Zeszyty Naukowe Wydziału Elektrotechniki i Automatyki Politechniki Gdańskiej Nr 20

*XIV Seminarium*
*ZASTOSOWANIE KOMPUTERÓW W NAUCE I TECHNICE' 2004*
Oddział Gdański PTETiS

# SYSTEM ROZPOZNAWANIA DŹWIĘKÓW INSTRUMENTÓW MUZYCZNYCH

**Piotr DALKA[1], Marcin DĄBROWSKI[2]**

Katedra Systemów Multimedialnych, Wydział Elektroniki, Telekomunikacji i Informatyki,
Politechnika Gdańska, ul. G. Narutowicza 11/12, 80-952 Gdańsk
tel: (058) 347 1301
e-mail: 1. dalken@sound.eti.pg.gda.pl, 2. bigyo@sound.eti.pg.gda.pl

Niniejszy referat przedstawia działanie systemu automatycznego rozpoznawania pojedynczych dźwięków instrumentów muzycznych. System składa się z trzech bloków: detekcja częstotliwości podstawowej, parametryzacja dźwięków i klasyfikacja. W algorytmie detekcji wykorzystano zmodyfikowany algorytm Schroedera. Parametryzację przeprowadzono głównie w oparciu o parametry zdefiniowane w standardzie MPEG-7. Na potrzeby systemu zaimplementowano trzy algorytmy klasyfikacji bazujące na sieciach neuronowych oraz wykorzystujące metodę najbliższego sąsiada. W referacie porównano wyniki klasyfikacji uzyskane w oparciu o opisane algorytmy oraz oceniono przydatność parametrów MPEG-7 do rozpoznawania dźwięków instrumentów muzycznych.

## 1. WPROWADZENIE

Opracowanie systemów automatycznego rozpoznawania dźwięków instrumentów muzycznych staje się z biegiem czasu coraz większą koniecznością. Systemy takie, w połączeniu z algorytmami rozplotu dźwięków instrumentów muzycznych i rozpoznawania linii melodycznej umożliwią przeprowadzanie szybkiego indeksowania nagrań muzycznych w komputerowych bazach danych, a następnie efektywne ich przeszukiwanie.

Dźwięki instrumentów muzycznych cechują się ogromną różnorodnością. W praktyce liczba instrumentów muzycznych jest nieograniczona. Wymaga to na wstępie ograniczenia wielkości zbioru rozpoznawanych instrumentów do skończonej, stosunkowo niewielkiej liczby elementów. Dźwięki wydobywane przez dwa różne typy instrumentów mogą być bardzo podobne do siebie. Ponadto cechy dźwięku instrumentu muzycznego zależą od sposobu artykulacji oraz wysokości dźwięku. Dodatkowo różnice konstrukcyjne między takimi samymi instrumentami sprawiają, że nigdy nie brzmią one identycznie. Z tego powodu poprawna identyfikacja instrumentu na podstawie pojedynczego dźwięku jest skomplikowanym zadaniem.

Eksperymenty przedstawione w niniejszym referacie zostały przeprowadzone na bazie próbek 10 instrumentów muzycznych: fagotu, klarnetu B, oboju, puzonu tenorowego, rogu

(waltorni), saksofonu altowego, skrzypiec, trąbki, tuby F oraz wiolonczeli. Większość próbek dźwiękowych (80%) pochodziła z Katalogu Dźwięków Instrumentów Muzycznych, który powstał w Katedrze Systemów Multimedialnych Politechniki Gdańskiej. Zestaw ten uzupełniono dźwiękami ze zbioru *McGill University Master Samples* (*MUMS*). W eksperymentach wykorzystano ponad 3500 próbek dźwiękowych o różnej artykulacji i dynamice. Wszystkie dźwięki charakteryzowały się częstotliwością próbkowania 44,1 kHz.

## 2. ALGORYTM DETEKCJI CZĘSTOTLIWOŚCI PODSTAWOWEJ

Kluczowym zagadnieniem w rozpoznawaniu dźwięków instrumentów muzycznych jest ich parametryzacja. W celu wyznaczenia wielu parametrów niezbędna jest znajomość wartości częstotliwości podstawowej dźwięku.

