**LABORATORIUM 9**

**1.Sygnał audio**

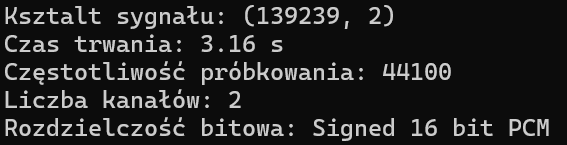
Zadanie 1. Korzystając z mikrofonu w komputerze lub dyktafonu w smartfonie nagraj sekwencję audio ”Jestem studentem informatyki”.

Zadanie 2. Wczytaj nagrany plik dźwiękowy do zmiennej s. Odczytaj/oblicz podstawowe parametry sygnału (czas trwania, częstotliwość próbkowania, rozdzielczość bitowa, liczba kanałów).

Obraz zawierający tekst, zrzut ekranu, Czcionka

Zawartość wygenerowana przez sztuczną inteligencję może być niepoprawna.KOD NAPISANY

REZULTAT



Parametry sygnału:

Czas trwania – można go obliczyć, dzieląc liczbę próbek (pierwszy wymiar kształtu tablicy s) przez częstotliwość próbkowania.

Częstotliwość próbkowania – jest zwracana jako drugi wynik funkcji sf.read() i oznacza, ile próbek dźwięku jest rejestrowanych na sekundę.

Liczba kanałów – to drugi wymiar naszych próbek. Dla sygnału mono będzie to wartość jeden my mamy sygnał stereo który ma dwa kanały.

Rozdzielczość bitowa – za pomocą funkcji sf.info można uzyskać rozdzielczość bitowa.Wynik pokazuje, że sygnał jest zakodowany w formacie PCM, a każda próbka ma 16 bitów (16-bitowa rozdzielczość próbkowania).

Zadanie 3

Wyświetlić sygnał tak, aby na osi poziomej znajdowała się jednostka czasu [ms] (konieczność przeliczenia zakresu). Jeżeli po wczytaniu sygnał nie jest znormalizowany, należy przeprowadzić normalizację wartości do zakresu [-1;1]. W przypadku sygnału stereo wykorzystaj dowolny z kanałów: lewy lub prawy

Obraz zawierający tekst, zrzut ekranu, Czcionka

Zawartość wygenerowana przez sztuczną inteligencję może być niepoprawna.

REZULTAT

Obraz zawierający diagram, Wykres, zrzut ekranu, tekst

Zawartość wygenerowana przez sztuczną inteligencję może być niepoprawna.

Zadanie 4. Sprawdź, czy dynamika sygnału jest odpowiednia? Czy zakres amplitudy jest odpowiednio wykorzystany? Czy nie występuje przesterowanie? Jaka jest amplituda szumu na początku i na końcu nagrania? Czy szum ma charakter losowy?

**Czy dynamika sygnału jest odpowiednia?**

Tak dynamika sygnału jest odpowiednia. Jest dużo wyraźnych zmian między cichymi a głośnymi fragmentami dźwięku.

**Czy zakres amplitudy jest odpowiednio wykorzystany?**

Zakres jest w miarę dobrze wykorzystany. Wartości mieszczą się w przedziale od -0.6 do 0.6 z zakresu |-1,1|. Można by było dodać trochę głośności do wypowiedzi by lepiej wykorzystać cały dostępny zakres amplitudy.

**Czy nie występuje przesterowanie?**

Nie występuje przesterowanie. Nigdzie sygnał nie przekracza zakresu |-1,1|

**Jaka jest amplituda szumu na początku i na końcu nagrania? Czy szum ma charakter losowy?**

Amplituda szumu na początku i końcu nagrania jest niewielka i stanowi drobne zakłócenia tła. Szum ma charakter losowy, nie widać żadnych powtarzalnych wzorców ani schematów w jego przebiegu.

