Analiza i Przetwarzanie Dźwięku - Projekt 1

Anna Hoang
305922

1 Opis aplikacji

Niniejszy projekt został napisany w programie MATLAB R2021a (wersja 9.10). Główną aplikację stanowi Live Script (format .mlx), który zawiera wizualizację wybranych parametrów, przebiegów czasowych wczytanych plików audio oraz charakterystycznych obszarów w analizie sygnałów dźwiękowych. Do obliczeń posłużyły pomocnicze skrypty i funkcje (*.m) w tym gotowe narzędzia z Signal Processing Toolbox i podstawowej biblioteki MATLABa.

Do empirycznego wykrycia pewnych zjawisk dźwiękowych na wybranej części danych użyty został Audio Labeler(?) z dodatku Audio Toolbox. Obserwacje te wykorzystane są następnie do automatycznej detekcji tychże zjawisk również dla innych plików dźwiękowych w głównej aplikacji.

2 Opis metod

2.1 Wyznaczanie parametrów sygnału

Sygnał dźwiękowy można opisać różnymi parametrami i tymi parametrami, którym poświęcony jest niniejszy projekt, są te określone w dziedzinie czasu. Pracując z plikami audio bada się cechy spróbkowanego sygnału dźwiękowego. Każda z próbek ma amplitudę i przebieg czasowy stanowi wykres tychże wartości względem czasu.

Krótkookresowe własności są wyznaczane dla każdej z tzw. ramek, czyli fragmentów o długości kilkudziesięciu ms zawierających określoną liczbę próbek. Długookresowe cechy są badane natomiast na dłuższych fragmentach — klipach trwających po kilka sekund i zawierających wspomniane ramki.

2.1.1 Określenie ramek sygnału

Sposób, w jaki są dobrane te fragmenty trwające 10-40ms ma znaczenie w praktycznych zastosowaniach i są ku temu poświęcone specjalne tzw. funkcje okna, jednak na potrzeby tego projektu posłużono się prostym modelem ramek o długości 10ms, bez nakładania się na siebie ramek.

2.1.2 Podstawowe parametry na poziomie ramki

Funkcje opisujące krótkookresowe cechy zadanego sygnału, mają zadane wzory i ich argumentami są ciągi kolejnych N próbek każdej z ramek $\{s_n\}$. Pierwszym parametrem jest **energia** (Short Time Energy), który definiujemy jako

$$STE(n) = \frac{1}{N} \sum_{i=1}^{N} s_n(i)^2. \tag{1}$$

Na wartości energii bazuje kolejny parametr - **głośność** i stanowi jego pierwiastek dla każdej ramki, czyli

$$p(n) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} s_n(i)^2}. \tag{2}$$

Istotną miarą przy wykrywaniu dźwięczności, bezdźwięczności oraz ciszy razem z parametrami powyżej jest **liczba zmian znaku sygnału** (Zero Crossing Rate). Przedstawia ją poniższy wzór

$$ZCR(n) = \frac{fs}{2N} \sum_{i=2}^{N} |sgn(s_n(i)) - sgn(s_n(i-1))|. \tag{3}$$

Do opisu wysokości dźwięku służy częstotliwość tonu podstawowego (Fundamental Frequnecy F0). Ten parametr określa również częstotliwości wyższych składowych harmonicznych, od których zależy wysokość i barwa dźwięku. Dla nieregularnych sygnałów można jedynie dokonać przybliżenia tejże częstotliwości za pomocą wzorów np. na autokorelację

$$R_n(r) = \frac{1}{N-r} \sum_{i=1}^{N-r} s_n(i) s_n(i+r) \tag{4}$$

gdzie r stanowi przesunięcie czasowe względem wartości, którego maksimum lokalne z funkcji autokorelacji można przybliżać jako szukane F0 czyli

$$F0(n) = \max_{r=0,1,\ldots,m-1} R_n(r) \tag{5}$$

dla $m < N$ przesunięć.

2.1.3 Inne parametry użyte w eksperymentach

W celu rozróżnienia fragmentów sygnału zawierającego mowę od fragmentów z muzyką analizuje się niskie wartości energii ramek w wybranym klipie. Wykorzystywany do tego jest odsetek liczby ramek o energii poniżej połowy średniej energii w obrębie całego klipu (Low Short Time Energy Ratio) czyli

$$LSTER(n_c) = \frac{1}{N_f} \sum_{j=1}^{N_f} \left[1 + sgn\left(\frac{1}{2} avSTE(n_c) - STE(j) \right) \right] \tag{6}$$

gdzie N_f - liczba ramek, a n_c - numer klipu oraz $avSTE(n_c)$ to jego średnie STE.

