programowaniu i językach programowania, DSP i multimediach.

e-mail: krzysiekkt@ukw.edu.pl

Prof. dr hab. Witold KOSIŃSKI



Jest pracownikiem naukowo-dydaktycznym w Polsko-Japońskiej Wyższej Szkole Technik Komputerowych w Warszawie oraz Instytucie Mechaniki Środowiska i Informatyki Stosowanej Uniwersytetu Kazimierza Wielkiego w Bydgoszczy. Sfery zainteresowania: matematyczne podstawy inteligencji obliczeniowej, sieci neuronowe, logika i liczby rozmyte, klasyfikacja i rozpoznawanie wzorców, falowe równania różniczkowe, mechanika i termodynamika continuum.

e-mail: wkos@pjwstk.edu.pl

Streszczenie

Klasyfikacją i agregacją danych multimedialnych zajmuje się standard MPEG-7, który dostarcza szereg podstawowych deskryptorów opisujących dźwięk. Wzorując się na istniejącym standardzie MPEG-7 stworzono nowe deskryptory rozpoznające konkretne instrumenty muzyczne. Głównym zadaniem postawionym w badaniach jest takie zdefiniowanie deskryptorów w przestrzeni widmowej, które w połączeniu z określonymi algorytmami przeszukiwań pozwolą na prawidłową interpretację źródła dźwięku z artykulacją piccicato. Do badań wybrano grupę strunowych instrumentów muzycznych znaną pod nazwą chordofonów.

Słowa kluczowe: sygnał audio, widmo sygnału, deskryptory, pizzicato, klasyfikacja, chordofony, MPEG-7 standard.

Automatic classification of string instruments in multimedia databases

Abstract

Classification and agregation of multimedia data used, for example in production processes of TV and radio stations is made by the use of MPEG-7 staandard. Searching process can be speed up if an appropriate labeling (indexing) of signals is used. The paper concerns determination of a set of descriptors in the spectrum domain which can allow to classify pizzicato sound signals generated by 8 different chordophone instruments being a subset of strings.

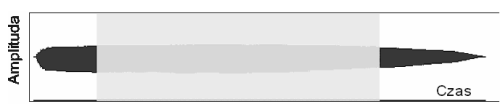
Keywords: audio signals, frequency domain, descriptors, pizzicato, classification, chordophones, MPEG-7 standard.

1. Analiza dźwięku wybranych instrumentów muzycznych

Większość dotychczasowych rozwiązań związanych z wydobywaniem danych multimedialnych bazuje na technice etykietowania przechowywanych informacji. Rozwiązanie to nie zawsze daje rzetelny wynik – tzn. wysyłane zapytania nie zawsze jest zgodne z oczekiwaniami. Kolejnym problemem, który występuje w procesie rozpoznawania sygnałów dźwiękowych, jest właściwa interpretacja źródła dźwięku. Rozpoznanie dźwięku pochodzącego np. z drgającej struny gitary może być bardzo trudne. Trudność ta najczęściej wynika z doskonałych procesorów muzycznych, za pomocą których z łatwością można „podrobić" oryginalny instrument. Drogą do rozwiązania problemu klasyfikacji i agregacji danych multimedialnych jest standard MPEG-7, który dostarcza szeregu podstawowych deskryptorów opisujących dźwięk.

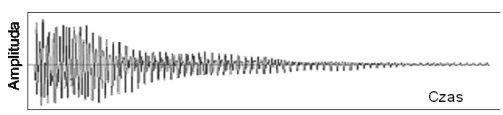
Dźwięki muzyczne wykazują okresowość przed osiągnięciem 10% swej maksymalnej amplitudy, a co istotniejsze, etap narastania dźwięku jest jedną z najistotniejszych cech dystynktywnych, pozwalających słuchaczowi sklasyfikować instrument. Posługując się takimi podstawowymi metodami analizy dźwięków jak transformata Fouriera oraz analiza falkowa można precyzyjnie określić skład widmowy analizowanej próbki. Przebiegi czasowe dźwięku badane są na podstawie analizy:

1. transjentu początkowego,
2. stanu quasi-ustalonego,
3. transjentu końcowego.