Do implementacji wybrano algorytm analizujący rozkład prążków widma bazujący na histogramie Schroedera. W trakcie prac nad algorytmem wprowadzono wiele modyfikacji, których celem było uzyskanie jak największej skuteczności działania, a w szczególności minimalizacja błędów oktawowych. Algorytm składa się z następujących bloków [1]:

1. Przygotowanie widma (obliczanie FFT, logarytmowanie, usuwanie trendu)
2. Wyznaczenie położenia prążków (kwantyzacja 1-bitowa, różniczkowanie)
3. Wyznaczenie częstotliwości podstawowej na podstawie analizy różnic między położeniami prążków.

Opracowany algorytm błędnie rozpoznał częstotliwość podstawową w przypadku 157 dźwięków spośród 3552, co daje skuteczność detekcji 95,6%. Wysokość wszystkich dźwięków oboju została wyznaczona prawidłowo. Z kolei najsłabszy wynik (87,3%) osiągnięto dla skrzypiec. Uzyskaną łączną skuteczność należy uznać za bardzo dobrą, biorąc pod uwagę zastosowanie częstotliwości podstawowej w algorytmach klasyfikacji dźwięków instrumentów muzycznych. Jednak w celu uniezależnienia wyników działania tych algorytmów od algorytmu wyznaczania częstotliwości podstawowej, wszystkie pozostałe eksperymenty zostały przeprowadzone na bazie dźwięków o prawidłowo określonej wysokości.

## 3. PARAMETRYZACJA DŹWIĘKÓW

Parametryzacja dźwięków instrumentów muzycznych prowadzi do określenia wektora cech dźwięku muzycznego. Wielkość tego wektora (liczba parametrów opisujących dźwięk) powinna być zminimalizowana, tzn. parametry silnie skorelowane ze sobą powinny być reprezentowane przez jeden parametr. Minimalizacja ta zwiększa uporządkowanie niesionej informacji oraz zmniejsza złożoność obliczeniową.

Zdecydowana większość analizowanych parametrów dźwięku została zdefiniowana w standardzie opisu danych multimedialnych MPEG-7 [2]. Są to [3]:

- *ASE* (ang. *Audio Spectrum Envelope*) – krótkookresowe widmo gęstości mocy sygnału, wykorzystujące skalę logarytmiczną na osi częstotliwości. W skład *ASE* w jednej ramce sygnału wchodzi jeden współczynnik oznaczający moc dla częstotliwości mniejszych od 62,5 Hz, seria współczynników reprezentujących moc w pasmach o szerokości ¼ oktawy w zakresie od 62,5 Hz do 16 kHz oraz jeden współczynnik dla pasma powyżej 16 kHz. Daje to w sumie 34 współczynniki dla jednej ramki sygnału. Średnie wartości każdego współczynnika oraz ich wariancje oznaczono odpowiednio jako $ASE_1 \ldots ASE_{34}$ i $ASEv_1 \ldots ASEv_{34}$.

- *ASC* (ang. *Audio Spectrum Centroid*) – środek ciężkości widma gęstości mocy o logarytmicznej skali częstotliwości. Wynikiem jest odległość w oktawach od referencyjnej częstotliwości 1 kHz. Średnią wartość *ASC* w czasie oraz jej wariancję oznaczono *ASC* i *ASC*v.
- *ASS* (ang. *Audio Spectrum Spread*) – wielkość odchylenia (drugi moment statystyczny) wartości skutecznej RMS widma gęstości mocy o logarytmicznej skali częstotliwości od środka ciężkości *ASC*. Średnią wartość *ASS* w czasie oraz jej wariancję oznaczono *ASS* i *ASS*v
- *SFM* (ang. *Spectral Flatness Measure*) – płaskość widma sygnału. Analiza przeprowadzana jest w pasmach częstotliwości o szerokości ¼ oktawy, w zakresie od 250 Hz do 16 kHz. W każdym paśmie płaskość jest zdefiniowana jako stosunek średniej geometrycznej i arytmetycznej próbek widma gęstości mocy. Uśrednione w czasie wartości *SFM* w poszczególnych pasmach oraz ich wariancje oznaczono odpowiednio $SFM_1 \ldots SFM_{24}$ i $SFMv_1 \ldots SFMv_{24}$.
- *LAT* (ang. *Log Attack Time*) – czas trwania transjentu wejściowego wyrażony w skali logarytmicznej.
- *SC* (ang. *Spectral Centroid*) – środek ciężkości widma w Hz obliczany jako ważona amplitudowo średnia częstotliwość próbek w widmie sygnału. Średnią wartość *SC* w czasie i jej wariancję oznaczono *SC* i *SC*v.
- *HSC* (ang. *Harmonic Spectral Centroid*) – środek ciężkości widma w Hz wyznaczony jako ważona amplitudowo średnia częstotliwość prążków w widmie sygnału. Średnią wartość *HSC* w czasie oraz jej wariancje oznaczono *HSC* i *HSC*v.
- *HSD* (ang. *Harmonic Spectral Deviation*) – odchylenie logarytmów amplitud prążków widma od logarytmu lokalnej obwiedni widma. Średnią wartość *HSD* w czasie i jej wariancję oznaczono *HSD* i *HSD*v.
- *HSS* (ang. *Harmonic Spectral Spread*) – ważone amplitudowo standardowe odchylenie amplitud prążków widma, odniesione do *HSC*. Średnią wartość *HSS* w czasie i jej wariancję oznaczono *HSS* i *HSS*v.
- *HSV* (ang. *Harmonic Spectral Variation*) – znormalizowana korelacja między amplitudami prążków widma w dwóch sąsiednich ramkach sygnału. Średnią wartość HSV w czasie i jej wariancję oznaczono *HSV* i *HSV*v.

