AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE
Wydział Elektrotechniki, Automatyki, Informatyki i Inżynierii Biomedycznej

Projekt dyplomowy

Określenie cech osobniczych mówcy na podstawie jego zarejestrowanych wypowiedzi

Determining the individual characteristics of the speaker on the basis of his recorded statements

Autor: *Łukasz Marcin Bednarek*
Kierunek studiów: Automatyka i Robotyka
Opiekun pracy: *dr inż. Andrzej Izworski*

Kraków, 2021

Spis treści

Wstęp

Praca ta została stworzona w celu przedstawienia sposobów na określenie cech osobniczych mówcy poprzez analizę jego zarejestrowanych wypowiedzi. Ukazane zostały metody oraz algorytmy wykorzystywane w procesie ekstrakcji parametrów głosu takich jak ton podstawowy, formanty oraz współczynniki MFC. W pracy zawarto także analizę skuteczności wykorzystania wymienionych cech do rozpoznawania płci oraz wieku mówcy przy wykorzystaniu Mieszanych Modeli Gaussowskich.

Automatyczne sposoby na rozpoznawanie cech osobniczych mówcy są tematem ciągle rozwijanym, a ich znaczenie wzrasta wraz ze wzrostem rozwoju interfejsów człowiek-maszyna. Określenie cech osoby komunikującej się z systemem jest niezbędne w celu ich adaptacji do potrzeb użytkownika.

Rozwój systemów pozwalających na ekstrakcję odpowiednich parametrów głosu oraz powiązania ich z jak największą liczbą cech osobniczych pożądany jest przez przemysł reklamowy, medyczny a także kryminalistykę oraz biometrię. Przykładami mogą być spersonalizowane reklamy stworzone na podstawie informacji o wieku oraz płci, diagnoza problemów zdrowotnych na podstawie odstępu parametrów głosu od normy, zawężanie liczby podejrzanych na podstawie nagrania głosu, a także tworzenie zabezpieczeń oraz kluczy opartych o parametry głosu.

W pierwszym rozdziale pracy omówione zostały wykorzystane bazy nagrań. Następnie wybrane metody przetwarzania wstępnego sygnału oraz cel ich zastosowania. Trzeci rozdział poświęcony został na opis sposobu tworzenia Mieszanych Modeli Gaussowskich wykorzystywanych w dalszej części pracy. W kolejnych pokazane zostały wybrane metody pozyskiwania parametrów głosu oraz ich analiza wraz z zastosowaniami do rozpoznawania płci oraz wieku osoby mówiącej, a także skuteczność ich wykorzystania w połączeniu z GMM.

1. Dane wykorzystane w badaniach

Do eksperymentów przeprowadzonych w pracy wykorzystane zostały dwie bazy nagrań głosowych. Pierwszą bazą były próbki głosu pozyskane w znanych warunkach przy wykorzystaniu aparatu Sony Nex-5, wraz z dołączonym mikrofonem Sony ECMSST1 Compact Stereo Microphone. Cztery badane osoby zostały poproszone o wymówienie kolejno zbioru samogłosek trzy razy zaczynając dwie sekundy po rozpoczęciu nagrywania. Badanymi osobami były dwie kobiety oraz dwoje mężczyzn w wieku dwadzieścia oraz czterdzieści osiem lat.

Jako drugą bazę danych wykorzystano publiczny zbiór nagrań głosu w języku polskim „Common Voice" stworzony dzięki inicjatywie Mozilli [1]. Baza liczyła 100082 nagrania, jednak do eksperymentów wykorzystane zostały tylko nagrania posiadające oznaczenia nadane przez autora dotyczące płci lub wieku osoby mówiącej. Do analizy płci użyto 2124 podpisane nagrania z czego 156 nagrań należało do kobiet. Ilość nagrań posiadających oznaczenia dotycząca wieku osoby mówiącej została przedstawiona w tabeli (Tab. 1).

Tabela 1. Ilość nagrań wykorzystanych do badania z podziałem na płeć oraz wiek mówcy.

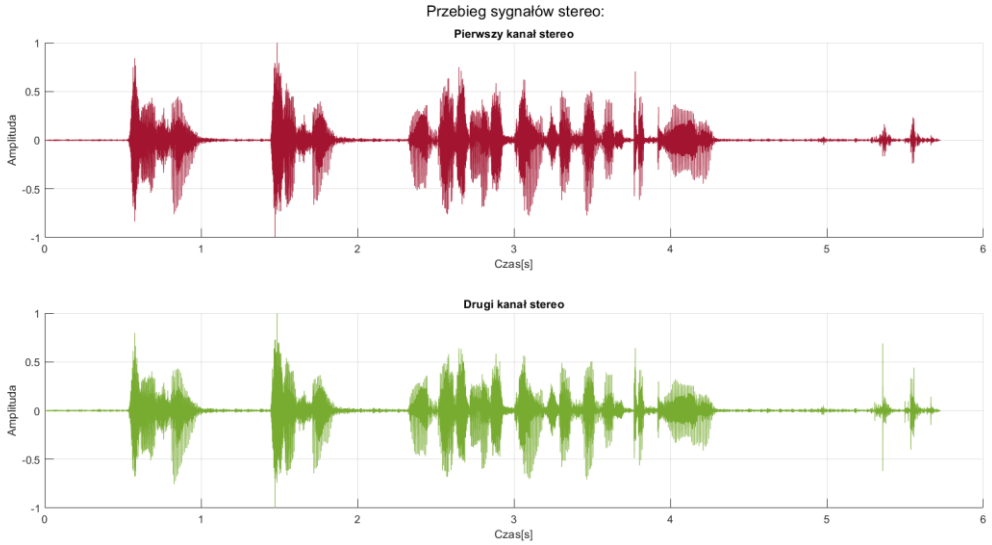
Wiek mówcy [lata]	Ilość nagrań		
	Męskie	Żeńskie	Suma
13-19	160	17	177
20-29	911	97	1008
30-39	812	41	853
40-49	72	0	72
50-59	6	1	7

2. Przetwarzanie wstępne sygnału

Aby poprawić jakość analizy danych pozyskanych z nagrania często potrzebne jest wstępne przetworzenie sygnału (ang. Signal Preprocessing). Termin ten odnosi się do działań mających na celu wzmocnienie pożądanych cech sygnału, wytłumieniu zakłóceń lub dostosowanie sygnału do dalszych działań. Operacje oraz metody, które zostały wykorzystane w trakcie badania celem przygotowania sygnału do dalszej analizy zostały opisane w tym rozdziale.

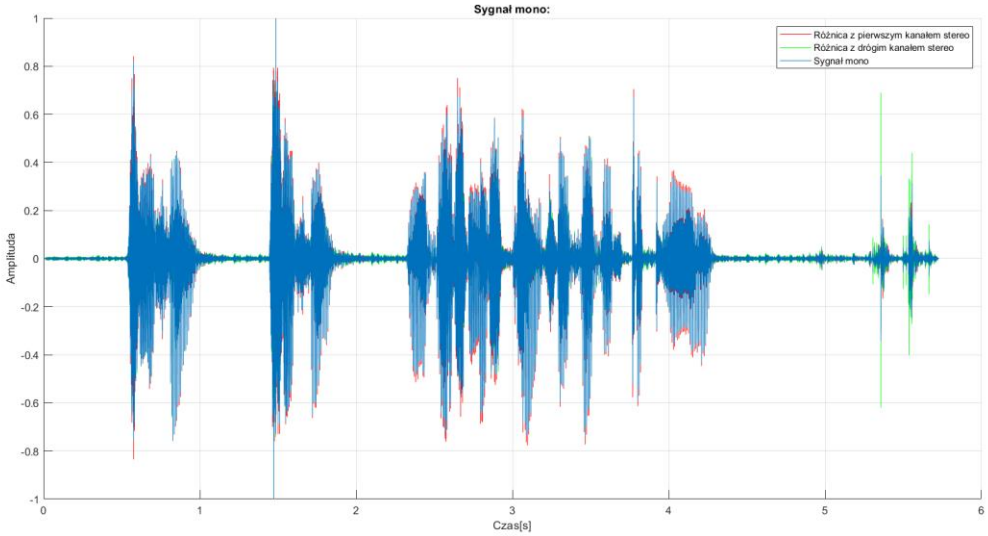
2.1. Przekształcenie stereo-mono

Nagrania stereofoniczne różnią się od sygnałów monofonicznych ilością kanałów, na których zapisany zostaje dźwięk. Sygnały audio pozyskane w technice mono posiadają tylko jeden kanał natomiast w technice stereo dźwięk zapisywany jest na dwóch lub więcej kanałach. Wielokanałowe rozwiązanie poprzez różnice dźwięku między kanałami pozwala lepiej oddać położenie dźwięku w przestrzeni. Właściwość ta nie jest jednak potrzebna podczas analizy nagrań głosu, które zostały przeprowadzone w dalszej części pracy, dlatego też w celu ułatwienia analizy oraz przyśpieszeniu działania algorytmów sygnał musi zostać ujednolicony do jednego kanału.



Rysunek 1. Przebieg sygnału audio pobranego z dwukanałowego mikrofonu.

Konwersja została wykonana poprzez obliczenie średniej wartości amplitudy sygnału każdej z próbek dla wszystkich kanałów. Obliczony w ten sposób sygnał był następnie wykorzystywany jako oryginalne nagranie.



Rysunek 2. Wynik przekształcenia wraz z zaznaczonymi różnicami.