**2. Zastosowanie okien kroczących**

Zadanie 1. Podzielić sygnał na ramki (okna) długości 10 ms i obliczyć dla każdej ramki dwie statystyki – funkcję energii E oraz funkcję przejść przez zero Z

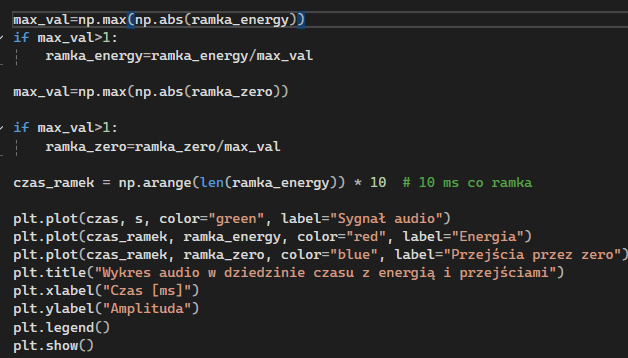
Obraz zawierający tekst, zrzut ekranu, oprogramowanie, Czcionka

Zawartość wygenerowana przez sztuczną inteligencję może być niepoprawna.

Obliczamy ilość próbek przypadających na jedno okno. Następnie przechodzimy przez sygnał w krokach równych tej liczbie próbek. W każdej iteracji wycinamy ramkę sygnału i zapisujemy ją do listy. Dla każdej ramki obliczamy jej energię oraz liczbę przejść przez zero, korzystając z wcześniej zdefiniowanych funkcji. Wyniki zapisujemy odpowiednio do list ramka\_energy i ramka\_zero.

Zadanie 2.Uzyskane funkcje (wektory) energii E oraz przejść przez zero Z należy znormalizować. Dokonać wizualizacji tych funkcji.

KOD NAPISANY



Sprawdzamy, a w razie potrzeby normalizujemy energie i przejście przez zero.Tworzymy tablice z czasami odpowiadającymi każdej ramce, przyjmując długość ramki10ms . Następnie wizualizujemy

REZULTAT

Obraz zawierający zrzut ekranu, Wykres, diagram, tekst

Zawartość wygenerowana przez sztuczną inteligencję może być niepoprawna.

Zadanie 3. Na co wskazują maksima oraz minima obliczonych funkcji? Czy można ich użyć do podziału sygnału na segmenty dźwięczne i bezdźwięczne? Jakie typy głosek można rozdzielić automatycznie na podstawie funkcji Z i E? Czy istnieją sekwencje głosek nierozdzielnych?

**Na co wskazują maksima oraz minima obliczonych funkcji?**

Maksima energii oznaczają głośne fragmenty sygnału, tam energia jest największa. Natomiast minima oznaczają ciche fragmenty.

W przypadku liczby przejść przez zero, wysoka wartość oznacza, że sygnał często zmienia znak co często odpowiada dźwiękom bezdźwięcznym takich jak sz, s, f. lub szumowi. Niska wartość pokazuje małą ilość drgań czyli dźwięki dźwięczne takie jak samogłoski.

**Czy można ich użyć do podziału sygnału na segmenty dźwięczne i bezdźwięczne?**

Można tego użyć do podziału segmentów na dźwięczne i bezdźwięczne. Wystarczy ustalić odpowiednie progi dla funkcji, żeby zidentyfikować fragmenty z mową i wyróżnić w nich dźwięczne i bezdźwięczne głoski. Używając kombinacji energii i przejść przez zero.

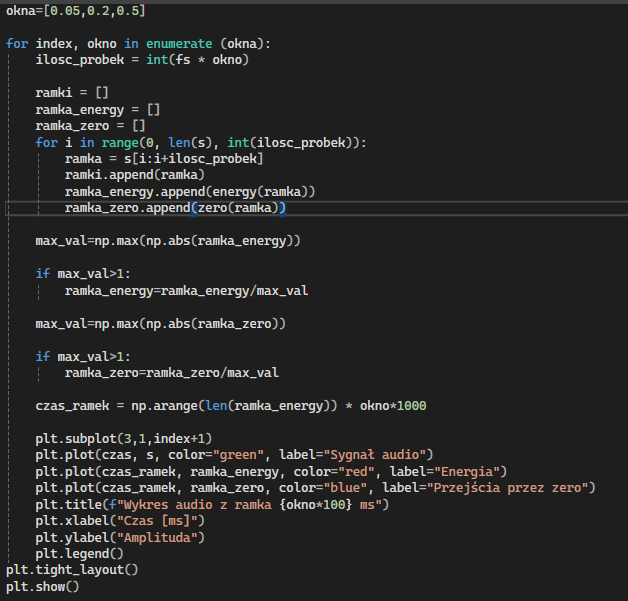
**Jakie typy głosek można rozdzielić automatycznie na podstawie funkcji Z i E?**

Jak opisywałem powyżej można automatycznie rozdzielić na głoski dźwięczne takie jak m,n,l,a jak i bezdźwięczne s,f,sz

**Czy istnieją sekwencje głosek nierozdzielnych?**

W rozróżnieniu głosek może wystąpić przy sekwencjach głosek o podobnych cechach akustycznych. Przykładowo dwie bezdźwięczne s i f lub dwie dźwięczne a-m. Wtedy rozdzielenie może być niedokładne co można zdefiniować jako głoski nierozdzielne.