Rys. 1. Przykład wykresu postaci czasowej dźwięku waltorni, a razkreślne (440Hz). Zacieniowano stan quasi-ustalony
Fig. 1. Example of waveform of a[1] (440hz) of French horn. Shaded region is the steady stage

W przypadku analizy sygnałów pochodzących z grupy instrumentów szarpanych należy brać pod uwagę tylko transjent końcowy – stan quasi-ustalony w tym przypadku nie występuje, (co jest cechą charakterystyczną tych instrumentów).



Rys. 2. Przykład wykresu postaci czasowej dźwięku altówki, a razkreślne (440Hz)
Fig. 2. Example of waveform of a[1] (440hz) of viola

2. Przyjęte metody badawcze

W celu odszukania wektora cech wybranej grupy instrumentów przeprowadzano analizę zarówno postaci czasowej jak i widmowej dźwięków. Do badań przeznaczono 840 próbek dźwięków zawierających się w zakresach 4 oktaw:

Wielka (A=110Hz)
Mała (a=220Hz)
Razkreślna (a^1 = 440Hz)
Dwukreślna (a^2 = 880Hz)

Zakres częstotliwości badanych próbek: 65,41 Hz<f< 987,77Hz. Badano pojedyncze dźwięki do naturalnego wybrzmiewania nuty.

2.1. Parametryzacja w dziedzinie czasu

W celu właściwego opisu postaci czasowej sygnału dźwiękowego zdecydowano się wykorzystać dwa parametry:

1. ZC – (zerro crossing) gęstość przejść przez zero osi OX w zadanym oknie. Do analizy wybrano okno o długości 1500 próbek rozpoczynając od wartości max, a więc wykluczono transjent początkowy przebiegu.
2. l_{tk} – logarytm czasu wybrzmiewania dźwięku wyrażony zależnością:

$$l_{tk} = log(t_{tk} - t_{max}) \tag{1}$$

gdzie:

t_{max} – czas osiągnięcia maksymalnej amplitudy dźwięku,
t_{tk} – czas osiągnięcia progu 10% maksymalnej amplitudy dźwięku w transjencie końcowym.

2.2. Parametryzacja w dziedzinie widma

Widmo zawiera bardzo wiele szczegółów, a zatem do celów automatycznej klasyfikacji instrumentów muzycznych konieczna jest jego parametryzacja. W celu odszukania wektora cech widmowych wybranych instrumentów przeprowadzono szereg badań związanych z wyznaczaniem środka ciężkości widma, zawartości składowych parzystych lub nieparzystych, odszukanie prążków harmonicznych itp. Badania przeprowadzone zostały na wyciętym oknie sygnału o długości 11025 próbek mierzonym od wartości max. Wycięty fragment przebiegu postaci czasowej został poddany DFT, a jego widmo poddano szczegółowej analizie. Zastosowanie jednakowej długości okna podyktowane było koniecznością utrzymania jednakowej rozdzielczości widma wyrażonej zależnością:

$$f_r = \frac{f_s}{n} \tag{2}$$

gdzie:

f_r – rozdzielczość widma
f_s – częstotliwość próbkowania (44100)
n – ilość próbek (11025)

Podczas prowadzonych badań zdecydowano się na rozdzielczość widma fr=4Hz.

Stwierdzono, że dla celów automatycznej klasyfikacji badanych instrumentów niektóre metody badawcze nie przynoszą istotnych korzyści. Na przykład analizując wyniki uzyskane na podstawie metody momentów k-tego rzędu wyrażonej zależnością:

$$m_k = \sum_{i=1}^{n} A(i) \cdot f(i)^k \tag{3}$$

gdzie:

m_k – moment widmowy k-*tego* rzędu
$A(i)$ – amplituda i-tej składowej.
$f(i)$ – częstotliwość i-tego prążka widma

stwierdzono, że nie uzyskano wyników istotnie przyczyniających się do określenia cechy jednego instrumentu. Wyciągnięty wniosek poparto faktem, że zbyt duża część zakresów jest właściwa dla różnych instrumentów zagranych w różnych oktawach.