2.2. Filtr preemfazy

Filtrem preemfazy nazywany jest filtr o skończonej odpowiedzi impulsowej (FIR) z jednym współczynnikiem. Filtr ten stosowany jest w celu wzmocnienia spektrum sygnału o około 20dB na dekadę oraz wyrównania balansu energii spektrum między wysokimi oraz niskimi częstotliwościami. Wzmocnienie to pozwala zniwelować spadek spektrum sygnału głosowego w wysokich częstotliwościach wynikający z budowy narządu głosowego oraz wynoszącego także około 20dB na dekadę [2]. Wzór transformaty Z filtra preemfazy dany jest jako,

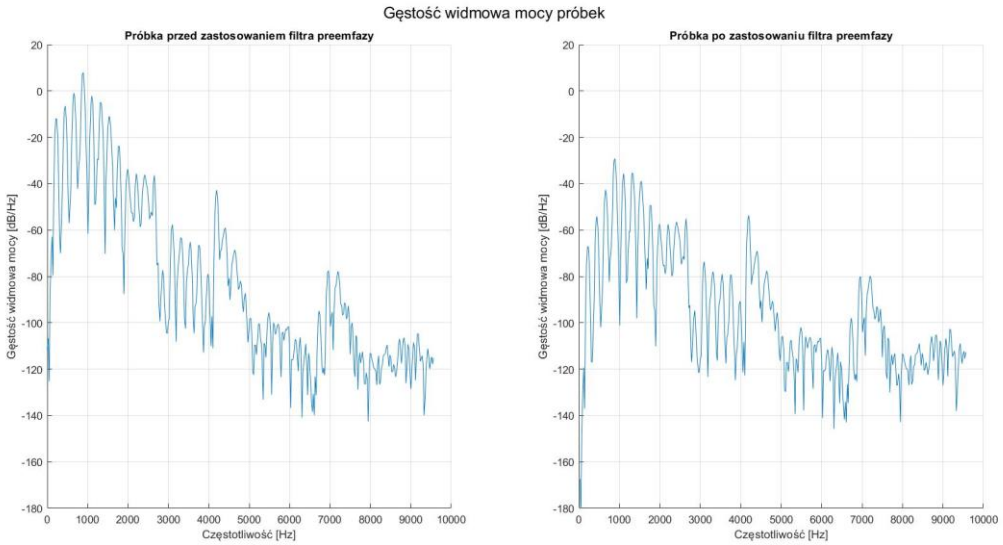
$$H_{pre}(z) = 1 + \alpha_{pre} z^{-1}, \tag{1}$$

Wiedząc, że transformata (1) jest wymierna wynik filtracji można zapisać jako,

$$Y_{pre}[n] = x[n] - \alpha x[n-1]\,, \tag{2}$$

Gdzie n to kolejna próbka sygnału x, oraz $0.9 \leq \alpha \leq 1$.

Wynik zastosowania filtru preemfazy można zaobserwować na rysunku (Rys. 3).



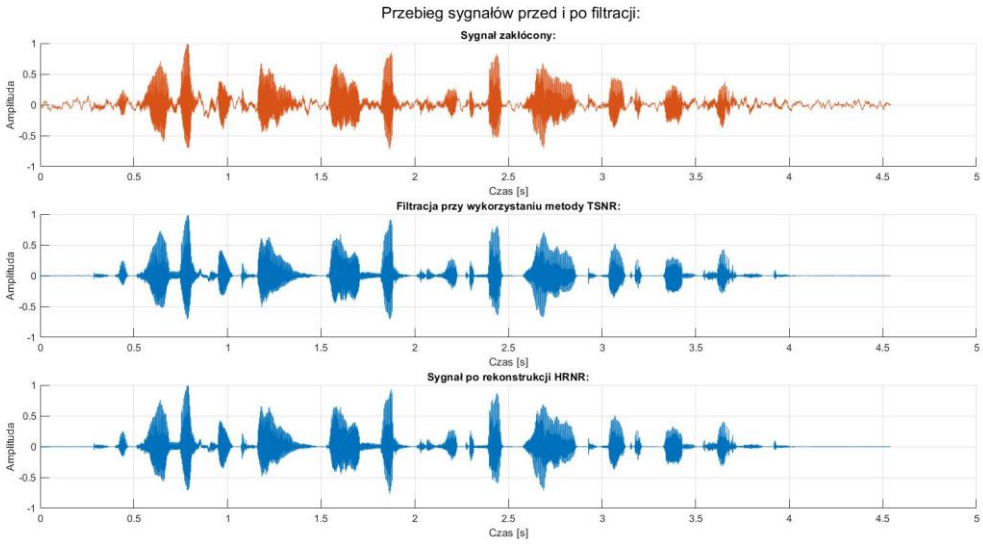
Rysunek 3. Porównanie gęstości widmowej sygnału mowy przed i po zastosowaniu filtra preemfazy.

2.3. Filtr TSNR

Jednym z większych problemów utrudniających poprawną analizę próbek głosowych jest występowanie zakłóceń sygnału. Najczęstszą przyczyną degradującą jakość nagrań głosu jest występowanie szumów tła. Szumy te dzielą się na stacjonarne i nie-stacjonarne oraz przyjmuje się, że nie są skorelowane z sygnałem mowy. Zakłada się również, że sygnał szumu jest addytywny względem sygnału mowy [3]. Algorytmy mające na celu redukcję szumów tła są ciągle rozwijane, ponieważ znajdują zastosowanie w dziedzinach takich jak komunikacja głosowa na odległość, rozpoznawanie głosu oraz innych, w których wymagana jest jak najlepsza jakość sygnału mowy [4]

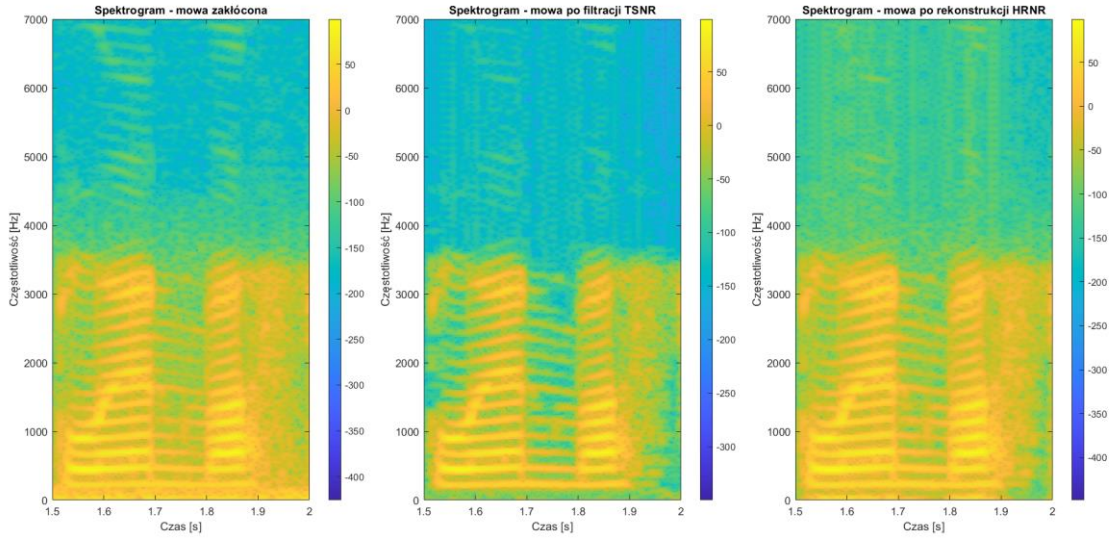
Do filtracji szumów tła próbek wykorzystanych w badaniach zastosowana została metoda Dwukrokowej Redukcji Szumu (ang. Two-Step Noise Reduction, TSNR) zaproponowana w [4]. Jest to metoda tłumienia wykorzystująca tylko jeden kanał, przez co podczas dokładnej estymacji szumu podatna jest na występowanie błędów. Wynikiem błędów tych jest przyjęcie osłabionych częstotliwości harmonicznych jako szumu powodując ich wytłumienie. Aby zrekonstruować wytłumione przez TSNR częstotliwości harmoniczne wykorzystuje się algorytm regeneracji harmonicznej (ang. Harmonic Regeneration Noise Reduction, HRNR) [3].

Aby sprawdzić sprawność działania algorytmu przeprowadzono kilka badań na zaszumionych próbkach pozyskanych z [6]. Obserwując przebieg wykresów w czasie można zauważyć różnicę przed i po filtracji.



Rysunek 4. Przebieg sygnału audio w czasie przed i po filtracji.

Różnice między filtracją przed oraz po rekonstrukcji HRNR można lepiej zaobserwować analizując spektrogramy sygnałów przedstawione na rysunku (Rys. 5).



Rysunek 5. Spektrogramy sygnałów przed oraz po filtracji.

3. Modele GMM

Gaussowski model mieszany (ang. Gaussian mixture model), w skrócie GMM jest parametryczną funkcją gęstości prawdopodobieństwa opisaną jako suma ważona gęstości komponentów Gaussowskich. W GMM zakłada się, że punkty danych generowane są poprzez mieszanie skończonej ilości dystrybucji Gaussowskich z nieznanymi parametrami. Modele te są powszechnie wykorzystywane w systemach biometrycznych wykorzystujących cechy spektralne głosu do rozpoznawania mówcy [6][7]. Modele GMM podczas eksperymentów przeprowadzonych w pracy wykorzystywane były do rozpoznawania płci oraz wieku mówcy na podstawie wektorów cech dla nagrań głosowych należących do bazy danych Mozilla Common Voice. Sposób w jaki tworzone były modele GMM został opisany rozdziale 3.1.

