

PORÓWNANIE SKUTECZNOŚCI KLASYFIKACJI GATUNKÓW MUZYCZNYCH W PRZY POMOCY SIECI NEURONOWEJ W ZALEŻNOŚCI OD TYPU DANYCH WEJŚCIOWYCH

Maciej BLASZKE^{1,2}
Damian KOSZEWSKI^{1,2}
Bożena KOSTEK²

¹Katedra Systemów Multimedialnych, Wydział Elektroniki, Telekomunikacji i Informatyki,
Politechnika Gdańska, Narutowicza 11/12, 80-233 Gdańsk

²Laboratorium Akustyki Fonicznej, Wydział Elektroniki, Telekomunikacji i Informatyki,
Politechnika Gdańska, Narutowicza 11/12, 80-233 Gdańsk

mblaszke@multimed.org; damkosze@multimed.org; bokostek@audioakustyka.org

Rozpoznawanie gatunku muzycznego jest jednym z podstawowych elementów inteligentnych systemów tworzenia automatycznych list muzyki. Platformy strumieniowe oferujące taką usługę wymagają rozwiązań, które pozwolą jak najdokładniej określić przynależność utworu do gatunku muzycznego. Aktualny stan wiedzy wskazuje, że najskuteczniejszym klasyfikatorem są sztuczne sieci neuronowe (w tym w wersji uczenia głębokiego), dla których wejście stanowi spektrogram (postać 2D wektora wejściowego), współczynniki MFCC czy wektor parametrów.

Ogólnodostępne narzędzia (np. MIRTtoolbox, czy Python/libROSA) pozwalają na wyznaczenie redundantnej listy parametrów. Wynikowy wektor jest zawsze nadmiarowy ze względu na obecność wysoce skorelowanych parametrów. Przykładem dokładnie wyspecyfikowanego wektora parametrów może być standard MPEG-7 opisany w normie ISO/IEC 15938. Jego zastosowanie pozwala uzyskać nawet 100% skuteczność w kontekście klasyfikacji instrumentu muzycznego. Zawiera on w sobie deskryptory opisujące zarówno przebieg czasowy sygnału, jak i jego strukturę. Można je podzielić na 6 głównych grup:

1. *Basic* - bazujące na wartości próbek sygnału fonicznego (*AudioWaveform Descriptor*, *AudioPower Descriptor*)
2. *Basic Spectral* - Prosta czasowo-częstotliwościowa analiza sygnału (*AudioSpectrumEnvelope*, *AudioSpectrumCentroid*, *AudioSpectrumSpread*, *AudioSpectrumFlatness*)
3. *Spectral Basis* - jednowymiarowa widmowa projekcja sygnału przygotowana w szczególności w celu łatwiejszej klasyfikacji sygnału (*AudioSpectrumBasis*, *AudioSpectrumProjection*)
4. *Signal Parameters* - Informacja o okresowości sygnału (*AudioFundamentalFrequency*, *AudioHarmonicity*)
5. *Timbral Temporal* - charakterystyka czasowo oraz barwa muzyczna (*LogAttackTime*, *TemporalCentroid*)
6. *Timbral Spectral* - deskryptory opisujące zależności liniowo-częstotliwościowe w sygnale (*SpectralCentroid*, *HarmonicSpectralCentroid*, *HarmonicSpectralDeviation*, *HarmonicSpectralSpread*, *HarmonicSpectralVariation*).

Główną motywacją do prowadzenia badań jest potrzeba każdorazowego tworzenia narzędzi do wyznaczania deskryptorów standardu MPEG-7. Istniejące oprogramowanie (różne środowiska programistyczne) zwykle nie zapewnia pełnej możliwości ich wykorzystania. Ponadto, często stosuje się dodat-

kowe, tzw. dedykowane do rozwiązywanego problemu deskryptory. Kolejnym problemem, który się pojawia jest nadmiarowość wektora cech. Aby uzyskać wysoką skuteczność klasyfikacji niezależnie od stosowanego algorytmu, należy dodatkowo stosować metody pozwalające na redukcję proponowanych cech (np. analizę składowych głównych, PCA – *Principal Component Analysis* czy analizę korelacyjną). Autorzy poprzednich prac opracowali sztuczną sieć neuronową, która z 5% błędem je wyznacza. Zaproponowaną sieć można wykorzystać zarówno jako niezależną metodę wyznaczania parametrów, jak również jako część bardziej złożonej architektury, np. jako jej wejście. Przykładem takiej sieci może być między innymi klasyfikator gatunku muzycznego.

W pracy została wykorzystana baza danych SYNAT zawierająca przykłady utworów muzycznych wraz z wyznaczonymi dla nich zgodnie z definicją deskryptorami MPEG-7 oraz dokładną informację o gatunku do którego należy dane nagranie. Ze względu na ograniczenia sieci neuronowej do aproksymacji niskopoziomowych deskryptorów do dalszych eksperymentów wybrano tylko nagrania o długości przynajmniej 8192 próbek. Następnie dokonano losowego podziału na trzy podzbiory: treninowy, walidacyjny i ewaluacyjny.

Do porównania wykorzystano 5 typów danych reprezentujących sygnał:

1. Przebieg czasowy;
2. Spektrogram;
3. Współczynniki MFCC;
4. Deskryptory MPEG-7 z bazy SYNAT;
5. Deskryptory MPEG-7 wyznaczone przy pomocy sieci neuronowej.

Dla całego wybranego zbioru wyznaczono wszystkie wymienione wyżej typy danych wykorzystując 8192 próbek ze środka nagrania. Dalsze eksperymenty wykonano z wykorzystaniem głębokiej sieci neuronowej dokonującej klasyfikacji do którego gatunku należy dane nagranie. W każdym przypadku budowa sieci była jednakowa w obrębie warstw ukrytych, a różnice były w obrębie warstw wejściowych, co było wymagane ze względu na różną długość i typy wektora danych.

Genre recognition is one of the core components of intelligent automatic music list creation systems. Streaming platforms offering such a service require solutions that will allow for the most accurate determination to what music genre a given song belongs. The current state of the art shows that the most effective classifier is an artificial neural network (including its deep learning version). In such cases, the input is often a 2d representation, i.e., a spectrogram, MFCC coefficients, or a feature vector. In the previous work, the authors developed an artificial neural network which, with a 5% error, determines a set of the MPEG-7 descriptors. They can be used, among other things, as input data to the music genre classifier. This work presents a comparison of the effectiveness of such a classifier depending on the type of input data such as raw waveform or spectrogram representation, MFCC, feature vector containing MPEG-7 standard low-level descriptors, or descriptors calculated with the use of a deep neural network.