Zadanie 4. Zbadać jaki wpływ na segmentowanie sygnału ma długość okna. Sprawdzić i pokazać wyniki dla okien równych: 5ms, 20ms i 50ms. Jakie elementy nagrania zostają rozdzielone przy bardzo długim, a jakie przy bardzo krótkim oknie?

KOD NAPISANY

REZULTAT

Obraz zawierający tekst, zrzut ekranu, diagram, Wykres

Zawartość wygenerowana przez sztuczną inteligencję może być niepoprawna.

Jak widać przy wizualizacji przy krótkim oknie wykrywamy szybkie zmiany. Wykrywamy ich dużo przez co wzrasta wrażliwość na zakłócenia np. szumy. Przy oknie 20ms wykrywanie dźwięków jest stabilniejsze. Dobrze są wykrywane dłuższe dźwięki. Przy długim oknie takim jak 50 ms spółgłoski, zlewają się z otoczeni, widoczne są bardziej ogólne zmiany głośności np. pauzy.

Zadanie 5. Podzielić sygnał na nakładające się ramki np. w stopniu 50% i przeprowadzić powyższe analizy. Jak wpływa nakładanie ramek na precyzję granic segmentów?

ZMIANA W KODZIE





W celu podzielenia sygnału zmniejszyłem krok pętli o połowę oraz przy tworzeniu czasu dla każdej ramki podzieliłem ją przez dwa.

REZULTAT

Obraz zawierający tekst, zrzut ekranu, diagram, Czcionka

Zawartość wygenerowana przez sztuczną inteligencję może być niepoprawna.

Nakładanie ramek poprawia precyzję wykrywania granic segmentów, ponieważ ramki częściowo się pokrywają, co pozwala lepiej wychwycić szybkie zmiany w sygnale i uniknąć utraty ważnych informacji. Dzięki temu segmentacja jest dokładniejsza i bardziej płynna.

**3. Analiza częstotliwości**

Zadanie 1. Zlokalizować (manualnie) fragment nagrania stanowiący samogłoskę ustną i skopiować jej fragment długości 2048 próbek do nowej zmiennej.

Obraz zawierający tekst, Czcionka, zrzut ekranu, typografia

Zawartość wygenerowana przez sztuczną inteligencję może być niepoprawna.

Wybrałem fragment który występuje w czasie 1350 ms, ponieważ to tam występuje największa energia.

Zadanie 2. Dokonać maskowania sygnału oknem Hamminga (wykorzystać wbudowaną funkcję numpy.hamming).



Zadanie 3. Obliczyć logarytmiczne widmo amplitudowe dla badanego (zamaskowanego) okna. Przydatne funkcje: numpy.log oraz numpy.abs.

Obraz zawierający tekst, Czcionka, zrzut ekranu, linia

Zawartość wygenerowana przez sztuczną inteligencję może być niepoprawna.

WIZUALIZACJA Zadania 1-3

KOD NAPISANY

Obraz zawierający tekst, zrzut ekranu, oprogramowanie, Czcionka

Zawartość wygenerowana przez sztuczną inteligencję może być niepoprawna.

Obraz zawierający tekst, diagram, zrzut ekranu, Wykres

Zawartość wygenerowana przez sztuczną inteligencję może być niepoprawna.REZULTAT

Zadanie 4. Wyświetlić widmo w taki sposób, aby na osi poziomej znajdowała się jednostka częstotliwości Hz w zakresie 0 - 10000 Hz. Schematyczne zobrazowanie wykonanych do tej pory zadań ukazane zostało na poniższym rysunku.

MODYFIKACJA KODU

Obraz zawierający tekst, zrzut ekranu, Czcionka

Zawartość wygenerowana przez sztuczną inteligencję może być niepoprawna.