3.1. Tworzenie modelu GMM

Przyjmując M jako ilość komponentów Gaussowskich model GMM można zapisać jako ich sumę ważoną zgodnie ze wzorem (3), [6]

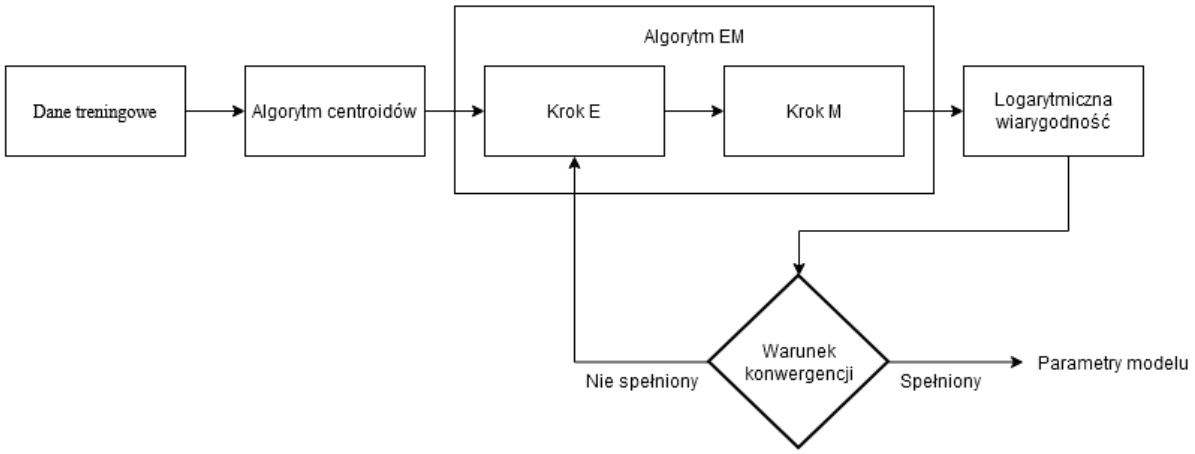
$$p(x|\lambda) = \sum_{i=1}^{M} \omega_i g(x|\mu_i, \Sigma_i), \tag{3}$$

gdzie x to d wymiarowy wektor danych wejściowych. Przyjmując, że i to indeks i-tego komponentu Gaussowskiego, ω_i to jego waga, μ_i to d wymiarowy wektor jego średnich oraz Σ_i to jego macierz kowariancji. Funkcja gęstości dla każdego d-wymiarowego komponentu gaussowskiego oznaczona we wzorze (3) jako g, dana jest wzorem,

$$g(x|\mu, \Sigma) = \frac{1}{2\pi^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right), \tag{4}$$

gdzie $|\Sigma|$ to wyznacznik macierzy kowariancji Σ. Parametry dla każdego modelu zapisujemy jako,

$$\lambda = \{\omega_i, \mu_i, \Sigma_i\} \; dla \; i = 1, \dots, M. \tag{5}$$



Rysunek 6. Schemat blokowy wyznaczania parametrów modelu GMM

Do wyznaczenia optymalnych parametrów modelu powszechnie stosowany jest dwukrokowy algorytm EM (ang. Expectation Maximisation Algorithm), [8]. Algorytm ten potrzebuje jednak wartości początkowych, aby rozpocząć wykonywanie obliczeń. W tym celu do estymacji parametrów początkowych modelu wykorzystano algorytm centroidów dostępny w środowisku Matlab jako komenda „kmeans".

Krok E algorytmu polega na obliczeniu możliwej zmiany parametrów γ modelu z wykorzystaniem obecnych parametrów. Przyjmując, N jako ilość $1 \times d$ wymiarowych wektorów wykorzystanych do budowy modelu oraz M jako ilość komponentów gaussowskich modelu, wartość γ możne zostać obliczona Korzystając ze wzoru,

$$\gamma(n,i) = \frac{\omega_i g(x_n|\mu_i,\Sigma_i)}{\sum_{j=1}^{M} \omega_j g(x_n|\mu_j,\Sigma_j)}, \tag{6}$$

gdzie $n = 1, \dots, N$ oraz $i = 1, \dots, M$. Po wykonaniu obliczeń algorytm przechodzi do kroku M.

W kroku M obliczone zostają nowe parametry modelu na podstawie wartości γ obliczonej w kroku E. Parametry te oblicza się korzystając ze wzorów,

$$\mu_i^{nowe} = \frac{1}{N_i} \sum_{n=1}^{N} \gamma(n,i) x_n, \tag{7}$$

$$\Sigma_i^{new} = \frac{1}{N_i} \sum_{n=1}^{N} \gamma(n,i)(x_n - \mu_i^{new})(x_n - \mu_i^{new})^T, \tag{8}$$

$$\omega_i^{new} = \frac{N_i}{N}, \tag{9}$$

gdzie,

$$N_i = \sum_{n=1}^{N} \gamma(n,i). \tag{10}$$

Po wyznaczeniu nowych parametrów modelu jego jakość zostaje oszacowana z wykorzystaniem logarytmicznej wiarygodności danej wzorem,

$$\ln p(X|\mu, \Sigma, \omega) = \sum_{n=1}^{N} \ln(\sum_{i=1}^{M} \omega_i g(x_n|\mu_i, \Sigma_i), \tag{11}$$

Ostatecznie sprawdzany jest warunek konwergencji. Jeśli nie został on spełniony algorytm wraz z nowymi parametrami wraca do kroku E algorytmu EM. Algorytm kończy się w momencie spełnienia warunku.

4. Ton podstawowy

Tonem podstawowym nazywa się falę harmoniczną o najmniejszej częstotliwości w szeregu harmonicznym, którą zwykle oznacza się jako F0. Za wysokość F0 odpowiada wielkość oraz grubość fałdu głosowego, który w okresie dojrzewania u mężczyzn rozrasta się o około 60% bardziej niż u kobiet. Masywność fałdów głosowych przekłada się bezpośrednio na zmniejszenie częstotliwości ich wibracji co powoduje obniżenie tonu F0 u mężczyzn [9]. Wielkość fałdów głosowych nie jest jednak jedynym czynnikiem wpływającym na wysokość tonu podstawowego. Istnieją również badania pokazujące, że różnice w F0 mogą zależeć także od pochodzenia [10] lub nawyków, na przykład palenia [11].

Inną cechą mającą wpływ na wysokość F0 jest wiek badanej osoby. Znane jest występowanie trendu łączącego wiek z wysokością tonu podstawowego, polegającego na jego zmianie wraz ze starzeniem się organizmu [12]. Fałdy głosowe mężczyzn po osiągnięciu wieku średniego zaczynają tracić swoją objętość stając się cieńsze oraz sztywniejsze co zwykle prowadzi do obserwacji niewielkiego wzrostu F0 u mężczyzn wraz z wiekiem. Kobiety przechodzą bardziej drastyczne zmiany związane ze zmianami hormonalnymi w okresie menopauzy. Zmiany te obejmują zwiększenie objętości oraz sztywności fałdów głosowych oraz ich wysuszenie, co prowadzi do znacznego zmniejszenia wysokości tonu podstawowego [13].

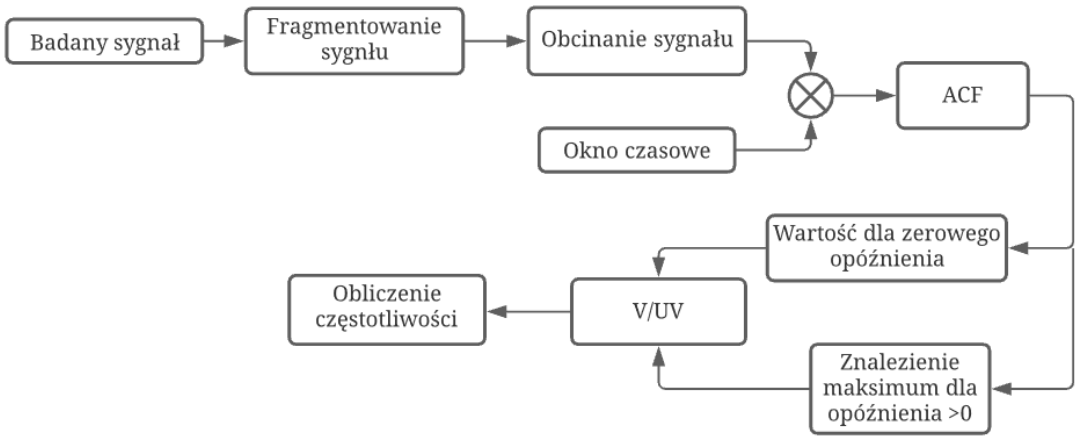
4.1. Obliczanie tonu podstawowego

Istnieje wiele algorytmów pozwalających na wyznaczenie tonu podstawowego mówcy na podstawie próbek pozyskanych z nagrań jego głosu [14]. W trakcie eksperymentów do analizy tonu podstawowego wykorzystane zostały dwa sposoby różniące się od siebie dziedziną, w której przeprowadzane było badanie f_0. Pierwszy ze sposobów pozwalał na przeprowadzenie analizy w dziedzinie czasu przy wykorzystaniu metody autokorelacji, drugi natomiast przeprowadzany był w dziedzinie częstotliwości przy wykorzystaniu cepstrum mocy sygnału.

4.1.1. Obliczanie tonu podstawowego z wykorzystaniem autokorelacji

Autokorelacja jest powszechnie wykorzystywaną metodą na estymację tonu podstawowego w dziedzinie czasu. Założeniem metody jest istnienie tonu podstawowego jako maksimum o najwyższej amplitudzie dla sygnału wejściowego przemnożonego przez ten sam sygnał opóźniony o czasowy odpowiednik częstotliwości tonu podstawowego. Algorytm AFC wraz z dodatkowymi modyfikacjami został stworzony na podstawie pracy [15].