Stwierdzono również, że jedną z istotniejszych grup deskryptorów charakteryzujących cechy widma są parametry tristimulus (Tr_1, Tr_2, Tr_3) opisywane zależnościami:

$$Tr_1 = \frac{A(1)^2}{\sum_{i=1}^{n} A(i)^2} \tag{4}$$

$$Tr_2 = \frac{\sum_{i=2}^{4} A(i)^2}{\sum_{i=1}^{n} A(i)^2} \tag{5}$$

$$Tr_3 = \frac{\sum_{i=5}^{n} A(i)^2}{\sum_{i=1}^{n} A(i)^2} \tag{6}$$

gdzie:

$A(i)$ – amplituda i-tej składowej
n – ilość próbek sygnału

Wykorzystując grupę parametrów tristimulus można rozróżnić dźwięki analizując zawartość grup harmonicznych widma w poszczególnych zakresach częstotliwości. Stwierdzono również, że klasyczne parametry tristimulus mogą okazać się mało efektywne w przypadku zastosowania progowania widma oraz stwierdzono, że szczególnie Tr_1 wnosi mało istotne informacje. Wniosek ten można wyciągnąć uwzględniając fakt, że w widmie mogą się zawierać niskie częstotliwości związane z zakłóceniami powstałymi podczas rejestrowania dźwięku (np. przypadkowe uderzenie w mikrofon lub pudło rezonansowe), które są brane pod uwagę podczas obliczania Tr_1 oraz Tr_2. W związku z opisanymi możliwościami uzyskania mało precyzyjnych wyników, zdecydowano się dokonać modyfikacji grupy parametrów tristimulus uwzględniając prążek o wartości maksymalnej jako najistotniejszą informację widma. Zdecydowano się przedstawić opisywaną grupę parametrów zależnościami:

$$NTr_1 = \frac{A(\max)^2}{\sum_{i=1}^{n} A(i)^2} \tag{7}$$

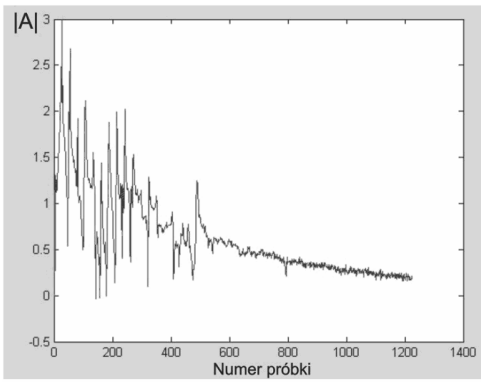
$$NTr_2 = \frac{\sum_{i=\max}^{2\cdot\max} A(i)^2}{\sum_{i=1}^{n} A(i)^2} \tag{8}$$

$$NTr_3 = \frac{\sum_{i=2\cdot\max}^{n} A(i)^2}{\sum_{i=1}^{n} A(i)^2} \tag{9}$$

gdzie:

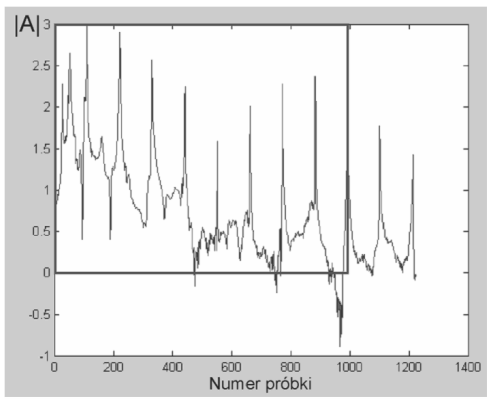
max – indeks prążka o maksymalnej wartości
$A(max)$ – amplituda prążka o maksymalnej wartości.

Analizując rozkład częstotliwościowy badanych widm zdecydowano się wprowadzić podział widma na 10 kolumn - po 100 próbek każda. Wszystkie wycięte kolumny widma poddano analizie. Stwierdzono również, że analiza widma w wyższych partiach częstotliwości nie przynosi ciekawych informacji, co przedstawiono graficznie:

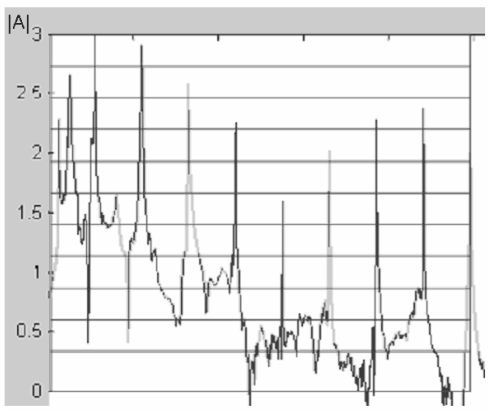


Rys. 3. Postać widmowa dźwięku a razkreślne dla wiolonczeli
Fig. 3. Example of spectrum of a^1 (440hz) of cello

Z powyższego przykładu można wyczytać, że analiza widma powyżej 1000 próbki sprowadza się do analizy szumu, co nie jest interesujące dla prowadzonych badań. W związku z tym zdecydowano się przeznaczyć do analizy okno widma zawarte między 1 a 1000 próbką – a zatem badano rozkład częstotliwościowy w zakresie do 4kHz.