REZULTAT

Obraz zawierający tekst, zrzut ekranu, Wykres, diagram

Zawartość wygenerowana przez sztuczną inteligencję może być niepoprawna.

Zadanie 5. Z wykresu odczytać F0, czyli częstotliwość podstawową (częstotliwość drgania strun głosowych) na podstawie pierwszego dominującego maksimum w przebiegu widma.

Obraz zawierający tekst, diagram, linia, Wykres

Zawartość wygenerowana przez sztuczną inteligencję może być niepoprawna.

Częstotliwością podstawową jest pierwszy pik który zaczyna się na około 190 hz. Jest to nasze pierwsze dominujące maksimum z widma. Jest to również pierwsza częstotliwość serii maksimów w strukturze harmonicznej.

Zadanie 6. Sprawdzić, które z badanych głosek posiadają strukturę harmoniczną i można odczytać z nich F0 (częstotliwość drgania strun głosowych)? Czym różni się wizualnie widmo głosek dźwięcznych od bezdźwięcznych? Czy widma samogłosek różnią się między sobą? W jaki sposób? Czy częstotliwość F0 jest stała dla tego samego nagrania, ale różnych głosek?

**Które głoski posiadają strukturę harmoniczną i można odczytać F0?**

Strukturę harmoniczną i wyraźne F0 mają głoski dźwięczne, szczególnie samogłoski oraz dźwięczne spółgłoski jak m, n, l, r , ponieważ są one wynikiem drgań strun głosowych. Bezdźwięczne nie mają struktury harmonicznej

**Czym różni się widmo głosek dźwięcznych od bezdźwięcznych?**

Widmo głosek dźwięcznych ma wyraźne, równomiernie rozmieszczone piki harmoniczne. Widmo bezdźwięcznych jest bardziej rozmyte, szumowe, bez wyraźnych harmonicznych.

**Czy widma samogłosek różnią się między sobą? W jaki sposób?**

Tak, każda samogłoska ma inny układ formantów, głównie F1 i F2, które nadają charakterystyczne brzmienie każdej samogłosce. Na widmie widać je jako piki o różnym położeniu.

**Czy częstotliwość F0 jest stała dla tego samego nagrania, ale różnych głosek?**

Nie, F0 może się lekko zmieniać między głoskami, nawet w tym samym nagraniu, zależnie od intonacji, akcentu itp.. Jednak zazwyczaj w jednym krótkim fragmencie mowy pozostaje stabilnie zbliżona.

**4. Rozpoznawanie samogłosek ustnych /a, e, i, o u, y/**

Zadanie 1. Zlokalizować fragment nagrania stanowiący samogłoskę ustną i skopiować jej fragment długości 2048 próbek do zmiennej okno.

Obraz zawierający Czcionka, tekst, zrzut ekranu, Grafika

Zawartość wygenerowana przez sztuczną inteligencję może być niepoprawna.

Ponownie wybieramy ten sam fragment.

Zadanie 2. Wyznaczyć dla tego fragmentu p=20 współczynników liniowego filtra LPC. Wykorzystać należy gotową funkcję obliczającą Liniowe Kodowanie Predykcyjne (LPC) z pakietu librosa.



Zadanie 3. Dowiedzieć się i krótko opisać, co to jest i do czego służy Liniowe Kodowanie Predykcyjne

Liniowe Kodowanie Predykcyjne (LPC) to metoda przewidywania kolejnych próbek sygnału na podstawie wcześniejszych. Umożliwia kompresję i analizę mowy, opisując właściwości toru głosowego. Stosowane jest m.in. w syntezie, rozpoznawaniu i kompresji mowy.

Zadanie 4. Uzupełnić otrzymany wektor a długości p, zerami do długości okna sygnału (2048).

Obraz zawierający tekst, Czcionka, zrzut ekranu, numer

Zawartość wygenerowana przez sztuczną inteligencję może być niepoprawna.

Zadanie 5. Wyznaczyć wygładzone widmo amplitudowe na bazie wektora a: widmoLPC = log(abs(fft(a))). Odbić otrzymane wygładzone widmo w poziomie (pomnożyć przez -1). Nałożyć wykres na właściwe widmo amplitudowe fragmentu sygnału tak, aby oba znalazły się na podobnej wysokości na osi y (prawdopodobnie będzie konieczność przeskalowania lub/i przesunięcia). Należy zadbać o prawidłową oś częstotliwości [Hz]! Przykład na poniższym rysunku.