Aby przyśpieszyć oraz poprawić działanie algorytmu stosowana jest metoda krótko-czasowej autokorelacji (short-time ACF) polegającą na wyznaczaniu F0 dla każdego fragmentu podzielonego wcześniej sygnału. Stosując fragmentowanie sygnału badane są tylko częstotliwości w zakresie od częstotliwości próbkowania sygnału do częstotliwości próbkowania sygnału podzielonej przez długość okna czasowego. Taki zabieg pozwala nam na pominięcie badania niskich częstotliwości, które nie mogą być tonem podstawowym natomiast mogą pogorszyć otrzymane wyniki oraz wydłużyć czas obliczeń.



Rysunek 7. Schemat blokowy wyznaczania F0 z wykorzystaniem ACF

Badany sygnał zostaje podzielony na fragmenty o długości z zakresu 25-40 ms, następnie wykorzystując technikę obcinania centrum (ang. Center Clipping) daną wzorem,

$$y(n) = clc[x(n)] = \begin{cases} x(n) - C_L & dla \quad x(n) \geq C_L \\ 0 & dla \quad |x(n)| < C_L, \\ x(n) + C_L & dla \quad x(n) \leq -C_L \end{cases} \tag{12}$$

dla sygnału $x(n)$ obliczona zostaje jego przycięta wersja $y(n)$. Zabieg ten pozwala pominąć wartości sygnału dla których amplituda jest mniejsza niż wartość C_L.

Próg C_L ustawiony został jako 60% wartości maksymalnej amplitudy sygnału w badanym fragmencie. Następnie sygnał $y(n)$ zostaje pomnożony przez okno czasowe o odpowiedniej długości po czym obliczane zostają wartości funkcji autokorelacji $R[k]$ danej wzorem,

$$R[k] = \sum_{m=0}^{N-k-1} s[m]s[m+k], \tag{13}$$

Gdzie N to ilość próbek w oknie czasowym, $s[m]$ to wartość sygnału dla próbki o indeksie m a k to ilość próbek o którą opóźniony jest sygnał s.

Badając próbkę w zakresie $k = [0, N-1]$ można zbadać występowanie interesujących częstotliwości. Dla $k = 0$ otrzymamy wartość energii autokorelacji badanej ramki sygnału, natomiast badając k, dla którego funkcja przyjmuje maksimum w granicach $k = [k_{lb}, k_{ub}]$, otrzymamy opóźnienie k_{f0} odpowiadające wartości tonu podstawowego. Wartości k_{lb} oraz k_{ub} wyznacza się z wykorzystaniem wzorów,

$$k_{lb} = \frac{f_s}{f_{0l}} \; oraz \; k_{ub} = \frac{f_s}{f_{0u}}, \tag{14}$$

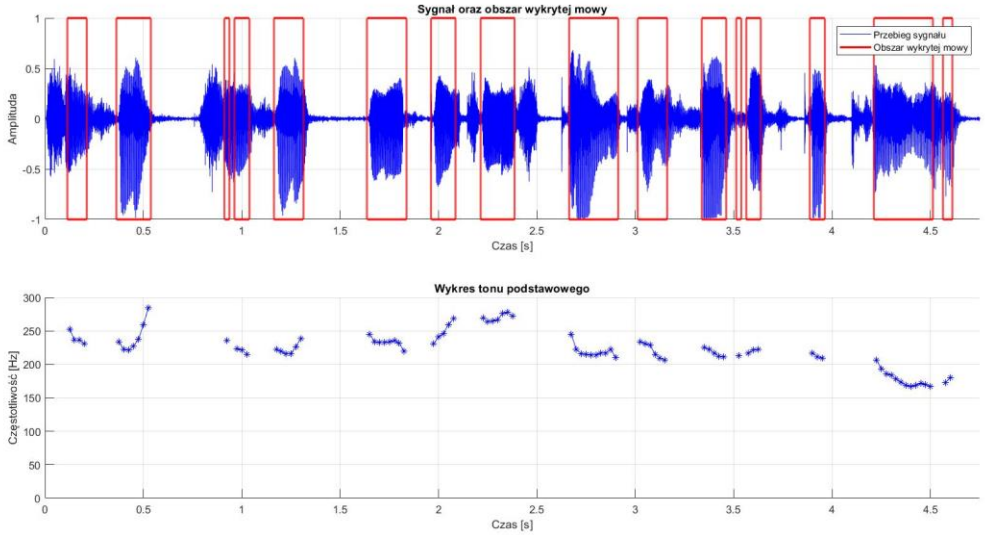
Gdzie f_{0l} oraz f_{0u} to odpowiednio minimalna oraz maksymalna częstotliwość, dla której możliwe jest wystąpienie tonu podstawowego a f_s to częstotliwość próbkowania sygnału. Przeliczenie znalezionego opóźnienia k_{f0} w dziedzinę częstotliwości wykonuje się korzystając ze wzoru,

$$f_0 = \frac{f_s}{k_{f0}}, \tag{15}$$

Przed wyznaczeniem tonu podstawowego podejmowana jest także decyzja, czy w badanym fragmencie występuje mowa (Voiced/Unvoiced Decision). Przy podjęciu decyzji sprawdza się warunek,

$$R\left[k_{f0}\right] \geq 0.55R[0], \tag{16}$$

Jeżeli warunek ten nie został spełniony, badana ramka jest odrzucana. Po sprawdzeniu wszystkich ramek obliczony zostaje średni ton podstawowy. Takie rozwiązanie pozwala na zmniejszenie błędu wyznaczonego tonu podstawowego poprzez badanie tylko „pewnych" ramek.

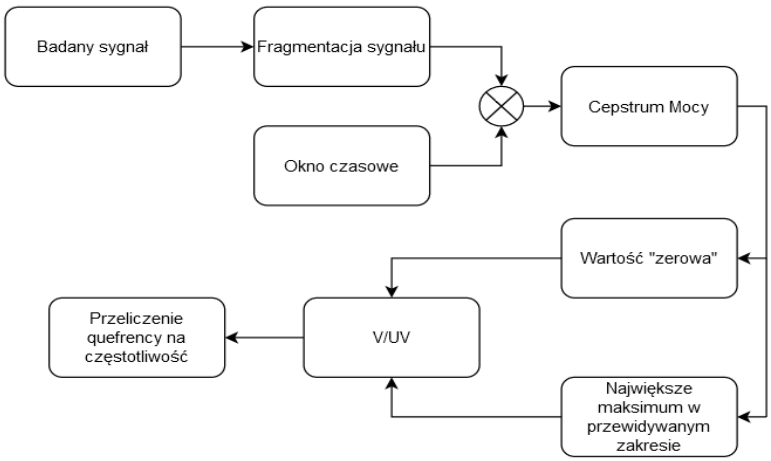


Rysunek 8. Wykres obrazujący działanie algorytmu ACF

4.1.2. Obliczanie tonu podstawowego z wykorzystaniem Cepstrum

Cepstrum sygnału to wynik odwrotnej transformaty Fouriera z logarytmu spektrum badanego sygnału. Operacja ta po raz pierwszy została zdefiniowana przez B. Bogerta, M. Healy'a oraz J. Tukey'a w pracy [16] jako metoda wykrywania echa w sygnałach sejsmicznych. Możliwość badania okresowości fal o określonych częstotliwościach w sygnale sprawiła, że analiza głosu stała się jedną z pierwszych dziedzin, w której cepstrum zyskało popularność [17]. Do badania tonu podstawowego wykorzystuje się powszechnie cepstrum mocy sygnału dane wzorem,

$$C_p = |\mathcal{F}^{-1}\{\log(|\mathcal{F}\{f(t)\}|^2)\}|^2 \,. \tag{17}$$



Rysunek 9. Schemat blokowy algorytmu wykrywania F0 z wykorzystaniem cepstrum mocy sygnału

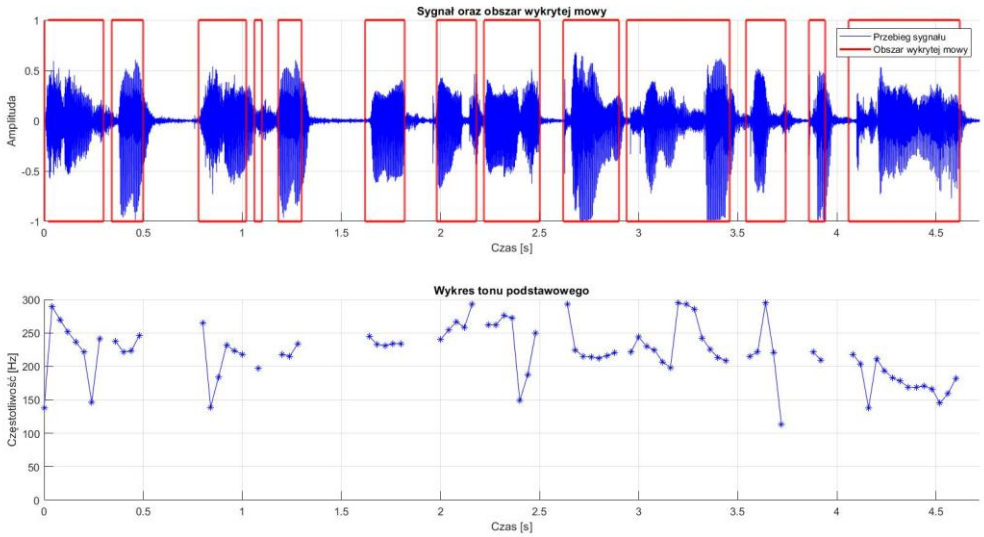
Podobnie jak w przypadku metody ACF badany sygnał zostaje podzielony na fragmenty pozwalając w ten sposób na pominięcie obliczeń dla bardzo niskich częstotliwości. Każdy z fragmentów zostaje przemnożony przez odpowiednie okno czasowe. Następnie obliczane zostaje cepstrum mocy (17) z którego wyciągane są wartości dla quefrency równej zero oraz wartość największego maksimum w przedziale quefrnecy odpowiadającym możliwym wystąpieniom tonu podstawowego. Przyjmując, że przewidywane wystąpienie f_0 znajduje się w przedziale $[f_{0_{lb}},\ f_{0_{ub}}]$, odpowiednie wartości quefrency wyznacza się korzystając ze wzoru,

$$quefrency = \frac{częstotliwość\ nagrywania}{szukana\ częstotliwość}. \tag{18}$$

Wiedząc, że wartość C_p dla $quefrency = 0$ przenosi informację o całkowitej energii autokorelacji sygnału w badanym fragmencie [18], podobnie jak w przypadku metody ACF po znalezieniu odpowiednich wartości podejmowana jest decyzja czy uzyskana wartość tonu podstawowego należy do fragmentu zawierającego mowę, czy nie (V/UV).