2.2 Progi parametrów do detekcji zjawisk

Nagrania na potrzeby projektu zostały wykonane w jednakowych warunkach, tym samym sprzętem, co stanowi pewną podstawę do uogólnienia pewnych przedziałów wartości wybranych parametrów potrzebnych do zaobserwowania w nich poszczególnych zjawisk.

Część plików jednej z osób mówiących została poddana ręcznej klasyfikacji. Pierwszym etapem było wyeksportowanie wartości parametrów sygnałów do arkusza kalkulacyjnego, z dodatkową kolumną na etykiety. Dodanie etykiet nastąpiło po empirycznej ocenie, jakie przedziały czasowe obejmowały partie spełniające wybrane cechy. Z tak sklasyfikowanych danych można było przybliżyć progi parametrów, dla których zachodzi wybrane zjawisko.

2.2.1 Cisza, bezdźwięczność i dźwięczność

Fragmenty bezdźwięczne charakteryzują się niską energią i wysokim ZCR, tym samym elementy dźwięczne będą się charakteryzować wyższym STE. Średnie i maksymalne wartości tych parametrów wyznaczone zostały dla każdego zjawiska, w tym ciszy ze względu na szum w tle nagrań. Jako przedziały klasyfikacji wybrane zostały eksperymentalnie kombinacje maksimów, minimów i wartości średnich. Na koniec te progi zostały zaaplikowane do pozostałych plików z bazy nagrań dla porównania.

	ZCR	STE
cisza	$\leqslant$ 1713.41	$\leqslant$0.002
bezdźwięczność	$\leqslant$ 1403.52	$\geqslant$0.025
dźwięczność	$\geqslant$ 1738.374	$\leqslant$0.002

Tabela 1: Wybrane progi dla ciszy, bezdźwięczności i dźwięczności.

2.2.2 Mowa a muzyka

Do rozróżnienia mowy od muzyki posłużono się miarą $LSTER$, która dla mowy powinna być dużo wyższa. Wybrano do empirycznego badania: pliki z samą mową (`"pobudka.wav", "wstajemy.wav"`) oraz samą melodią (`"budzik1.wav" "budzik2.wav"`) i zaaplikowano wyznaczone empirycznie progi $LSTER$ do rozróżnienia mowy od muzyki w pliku `"czas wstawać.wav"`, zawierającym elementy muzyczne jak i mowę.

	dolna wartość LSTER	górna wartość LSTER
cisza	-	0.5
muzyka	0.5	1.25
mowa	1.2	-

Tabela 2: Wybrane progi dla ciszy, bezdźwięczności i dźwięczności.

3 Wyniki i wnioski

3.1 Cisza, bezdźwięczność i dźwięczność

Dla jednego z plików, na którym zaaplikowane zostały wybrane progi, poprawność klasyfikacji wynosiła ok. 66% w wyspecyfikowanych ręcznie obszarach.