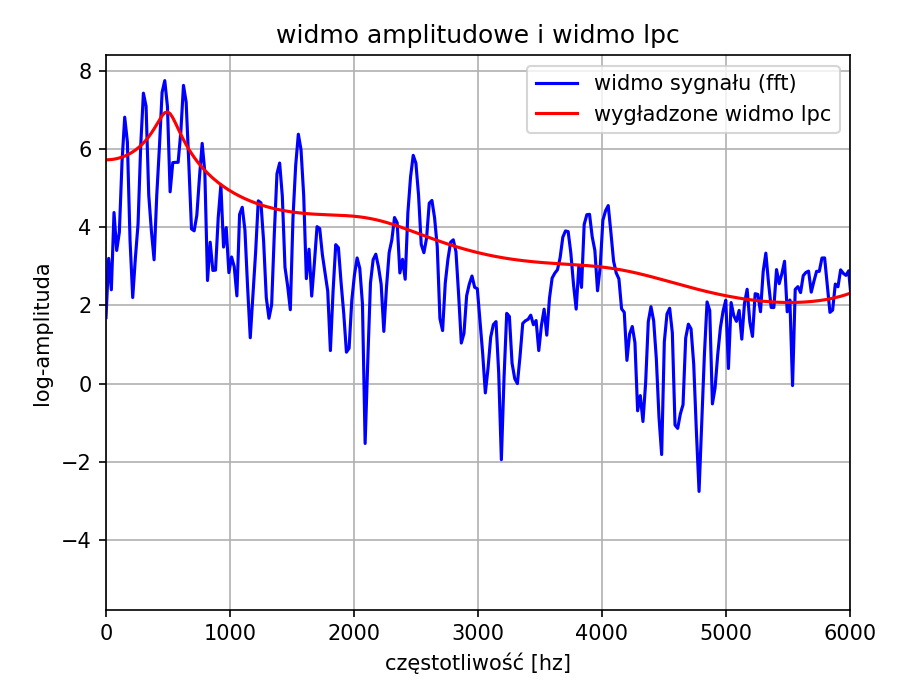
CAŁY KOD ZADANIA 1-5

Obraz zawierający tekst, zrzut ekranu, Czcionka

Zawartość wygenerowana przez sztuczną inteligencję może być niepoprawna.

Oprócz wcześniej wykonanych punktów, wyznaczamy wygładzone widmo amplitudowe za pomocą funkcji fft z scipy.fftpack. Następnie przekształcamy je na skalę logarytmiczną i mnożąc przez -1, odbijamy widmo w poziomie. W celu lepszego dopasowania na wykresie, każde widmo przeskalowujemy, odejmując jego średnią wartość. Na końcu prezentujemy wyniki na wykresie.

REZULTAT



Zadanie 6. Wyznacz dla kilku samogłosek częstotliwości F1 i F2. Odnieś wyznaczone wartości do wzorcowego wykresu. Jakie samogłoski udało się właściwie rozpoznać? Jak wpływa ilość wybranych współczynników filtra p na proces rozpoznawania? Przetestować kilka możliwości. W przypadku, gdy istnieje problem ekstrakcji niezakłóconych samogłosek z nagrania należy dokonać kolejnego nagrania audio, tym razem zawierającego izolowane samogłoski: /a,e,i,o,u,y/.

Część 1

Obraz zawierający tekst, zrzut ekranu, Czcionka

Zawartość wygenerowana przez sztuczną inteligencję może być niepoprawna.

Obraz zawierający tekst, diagram, Wykres, linia

Zawartość wygenerowana przez sztuczną inteligencję może być niepoprawna.REZULTAT

Z powodu występującego problemu z rozpoznaniem samogłosek z nagrania wykonałem kolejne nagranie audio, tym razem zawierające izolowane samogłoski. Jak widać na wykresie teraz można bez problemu wybrać fragmenty z samogłoskami i rozpoznania ich w następnych etapach.