Jeśli amplituda znalezionego maksimum podzielona przez wartość C_p dla $quefrency = 0$, jest mniejsza niż próg P_{vu} to próbka jest oznaczana jako niezawierająca mowy oraz obliczony z niej ton podstawowy nie jest zaliczany do wyniku. Krok ten jednak może zostać zastosowany tylko dla bazy nagrań zbieranych w jednakowych warunkach, przy wykorzystaniu tego samego urządzenia nagrywającego. W innym wypadku ręczne ustalenie progu P_{vu} jest nieosiągalne.

Następnie otrzymana wartość quefrency, dla której znaleziono maksimum zostaje przeliczona na częstotliwość z wykorzystaniem wzoru (18). Po zbadaniu wszystkich ramek ton podstawowy zostaje obliczony jako średnia wszystkich wyników.



Rysunek 10. Przykład wykrywania F0 z wykorzystaniem cepstrum

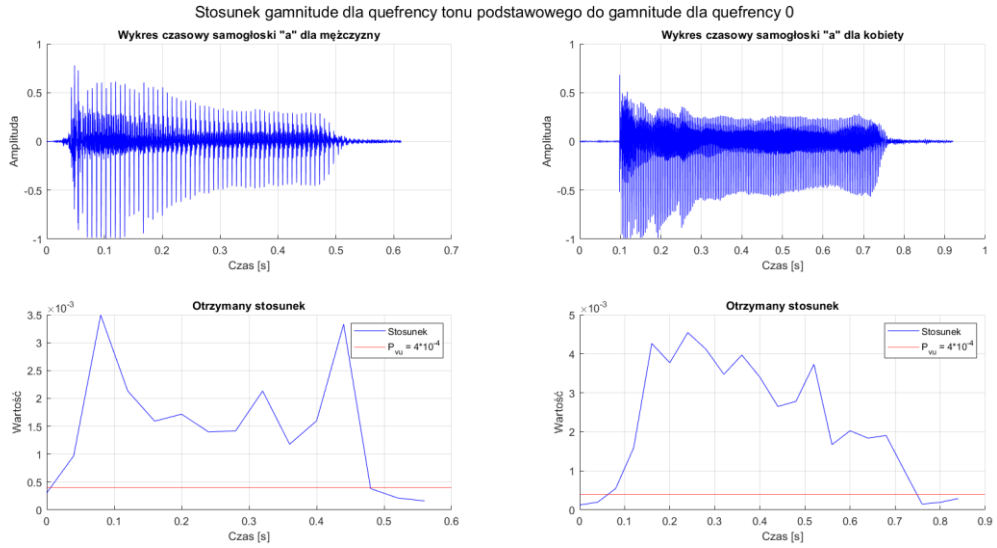
4.2. Analiza tonu podstawowego

Wykorzystując opisane wcześniej metody, wyznaczono ton podstawowy dla bazy samogłosek wypowiadanych przez dwie kobiety oraz dwóch mężczyzn. Każde nagranie została przefiltrowane algorytmem TSNR. Sygnał został podzielony na ramki długości 25 ms. Przyjęto, że poszukiwany ton podstawowy znajduje się w zakresie 80-350 Hz. Jako pierwszy sprawdzono algorytm autokorelacji.

Tabela 2. Ton podstawowy z podziałem na osoby oraz samogłoski dla metody ACF.

	Autokorelacja			
	Mężczyzna 1, 20 lat	Mężczyzna 2, 48 lat	Kobieta 1, 20 lat	Kobieta 2, 48 lat
„a”	122	126	258	230
„e”	123	128	258	229
„i”	131	138	259	229
„o”	125	128	260	229
„u”	130	134	261	235
„y”	129	133	258	234

Następnie dla tych samych nagrań wyznaczono ton podstawowy z wykorzystaniem cepstrum mocy sygnału. Ręcznie ustalono wartość progu $P_{vu} = 4 \cdot 10^{-4}$ na podstawie obserwacji wykresu (Rys. 11).



Rysunek 11. Przebiegi samogłoski "a" dla kobiety oraz mężczyzny z wyznaczonym stosunkiem oraz zaznaczonym wybranym progiem.

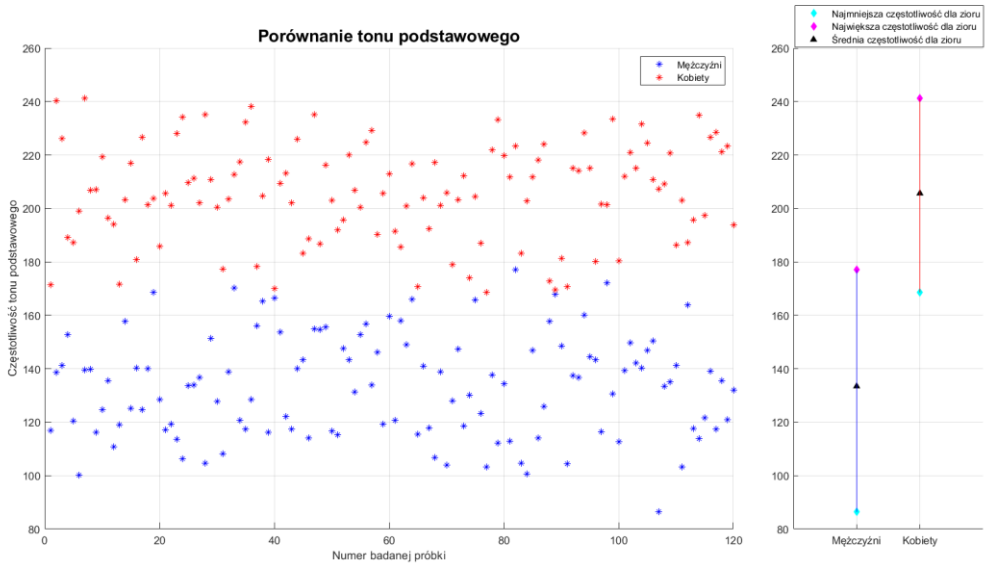
Otrzymane wyniki zostały zebrane w tabeli (Tab. 3)

Tabela 3. Ton podstawowy z podziałem na osoby oraz samogłoski dla metody CEPSTRUM.

	CEPSTRUM			
	Mężczyzna 1, 20 lat	Mężczyzna 2, 48 lat	Kobieta 1, 20 lat	Kobieta 2, 48 lat
„a”	122	125	255	234
„e”	123	131	246	220
„i”	130	138	228	217
„o”	125	127	245	226
„u”	130	133	242	232
„y”	128	127	244	225

Dla obu metod otrzymane różnice w F0 między kobietami oraz mężczyznami są na tyle wysokie, że określenie płci osoby na podstawie wysokości tonu podstawowego nie stanowi większego problemu. Analizując różnice dla wieku można zauważyć widoczny spadek F0 między 20 oraz 48 letnią kobietą. Dla mężczyzn widoczny jest nieznaczny wzrost F0.

Podczas testów działania algorytmów dla nagrań zawierających całe zdania zaobserwowano, że metoda CEPSTRUM radzi sobie gorzej niż metoda oparta na autokorelacji. Mając to na uwadze do obliczenia tonu podstawowego dla bazy głosów Mozilla Common Voice wykorzystano tylko algorytm ACF. Do analizy wykorzystano po sto losowo wybranych nagrań dla obu płci. Ponieważ przy takiej ilości nagrań ocena ich jakości nie była możliwa ręcznie, każde nagranie przefiltrowano algorytmem TSNR przyjmując pierwsze 100 ms jako szum. Otrzymane wyniki przedstawione zostały na rysunku (Rys. 12).