Dodatkowo oprócz parametrów zdefiniowanych w standardzie MPEG-7, wzięto też pod uwagę dwa inne deskryptory dźwięku:

- *KeyNum* – wysokość dźwięku w skali liniowej. Wyrażona jest jako numer dźwięku zgodny ze standardem MIDI [4].
- *Ev* – zawartość parzystych harmonicznych w widmie sygnału [5].

W wyniku analizy statystycznej, opartej o statystyki Behrensa-Fishera [6] oraz analizę korelacyjną [6], wyznaczono optymalny wektor parametrów stanowiący podstawę działania zaimplementowanych algorytmów klasyfikacji dźwięków instrumentów muzycznych:

[ $ASE_{2\ldots5,\ 8,\ 9,\ 18,\ 21,\ 23\ldots31,\ 33,\ 34}$, $ASEv_{5\ldots9,\ 21,\ 31,\ 34}$, *ASC*, *ASS*, *ASS*v, $SFM_{13\ldots19}$, $SFM_{21,\ 22,\ 24}$, *HSC*, *HSD*, *HSD*v, *HSS*, *HSS*v, *Ev*, *LAT*, *KeyNum* ]

## 4. ALGORYTMY KLASYFIKACJI

W celu automatycznego rozpoznawania dźwięków instrumentów muzycznych zaimplementowano trzy algorytmy. Dwa z nich działają w oparciu o sztuczne sieci

neuronowe, trzeci zaś jest realizacją metody minimalnoodległościowej (metoda najbliższego sąsiada).

### 4.1. Sieci neuronowe

Jako podstawowy algorytm rozpoznawania do zadań klasyfikacji dźwięków instrumentów muzycznych wykorzystano jednokierunkową sieć neuronową o trzech warstwach. Strukturę sieci zdefiniowano następująco:

- liczba neuronów w warstwie wejściowej równa liczebności wektora cech,
- liczba neuronów w warstwie ukrytej równa liczbie neuronów warstwy wejściowej,
- liczba neuronów wyjściowych równa ilości rozpoznawanych klas; każdy instrument jest reprezentowany przez jeden neuron w warstwie wyjściowej,
- neurony w warstwie wejściowej i wyjściowej posiadały unipolarną sigmoidalną funkcję aktywacji, a neurony w warstwie ukrytej – funkcję bipolarną.

Nauka sieci była przeprowadzana zgodnie z algorytmem wstecznej propagacji błędu EBP. Dodatkowo zastosowano algorytm kontroli procesu generalizacji sieci. Wektory wejściowe dzielone były w stosunku 1:1 na wektory uczące i sprawdzające. Przydział wektorów do każdej z tych grup był losowy. Proces nauki uznany był za zakończony, jeśli skumulowana wartość błędu odpowiedzi sieci na wektory uczące spadała poniżej założonego progu, lub gdy skumulowana wartość błędu odpowiedzi sieci na wektory sprawdzające wzrastała przez więcej niż 10 iteracji z rzędu. Trening sieci powtarzano 10 razy i najlepiej wytrenowana sieć była wykorzystywana do dalszych badań.