CZĘŚĆ 2

Iterujemy po wybranych fragmentach. Dla każdego obliczamy okno. Tworzymy zmienne które będą przechowywać p z odpowiadającą najmniejszą częstotliwością F2 jak i częstotliwości F1 i F2 odpowiadającym danemu p. Iterujemy po danych wartościach p. I dla każdej wartości sprawdzamy częstliwości i je wyświetlamy. Jeśli Częstoliwość F1 jest mniejsza niż obecnie zapisana inicjalizujemy zmienne. Na koniec każdej iteracji dodajemy najlepsze częstotliwości do x i y.

Dla każdego wskazanego fragmentu który ma wysoką energie odpowiednią dla samogłosek, wycinamy okno o długości 2048 próbek. Iterujemy po różnych wartościach p, obliczając współczynniki predykcyjne. Uzupełniamy je zerami do tej samej długości, aby wykonać FFT. Logarytmujemy i odwracamy widmo, a następnie wyszukujemy dwa pierwsze maksima, które traktujemy jako formanty F1 i F2. Dla każdej wartości p sprawdzamy, czy uzyskany F2 jest mniejszy niż wcześniej zapisany. Jeśli tak, zapisujemy aktualne F1, F2 i p jako najlepsze. Wykonujemy tą dodatkową akcje w celu zwizualizowania tych punktów. Wybieramy najniższe F2, ponieważ niektóre wartości p dają nierealistycznie wysokie częstotliwości, które uniemożliwiają rozpoznanie samogłoski. Po zakończeniu dla danego fragmentu, do list x i y dodajemy te najlepsze wartości formantów.

CZĘŚĆ 3

Obraz zawierający tekst, zrzut ekranu, Czcionka

Zawartość wygenerowana przez sztuczną inteligencję może być niepoprawna.

Wyświetlamy w podobny sposób jak zostały przedstawione poszczególne samogłoski na rysunku w karcie pracy.

Obraz zawierający tekst, zrzut ekranu, oprogramowanie

Zawartość wygenerowana przez sztuczną inteligencję może być niepoprawna.REZULTAT KODU

Obraz zawierający tekst, zrzut ekranu, Czcionka, czarne i białe

Zawartość wygenerowana przez sztuczną inteligencję może być niepoprawna.

**Jak wpływa ilość wybranych współczynników filtra p na proces rozpoznawania?**

Wartość p wpływa na dokładność odwzorowania widma. Małe p=15 nie uchwyci wszystkich formantów daję często zbyt duży przez to wynik. Zbyt duże P=70 staję się zbyt dokładny i prowadzi do zniekształconych, lub w ogóle niewyraźnych formant. Co powoduję brakiem znalezienia formanty jak dla wartości 1200 i 2400. Optymalny dobór p=25,30 pozwala na lepsze rozpoznanie głosek, bo widmo jest gładkie i zawiera realne formanty.

Obraz zawierający tekst, zrzut ekranu, wyświetlacz, oprogramowanie

Zawartość wygenerowana przez sztuczną inteligencję może być niepoprawna.WIZUALNY REZULTAT

**Jakie samogłoski udało się właściwie rozpoznać?**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Wartość** | **Oryginalna samogłoska** | **Częstotliwości [F1,F2]** | **Przewidziane** | **Czy przewidziane?** |
| 1200 | a | 689/1291 | a | **TAK** |
| 2400 | e | 581/1593 | e/a | **TAK** |
| 3700 | i | 215/2433 | i/y | **TAK** |
| 5250 | o | 516/831 | o | **TAK** |
| 7100 | u | 279/2433 | i/y | **NIE** |
| 8600 | y | 387/1658 | y | **TAK** |

Nie została jedynie przewidziana samogłoska u. Dla samogłoski zgadza się F1 natomiast nie zgadza się F2 dla żadnego p. Możliwe, że u wybrzmiało na audio bardzo podobnie jak y. Dla reszty przewidziano poprawnie. W przypadkach, gdzie częstotliwości F1 i F2 mieściły się w zakresach odpowiadających dwóm podobnym samogłoskom, np. **e/a**, zauważalna była różna skuteczność w zależności od wartości p. Przy niepoprawnie przewidzianej samogłosce niektóre p nie przewidywało tej samogłoski, a gdy trafiało w zakres tej samogłoski, to często na jego granicy. Ta samogłoska, która została ostatecznie poprawnie rozpoznana, miała wartości formantów lepiej dopasowane do typowych zakresów.