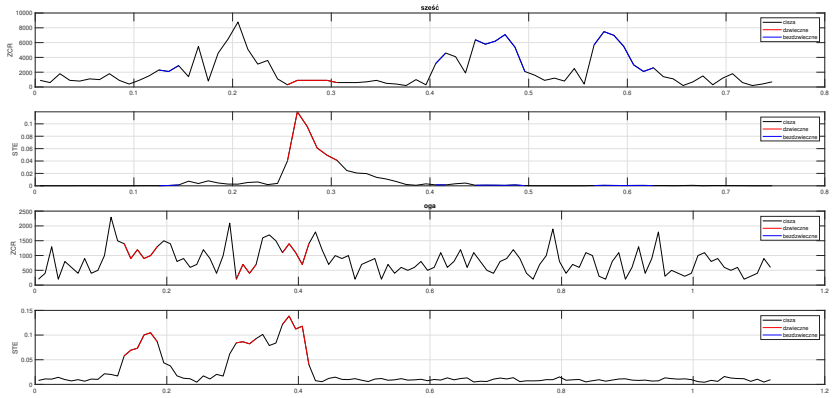
time	STE	ZCR	label	automatic	accurate
0,11	0,04	897,96	silence	silence	1
0,12	0,05	1496,61	silence	silence	1
0,13	0,04	2294,80	unvoiced	unvoiced	1
0,14	0,12	2095,25	unvoiced	unvoiced	1
0,15	0,30	2893,44	unvoiced	unvoiced	1
0,24	0,43	3591,86	unvoiced	silence	0,5
0,25	0,86	1097,51	unvoiced	silence	0,5
0,26	8,85	299,32	voiced	voiced	1
0,27	26,39	897,96	voiced	voiced	1
0,28	21,15	897,96	voiced	voiced	1
0,29	13,50	897,96	voiced	voiced	1
0,30	10,91	897,96	voiced	voiced	1
0,31	9,11	598,64	voiced	voiced	1
0,32	5,45	598,64	voiced	voiced	1
0,33	4,55	598,64	voiced	silence	0
0,34	4,35	698,42	voiced	silence	0
0,35	3,03	897,96	voiced	silence	0
0,36	2,42	498,87	voiced	silence	0
0,37	1,53	399,10	voiced	silence	0
0,38	0,48	199,55	unvoiced	silence	0,5
0,39	0,19	997,74	unvoiced	silence	0,5
0,40	0,70	299,32	unvoiced	silence	0,5

Tabela 3: Wyniki dla wybranych ramek z 40ms nagrania 'sześć.wav'.

Każde z nagrań miało bardzo różne zakresy parametrów ze względu na naturalną głośność, z jaką się wypowiadały osoby nagrywane oraz odległością, w której się znajdywały od mikrofonu. Można również zauważyć, że w wielu miejscach szum tła w wielu miejscach był niemal nierozróżnialny zarówno od mowy bezdźwięcznej, jak i tej dźwięcznej.

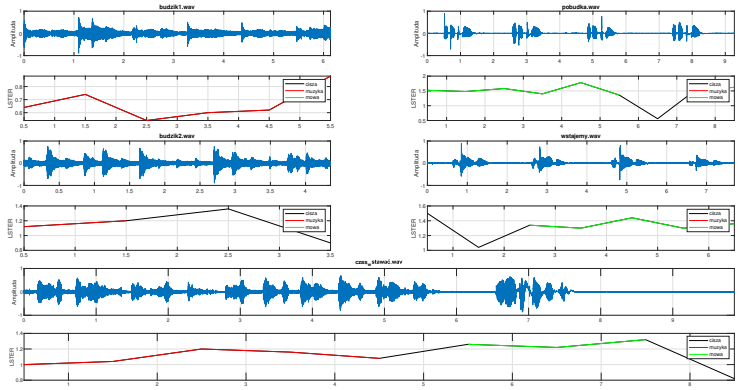
3.2 Mowa a muzyka

Do rozróżniania mowy od muzyki zastosowany został parametr $LSTER$, który mierzy się na poziomie klipów, co należało uwzględnić w eksperymentach, żeby nie dzielić plików o mniej niż 2sek. Mimo małej liczby danych wyjściowych, jakie się otrzymuje dla tej miary przy krótkich sygnałach audio, wyniki okazały się dość dobre. Nagrania na potrzeby tego eksperymentu były wykonane przez



Rysunek 1: Wygenerowane automatycznie obszary SUV.

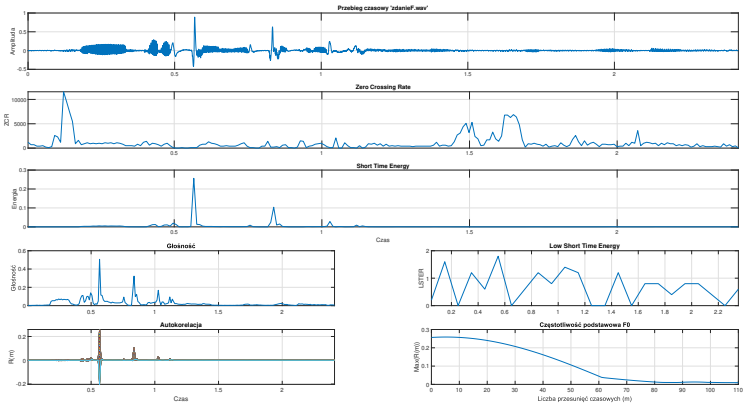
jedną osobę, a muzyka była grana z tego samego urządzenia, co mogło polepszyć dokładność wyników.