Tak zdefiniowana pojedyncza sieć neuronowa stanowi jednoetapowy klasyfikator dźwięków instrumentów muzycznych. Ponadto zaimplementowano również dwuetapowy system złożony z czterech sieci neuronowych. Sklasyfikowanie każdego dźwięku wymaga użycia dwóch sieci. Pierwsza z nich rozpoznaje rodzinę instrumentów, do której należy analizowany dźwięk. Następnie na podstawie wyniku tej klasyfikacji dźwięk jest rozpoznawany przez jedną z trzech sieci wyspecjalizowanych w identyfikacji dźwięków instrumentów należących do jednej z rodzin (smyczkowe, dęte drewniane lub dęte blaszane). Struktura każdej sieci neuronowej i zasady ich treningu były analogiczne, jak w przypadku pojedynczej sieci neuronowej.

### 4.2 Metoda najbliższego sąsiada

Dodatkowym algorytmem, który zaimplementowano i zoptymalizowano do zadania rozpoznawania dźwięków instrumentów jest minimalnoodległościowa metoda najbliższego najbliższego sąsiada [7]. Polega ona na znalezieniu wektora ze zbioru wzorców charakteryzującego się największym podobieństwem (najmniejszą odległością) w stosunku do wektora badanego. Przyjmuje się, że badany instrument należy do tej samej klasy, co znaleziony wektor wzorcowy.

Podobnie, jak w przypadku sieci neuronowej, wektory wejściowe dzielone były w stosunku 1:1 na wektory wzorcowe i sprawdzające; przydział wektorów do każdej z tych grup był losowy. Zastosowana została metryka Hamminga, dla której uzyskano we wstępnych badaniach najlepsze wyniki.

## 5. WYNIKI KLASYFIKACJI

W tablicy 1 zestawiono rezultaty rozpoznawania dźwięków 10 instrumentów muzycznych wszystkimi zaimplementowanymi algorytmami klasyfikacji.

Tablica 1. Porównanie skuteczności klasyfikacji zaimplementowanych algorytmów

| Instrument | Pojedyncza sieć neuronowa | | | Grupa sieci neuronowych | | | Metoda najbliższego sąsiada | | |
|---|---|---|---|---|---|---|---|---|---|
| | razem | błędy | % | razem | błędy | % | razem | błędy | % |
| fagot | 179 | 5 | 97,2 | 189 | 8 | 95,8 | 173 | 16 | 90,8 |
| klarnet | 195 | 27 | 86,2 | 189 | 7 | 96,3 | 185 | 10 | 94,6 |
| obój | 173 | 21 | 87,9 | 165 | 14 | 91,5 | 155 | 8 | 94,8 |
| puzon | 166 | 11 | 93,4 | 166 | 12 | 92,8 | 164 | 10 | 93,9 |
| róg | 166 | 23 | 86,1 | 163 | 14 | 91,4 | 179 | 17 | 90,5 |
| saksofon | 124 | 6 | 95,2 | 119 | 9 | 92,4 | 129 | 2 | 98,5 |
| skrzypce | 182 | 13 | 92,9 | 189 | 19 | 90,0 | 204 | 12 | 94,1 |
| trąbka | 138 | 5 | 96,4 | 142 | 11 | 92,3 | 137 | 0 | 100,0 |
| tuba | 159 | 4 | 97,5 | 161 | 3 | 98,1 | 158 | 0 | 100,0 |
| wiolonczela | 214 | 16 | 92,5 | 214 | 10 | 95,3 | 212 | 13 | 93,9 |
| razem | 1696 | 131 | 92,3 | 1697 | 107 | 93,7 | 1696 | 88 | 94,8 |

Spośród algorytmów opartych o sieci neuronowe, lepszy wynik osiągnęła grupa sieci neuronowych, pomimo, iż w tym przypadku każda próbka była klasyfikowana kolejno przez dwie sieci neuronowe, wskutek czego ich błędy się kumulowały. Jednak wynik 96,0% skuteczności rozpoznawania rodziny instrumentów oraz 97,3% skuteczności rozpoznawania instrumentu pod warunkiem, że rodzina została określona poprawnie sprawiły, że grupa sieci neuronowych poprawnie sklasyfikowała prawie 1,5 procenta więcej próbek dźwiękowych w porównaniu z pojedynczą siecią neuronową.

Oba algorytmy oparte na sieciach neuronowych najsłabszy wynik zanotowały dla dźwięków klarnetu i oboju (mylonych ze sobą wzajemnie) oraz rogu (mylonego z puzonem). Warto zauważyć, że rozkład wyników klasyfikacji poszczególnych instrumentów jest bardziej równomierny w przypadku grupy sieci neuronowych.