Rys. 4. Fragment widma gitary akustycznej z zaznaczonym obszarem przeznaczonym do analizy
Fig. 4. Acoustic guitar spectrum with the window considered



Rys. 5. Przykładowy podział widma na warstwy
Fig. 5. Example of partition of spectrum into layers

W trakcie badań zdecydowano się przeprowadzić analizę rozkładu częstotliwościowego badanego fragmentu widma. W tym celu dokonano podziału widma na 10 kolumn, w których zliczano zgromadzoną energie. Poza tym analizę skierowano na rozkład energetyczny w poszczególnych warstwach widma, badając ilość zgromadzonej energii w poszczególnych przedziałach.

3. Decyzja/Rozpoznanie

Proces klasyfikacji instrumentów oraz selekcji cech został przeprowadzony z wykorzystaniem ogólnie dostępnego pakietu WEKA. Wykorzystując wyniki uzyskane podczas analizy rozkładu częstotliwościowego oraz energetycznego otrzymano dla 8 klas instrumentów rozpoznawalność wahającą się w granicach od 61.9266 % do 77.7778 %. Analizując zebrane próbki dźwięku najczęściej posługiwano się metodą holdout, która polega na jednokrotnym podziale zbioru na część treningową i testową. Podczas prowadzonych badań zdecydowano się uwzględnić podział, który dla części testowej oscylował od 20 % do 40% populacji zbioru wejściowego. Fragment przykładowej macierzy przekłamań przedstawiono poniżej:

c	d	e	f	G	h		<--	Classified
75	12,5	0	0	0	0	c	=	gitara_elektryczna
0	90	0	0	0	0	d	=	gitara_basowa
0	0	75	25	0	0	e	=	Altowka
0	12,5	12,5	62,5	12,5	0	f	=	Skrzypce
0	0	0	14,29	85,71	0	g	=	Kontrabas
0	0	22,22	11,11	0	66,67	h	=	Wiolonczela

Rys. 6. Fragment macierzy przekłamań badanych instrumentów. Ogólna rozpoznawalność = 71.11 % przy podziale zbioru 60:40
Fig. 6. Part of error matrix obtained in the experiments. The level of recognition is 71.11%; proportion of training and testing sets is 60:40

Z powyższego fragmentu macierzy wynika, że przeznaczając 60% zbioru próbek na część treningowa zbioru a 40% na część testową najgorszą rozpoznawalnością charakteryzują się skrzypce, które poprawnie zostały zinterpretowane tylko w 62,5%. W 12,5% procentach próbki skrzypiec zostały zinterpretowane jako dźwięki kontrabasu, gitary basowej i altówki.

W dalszej pracy autorzy planują skupić swoja uwagę na optymalnym doborze szerokości warstw oraz ich ilości. Jako drogę do rozwiązania w/w problemu zaplanowano wykorzystanie histogramu amplitudy.

4. Literatura

[1] Jordi Bonada, Alex Loscos, Pedro Cano, Xavier Serra „Spectral Approach to the Modeling of the Singing Voice" p. 4-15 Proceedings of 111th AES Convention; 2001 September 21–24 New York, USA
[2] José M. Martínez "MPEG-7 Overview", Klangenfurt, July 2002. ISO/IECJTC1/SC29/WG11, N4980 pp. 1-96
[3] Xavier Serra, Xavier Amatriain, Jordi Bonada, Alex Loscos "Spectral Modeling for Higher-level Sound Transformations" Mosart Deliverable D22. Evaluation report of Timbre modeling, 2002.", url = "citeseer.ist.psu.edu/amatriain01spectral.html"
[4] Krzysztof Tyburek „Klasyfikacja cech instrumentów muzycznych w standardzie MPEG 7" s. 4-12 Lwów, czerwiec 2004

Artykuł recenzowany