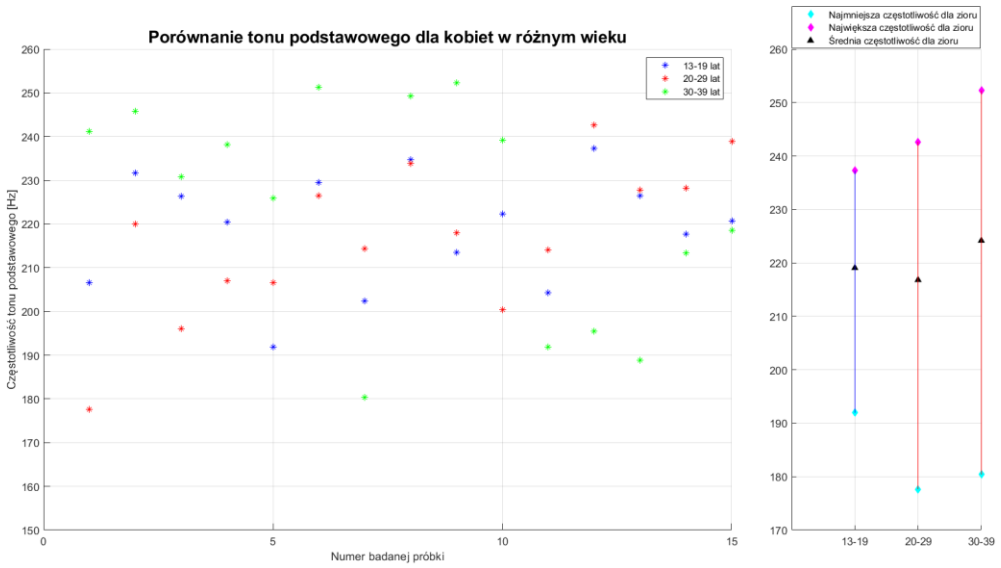
Rysunek 12. Wyniki analizy tonu podstawowego dla nagrań z bazy Mozilla Common Voice

Podczas testów okazało się, że najlepsze wyniki analizy uzyskano dla okna wynoszącego 30 ms, oraz dla ramek nie nachodzących na siebie. Zauważono także, że decyzja V/UV

algorytmu autokorelacji dla niektórych nagrań głosów męskich okazuje się zbyt rygorystyczna. Stosunek wartości autokorelacji tonu podstawowego $R[k_{f0}]$ męskiego głosu do energii autokorelacji badanego fragmentu $R[0]$ okazuje się czasem mniejszy niż 0,55. Aby poprawnie zbadać ton podstawowy w takich przypadkach obniżany jest próg V/UV aż ton podstawowy zostanie znaleziony. Dla nagrań żeńskich nie zaobserwowano problemów z progiem decyzji V/UV.

Wyniki otrzymane z badania pokazują wyraźne różnice w F0 między płciami, możliwe do zaobserwowania korzystając z fragmentów wypowiedzi nagranych w nieznanych warunkach, wykonanych z użyciem nieznanego urządzenia. Ton podstawowy mężczyzn zawierał się w przedziale [80-180Hz] ze średnią około 135Hz natomiast kobiet [170-240Hz] ze średnią ~208Hz. Wykorzystując modele GMM udało się uzyskać 98% dokładności rozpoznania płci dla zbioru treningowego złożonego z 90 wartości tonu podstawowego dla każdej płci. Liczebność zbioru testowego wynosiła 30.

Ponieważ ilość nagrań głosów kobiet powyżej 39 roku życia była niewystarczająca, sprawdzono czy różnice w F0 pozwalają zaobserwować różnice między trzema najliczniejszymi grupami wiekowymi dla kobiet, czyli 13-19, 20-29 oraz 30-39 lat. Otrzymane wyniki nie wykazały widocznej zmiany tonu podstawowego między osobami z tych grup, co wskazuje, że rozpoznawanie wieku mówcy w takich przedziałach może być nieosiągalne wykorzystując wyłącznie ton podstawowy.



Rysunek 13. Porównanie tonu podstawowego dla kobiet w różnym wieku

W przypadku mężczyzn, dla których ilość nagrań była większa, porównano wszystkie dostępne przedziały wiekowe. Dwa ostanie przedziały połączono razem tworząc przedział 40-59 lat.

Wyznaczony średni ton podstawowy mężczyzn z przedziałów wiekowych 13-19, oraz 20-29 jest niższy niż dla pozostałych przedziałów grup starszych mężczyzn. Nie można jednak wyznaczyć odpowiednich granic między F0 oraz odpowiednimi grupami wiekowymi (Rys. 14).

Rysunek 14. Porównanie tonu podstawowego dla mężczyzn w różnym wieku

4.3. Wnioski z analizy tonu podstawowego

Oba algorytmy wykrywania tonu podstawowego sprawdzają się poprawnie w przypadku badania tonu podstawowego dla nagrań pobranych w znanych warunkach. Wyniki dla badanego zbioru samogłosek były podobne dla każdej badanej osoby. Badając fragmenty wypowiedzi pochodzące z bazy Mozilla Common Voice można było zaobserwować znaczną różnicę w jakości wyników między algorytmami. Metoda cepstrum sprawdziła się w tych warunkach znacznie gorzej od metody autokorelacji, co spowodowane było poprzez brak możliwości poprawnego ustalenia progu P_{vu} dla nagrań o nieznanych parametrach. Mając ten fakt na uwadze można stwierdzić, że algorytm obliczania tonu podstawowego z wykorzystaniem ACF jest lepszym wyborem podczas badania nagrań nieznanego pochodzenia.

Ton podstawowy, jak pokazują wyniki, pozwala na ocenę płci mówcy na podstawie jego nagrania. Wykorzystując ton podstawowy do rozpoznawania płci udało się uzyskać 98% skuteczności poprawnego rozpoznania dla modeli GMM. Rozpoznanie wieku na podstawie tonu podstawowego nie jest jednak możliwe dla osób z przedstawionych w rozdziale 4.2 przedziałów wiekowych. Aby jednak całkowicie odrzucić ten pomysł potrzebne by było przeprowadzenie badań na większej ilości przedziałów wiekowych.

5. Formanty

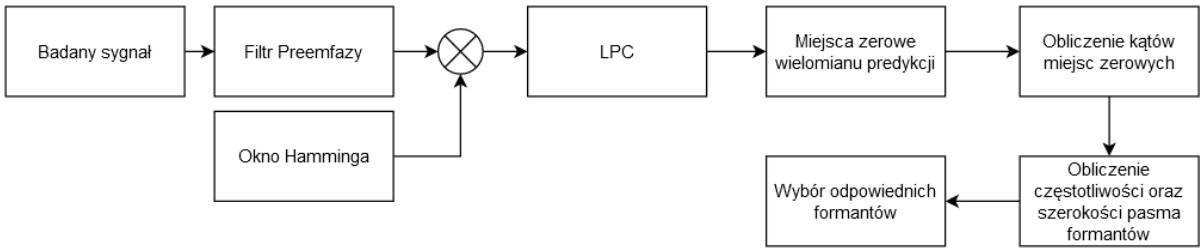
Formanty są to szerokie spektralne maksima powstające na skutek rezonansu akustycznego traktu głosowego. W granicach pasma częstotliwości formantów wszystkie tony składowe głosu ulegają szczególnemu wzmocnieniu. Na wysokość formantów wpływa długość traktu głosowego. Mężczyźni posiadają o około 15% dłuższy trakt głosowy co przekłada się na niższe częstotliwości formantów. Częstotliwości formantów powiązane są z barwą głosu, która jak pokazano w pracy [19] pozwala na rozpoznanie płci mówcy bez potrzeby oceny tonu podstawowego. [20], [21]

Do wyznaczania częstotliwości formantów najczęściej wykorzystywane jest liniowe kodowanie predykcyjne (LPC). Podczas analizy nagrań zastosowana została właśnie ta metoda.

5.1. Wykrywanie formantów z wykorzystaniem LPC

Liniowe kodowanie predykcyjne jest potężnym narzędziem pozwalającym na analizę sygnałów audio oraz ich przetwarzanie. Metoda ta zdobyła popularność dzięki możliwości wyjątkowo efektywnej oraz precyzyjnej estymacji parametrów mowy. LPC obliczane jest poprzez wykorzystanie kombinacji liniowych poprzednio zbadanych próbek sygnału do predykcji obecnej próbki. Jeżeli predykcja okaże się skuteczna to współczynniki poprzednich funkcji liniowych mogą zostać wykorzystane do reprezentacji badanego sygnału [22]. W trakcie badań do obliczenia LPC sygnału wykorzystano wbudowaną funkcję „lpc" środowiska Matlab.

Aby obliczyć formanty sygnału potrzebne jest obliczenie zespolonych miejsc zerowych r_n wielomianu predykcji otrzymanego z funkcji LPC [22]. Następnie wybierane są tylko r_n z takim samym znakiem przy współczynniku zespolonym, tak aby ograniczyć zakres szukanych częstotliwości od 0 do $\frac{f_s}{2}$, gdzie f_s to częstotliwości próbkowania sygnału.



Rysunek 15. Diagram wykrywania formantów z wykorzystaniem LPC

Dla wybranych zespolonych miejsc zerowych obliczony zostaję kąt (19), na podstawie którego oblicza się częstotliwości formantów (20). W tym samym kroku obliczane są także szerokości pasam (21) dla danych formantów.

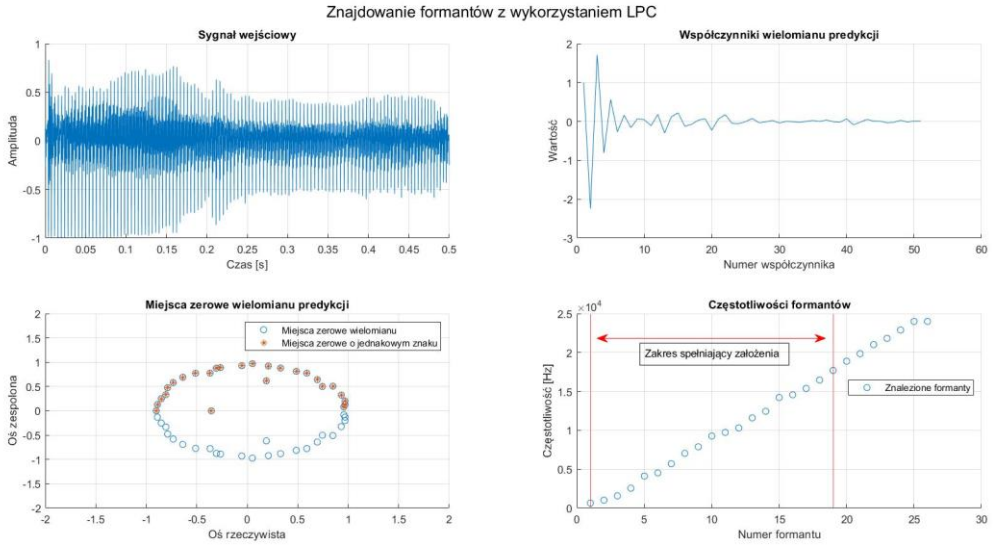
$$\theta_n = arctan2(Im(r_n), Re(r_n)), \tag{19}$$

$$F_n = \frac{fs}{2\pi}\theta_n, \tag{20}$$

$$B_n = -\frac{f_s}{2\pi}\ln(r_n), \tag{21}$$

Gdzie $arctan2$ to dwu argumentowa funkcja arcus tangens, oraz n to numer znalezionego miejsca zerowego.