Rysunek 2: Wygenerowane automatycznie obszary mowa/muzyka.

3.2.1 Głosy damskie i męskie

Niektóre źródła podają, że typowa częstotliwość podstawowa dla mowy dźwięcznej u mężczyzn jest w przedziale 0.085-0.155 kHz natomiast u dorosłych kobiet, ta wartość waha się między 0.165 a 0.255 kHz. Wyniki dla zastosowanych w programie wzorów są rzędu tej wielkości, w dużym przybliżeniu.

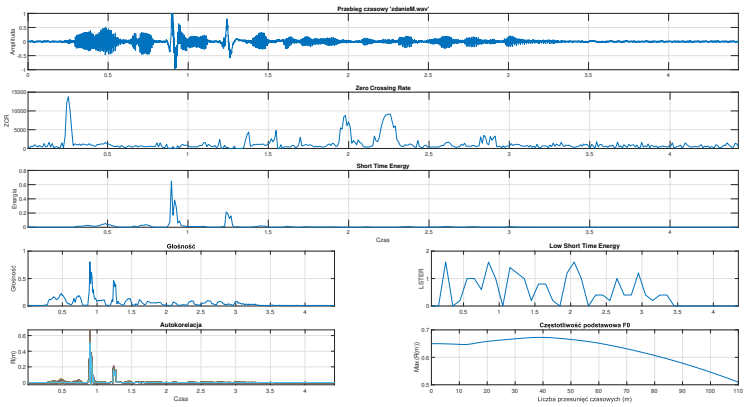


Rysunek 3: Parametry dla głosu damskiego.

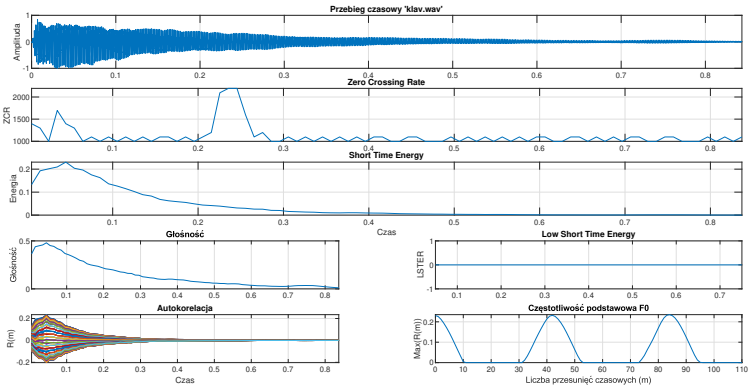
3.2.2 Różne instrumenty

Instrumenty muzyczne produkują dźwięk o określonej wysokości, którą przybliżaliśmy w parametrze F0. Są źródłami różnych fal akustycznych co widać po zestawieniu poniższych wykresów.

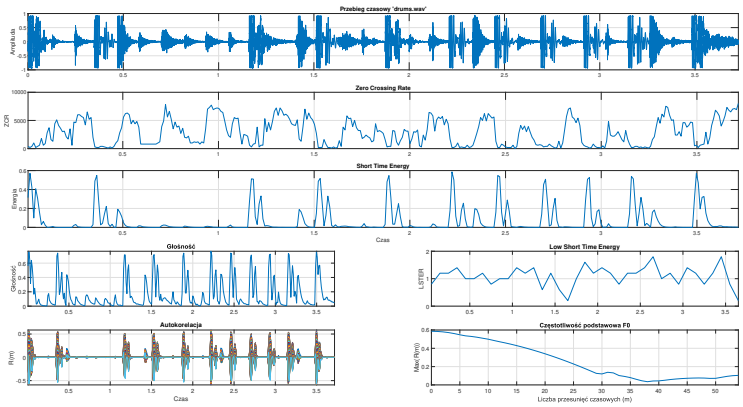
Różnice w zachowaniu pozostałych funkcji nie są drastyczne, ale po zmierzeniu częstotliwości podstawowej te melodie widać, że są inne.



Rysunek 4: Parametry dla głosu męskiego.



Rysunek 5: Pianino



Rysunek 6: Perkusja