Najlepszy wynik klasyfikacji osiągnięto jednak w metodzie najbliższego sąsiada. Wszystkie dźwięki trąbki i tuby zostały określone bezbłędnie. Z kolei najsłabszy rezultat uzyskano w przypadku fagotu, który był mylony z puzonem i tubą oraz rogu, błędnie rozpoznawanego jako puzon. Niekorzystną cechą jest istnienie wysokiej różnicy miedzy wynikiem klasyfikacji dla najlepiej i najgorzej rozpoznanego instrumentu.

## 6. WNIOSKI

Przeprowadzone eksperymenty wykazują, że parametry ze standardu MPEG-7 są bardzo skuteczne w zadaniach związanych z automatyczną klasyfikacją dźwięków instrumentów muzycznych. Parametry wyznaczane bezpośrednio w oparciu o widmo dźwięku (*ASE, ASC, ASS, SFM*) wydają się być bardziej istotne niż pozostałe. Ponadto są

one bardziej uniwersalne niż parametry wyznaczane w oparciu o rozkład prążków w widmie sygnału (*HSC, HSD, HSS, HSV*), gdyż mogą być zastosowane do rozpoznawania instrumentów dowolnego typu.

Zbadane algorytmy osiągnęły skuteczność wyraźnie wyższą od 90%, co jest zadowalającym wynikiem. Należy podkreślić, że algorytmy te działały w rygorystycznych warunkach: próbki dźwiękowe pochodziły z dwóch różnych źródeł, a ponadto tylko połowa wszystkich próbek wchodziła w skład zbioru uczącego/wzorca. Można zauważyć, że wynik klasyfikacji poszczególnych instrumentów jest ściśle zależny od pozostałych instrumentów tworzących bazę dźwięków. Instrumenty o podobnym brzmieniu (np. tuba i puzon) i rejestrach (np. puzon i fagot) były najczęściej ze sobą mylone.

Algorytm oparty na metodzie najbliższego sąsiada okazał się nieznacznie lepszy od algorytmów bazujących na sieciach neuronowych. Jednakże cenną cechą sieci neuronowych jest ich zdolność do generalizacji oraz znacznie mniejsza złożoność obliczeniowa procesu klasyfikacji w porównaniu z metodą najbliższego sąsiada.

## 7. PODZIĘKOWANIA

Badania były dofinansowane przez Ministerstwo Nauki i Informatyzacji w ramach grantu nr 4T11D 01422

## 8. BIBLIOGRAFIA

1. Dalka P., Dąbrowski M.: Opracowanie systemu automatycznego rozpoznawania dźwięków instrumentów muzycznych, Praca magisterska, KSM WETI PG, Gdańsk 2003.
2. Information Technology – Multimedia Content Description Interface – Part 4: Audio, International Organization For Standardization, ISO/IEC JTC 1/SC 29, June 2001.
3. Szczuko P., Dalka P., Dąbrowski M., Kostek B.: MPEG-7-based Low-Level Descriptor Effectiveness in the Automatic Musical Sound Classification, 116 Audio Eng. Convention, Preprint No. 6105, Berlin 2004.
4. Marchand S.: An efficient pitch-tracking algorithm using a combination of Fourier transforms, Proc. of Digital Audio Effects '01, Limerick, Ireland, December 6-8, 2001.
5. Kostek B., Wieczorkowska A.: Parametric Representation of Musical Sounds, Archives of Acoustics, vol. 22, No. 1, pp. 3-26, 1997.
6. Kostek B.: Soft computing in acoustics, Physica Verlag, New York, Heidelberg, 1999.
7. Tadeusiewicz R., Flasiński M., Rozpoznawanie obrazów, PWN, Warszawa 1991.

## A SYSTEM FOR MUSICAL INSTRUMENT SOUND CLASSIFICATION

This paper presents a system for the automatic classification of isolated musical instrument sounds. The system consists of three blocks: pitch detection, parametrization and classification. The pitch detection employs the modified Schroeder's algorithm. The parametrization of musical sounds is mainly based on descriptors contained in the MPEG-7 standard. For the purpose of automatic classification three decision algorithms, based on artificial neural networks (ANNs) and the nearest neighbor algorithm, are used. This paper contains a comparison of results obtained by these algorithms and an evaluation of MPEG-7 descriptors significance in the musical instrument sounds classification.