Obliczone formanty wybiera się na podstawie ich częstotliwości oraz szerokości pasma. Zwykle przyjmuje się $F_n > 90\ Hz$ oraz $B_n < 400\ Hz$.



Rysunek 16. Znajdowanie formantów z wykorzystaniem LPC

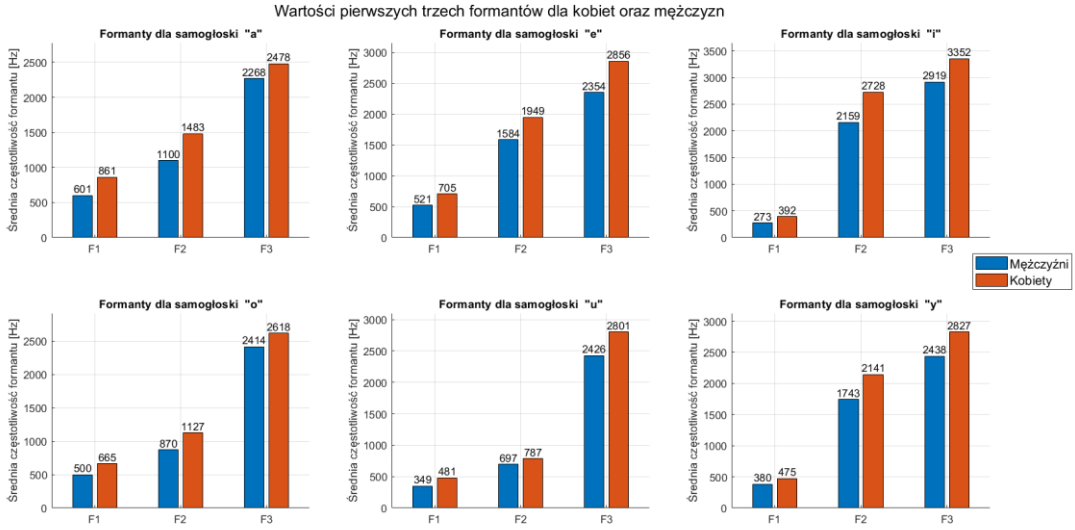
5.2. Analiza formantów

Wykorzystując przedstawiony wcześniej algorytm wyznaczono częstotliwości formantów dla bazy nagrań samogłosek. Wyniki zostały przedstawione w tabeli (Tab. 4).

Tabela 4. Wysokości formantów poszczególnych samogłosek dla dwóch kobiet oraz dwóch mężczyzn

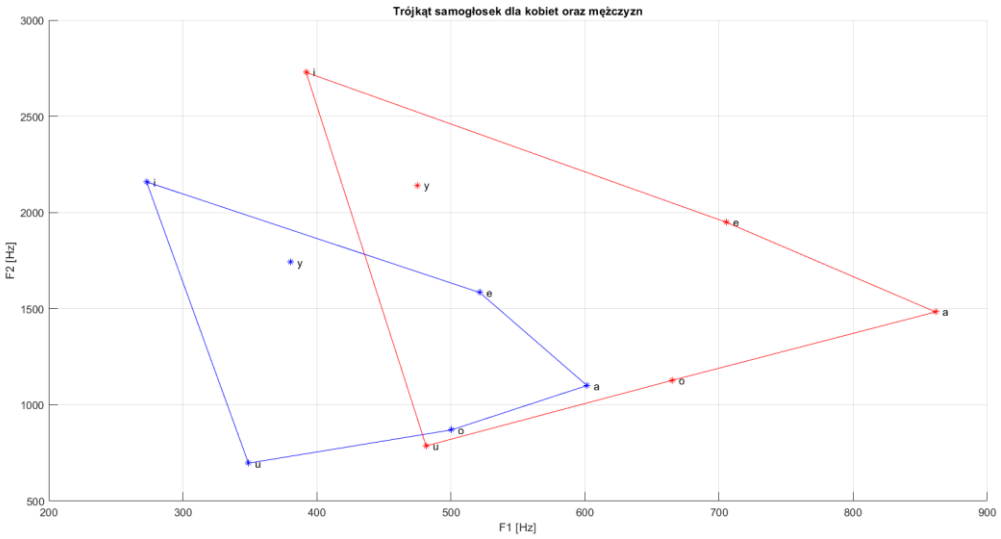
	Mężczyzna 1, 20 lat			**Mężczyzna 2, 48 lat**			**Kobieta 1, 20 lat**			**Kobieta 2, 48 lat**		
	$f_1\ [Hz]$	$f_2\ [Hz]$	$f_3\ [Hz]$	$f_1\ [Hz]$	$f_2\ [Hz]$	$f_3\ [Hz]$	$f_1\ [Hz]$	$f_2\ [Hz]$	$f_3\ [Hz]$	$f_1\ [Hz]$	$f_2\ [Hz]$	$f_3\ [Hz]$
„a”	554	1082	2298	649	1119	2238	833	1489	2443	890	1477	2514
„e”	506	1605	2386	537	1563	2323	751	1864	2800	660	2034	2913
„i”	268	2202	2970	278	2116	2869	384	2665	3256	400	2792	3449
„o”	493	939	2378	508	801	2450	631	1154	2410	699	1101	2826
„u”	349	769	2490	349	625	2363	505	794	2697	458	780	2905
„y”	371	1682	2482	390	1804	2395	478	2095	2863	472	2188	2792

Analizując wyniki przedstawione w tabeli (Tab. 4) można stwierdzić, że wysokości formantów są zależne od płci osoby mówiącej. Pierwsze trzy formanty dla kobiet są większe niż dla mężczyzn dla każdej samogłoski.



Rysunek 17. Średnie wartości trzech formantów dla kobiet oraz mężczyzn

Tworząc trójkąt samogłosek na podstawie pierwszych dwóch formantów można lepiej zaobserwować różnice między płciami.



Rysunek 18. Trójkąt samogłosek dla kobiet oraz mężczyzn

Na rysunku (Rys. 18) można zauważyć, że trójkąt samogłosek dla kobiet jest bardziej rozciągnięty oraz przesunięty w stronę wyższych częstotliwości.

Następnie dla zbiorów nagrań męskich oraz żeńskich z bazy danych Mozilla Common Voice przeprowadzono badanie polegające na ocenie możliwości rozpoznawania płci z wykorzystaniem wektora formantów występujących w mowie.

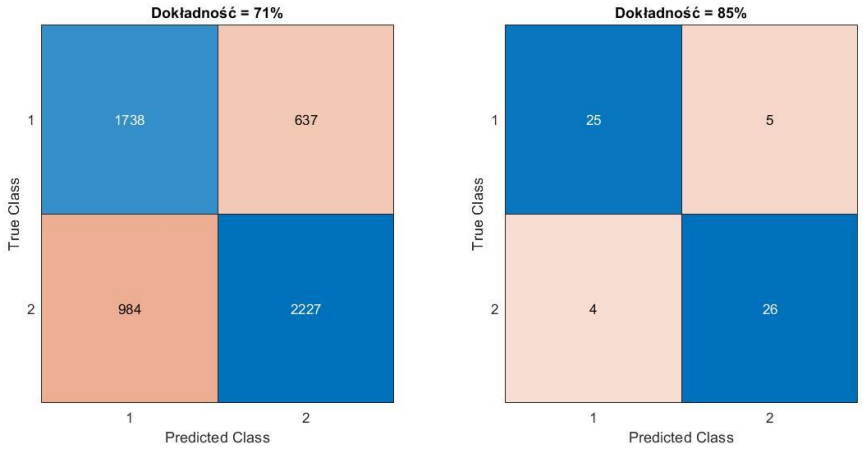
W celu badania losowo wybrano po 100 nagrań dla grupy mężczyzn oraz kobiet. Dla każdego z nagrań wyznaczono ton podstawowy z wykorzystaniem funkcji autokorelacji, tworząc przy tym wektor ramek nagrania spełniających założenia V/UV algorytmu. Następnie dla wszystkich ramek obliczono formanty z wykorzystaniem opisanego w punkcie 6.1 algorytmu. Oba zbiory następnie podzielono na 70 nagrań treningowych oraz 30 testowych.

Następnie utworzono dwa modele GMM osobno dla kobiet oraz mężczyzn, które wytrenowane zostały z wykorzystaniem formantów każdej ramki przypisanej do nagrania treningowego. Działanie wytrenowanego modelu sprawdzono na ramkach nagrań oznaczonych jako testowe.

Analiza została przeprowadzona dla liczby formantów należącej od 1 do 16. Sprawdzono, jak wygląda skuteczność modelu przy wykrywaniu płci dla różnych ilości komponentów gaussowskich wykorzystanych w budowie modelu.

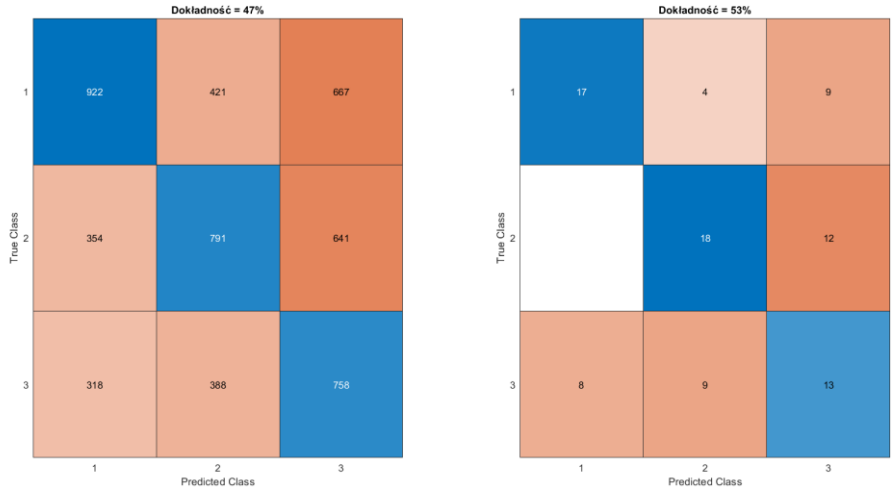
Z otrzymanych wyników zaobserwowano, że najlepsze wyniki klasyfikacji ramek otrzymano dla liczby badanych formantów równej 9 oraz liczby komponentów gaussowskich modelu równej 4.

Następnie każde nagranie było klasyfikowane jako męskie lub żeńskie w zależności od dominujących w nim ramek danej płci, co pozwoliło uzyskać dokładność 68-71% przekładającą się na 85% skuteczności rozpoznawania płci badanej osoby.



Rysunek 19. Macierze pomyłek rozpoznawania płci dla ramek (lewy) oraz nagrań (prawy). 1-mężczyźni, 2-kobiety

Eksperyment ten został powtórzony dla nagrań podzielonych na kategorie wiekowe. Do badania zostały wybrane trzy najliczniejsze grupy, czyli osoby w 13-19 lat, 20-29 lat oraz 30-39 lat. Liczebność każdej grupy wynosiła 130 nagrań z podziałem na 100 nagrań treningowych oraz 30 nagrań testowych. Największą dokładność, jaką udało się uzyskać przy rozpoznawaniu wieku z wykorzystaniem modeli GMM uczonych wektorem cech złożonym z formantów, wynosiła ~47% skuteczności przy rozpoznawaniu ramki co przekładało się na 48-53% skuteczności w rozpoznawaniu osoby.



Rysunek 20. Macierze pomyłek rozpoznawania wieku dla ramek (lewy) oraz nagrań (prawy). 1=13-19 lat, 2=20-29 lat, 3=30-39 lat

5.3. Analiza formantów – wnioski

Wykorzystując modele GMM udało się poprawnie rozpoznać płeć w 85% przypadków. Uzyskane wyniki pozwalają jednak stwierdzić, że analiza formantów może być wykorzystywana z powodzeniem do rozpoznawania płci mówcy z pominięciem obliczania tonu podstawowego lub jako czynnik decyzyjny, w przypadku braku możliwości pewnej klasyfikacji płci z wykorzystaniem tonu podstawowego.

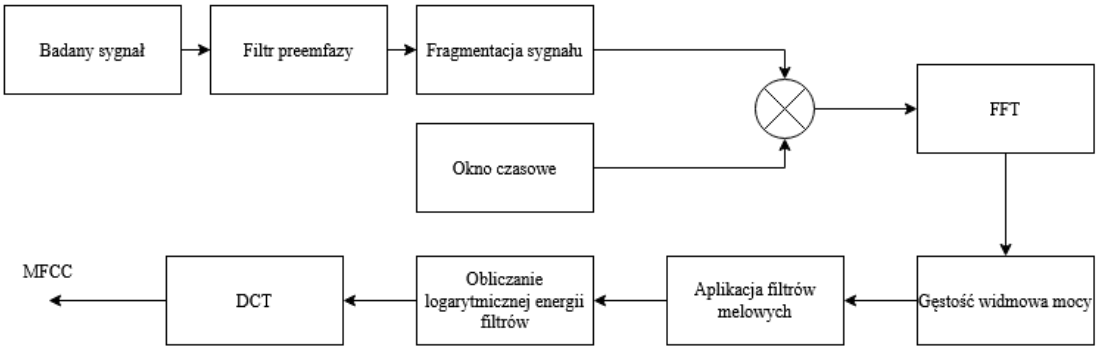
Wektor formantów nie odniósł jednak sukcesu jako wektor charakterystyk wykorzystanych w celu klasyfikacji wieku mówcy. Połączony z modelami GMM pozwolił na osiągnięcie dokładności wysokości 53%. Wynik ten pozwala jednak zaproponować istnienie zależności między wysokościami formantów dla osób różniących się wiekiem w odstępach 10 lat. Aby poprawnie rozważyć to stwierdzenie potrzebne byłoby przeprowadzenie badań na nagraniach pobranych w kontrolowanych warunkach. Dla analizy nagrań pobranych w nieznanych warunkach wymagane będzie zbadanie innej metody wyznaczania wektora charakterystyk.

6. Współczynniki MFC

MFCC, czyli Melowe Współczynniki Cepstralne zostały po raz pierwszy przedstawione przez S. Davisa oraz P. Mermelsteina w pracy [23] wydanej w roku 1980. Od momentu wydania ciągle są najczęściej wykorzystywaną metodą do ekstrakcji wektora cech z próbki mowy. Współczynniki MFC powszechnie wykorzystuje się jako wektor charakterystyk w celu klasyfikacji danych z wykorzystaniem uczenia maszynowego.

6.1. Wyznaczanie współczynników MFC

Melowe współczynniki cepstralne uzyskuje się poprzez obliczenie odwrotnej transformacji kosinusowej z logarytmu sumy energii dla każdego filtru melowego pomnożonego przez spektrum energii sygnału. Algorytm wyznaczania współczynników MFC został napisany w oparciu o [24].



Rysunek 21. Schemat blokowy wyznaczania MFCC

Pierwszym krokiem jest stworzenie banku filtrów melowych, które zostaną pomnożone przez spektrum gęstości widmowej mocy sygnału głosu. Na początku wybrana zostaje ilość współczynników MFC N, które należy wyznaczyć oraz zakres częstotliwości $[f_{lb}, f_{ub}]$, w których zostaną one wyznaczone. Następnie korzystając z wzoru (22) wartości f_{lb} oraz f_{ub} zostają przekształcone do skali melowej dając odpowiednio $M_{f_{lb}}$ oraz $M_{f_{ub}}$.

$$M(f) = 1127ln\left(1 + \frac{f}{700}\right), \tag{22}$$

$$f(M) = 700\left(e^{\frac{M}{1127}} - 1\right), \tag{23}$$

Pomiędzy granicami $M_{f_{lb}}$ oraz $M_{f_{ub}}$ wybiera się $N + 2$ równomiernie rozłożone punkty a następnie każda z wartości w wektorze zostaje przekształcona z powrotem do częstotliwości z wykorzystaniem wzoru (23) tworząc wektor częstotliwości melowych f_m.

Po znalezieniu wektora f_m należy przekonwertować go do najbliższych częstotliwości możliwych do zaobserwowania na spektrum gęstości widmowej sygnału o znanej rozdzielczości,

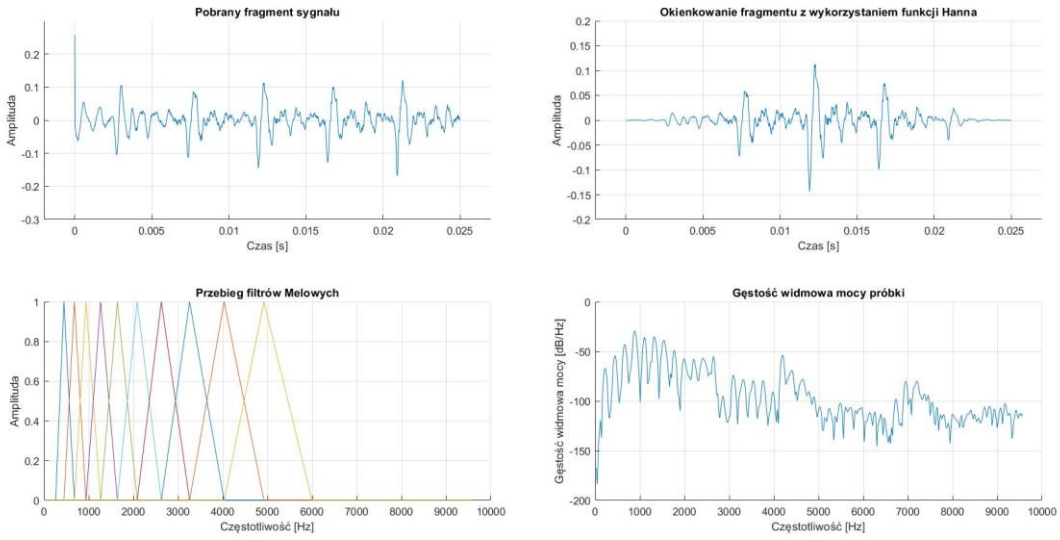
$$f_r(f_m) = Ent\left(\frac{(NFFT+1)f_m}{f_s}\right), \tag{24}$$

Gdzie $NFFT$ to długość okna szybkiej transformaty Fouriera a f_s to częstotliwość próbkowania sygnału.

Następnie korzystając ze wzoru (25) utworzony zostaje bank filtrów melowych H.

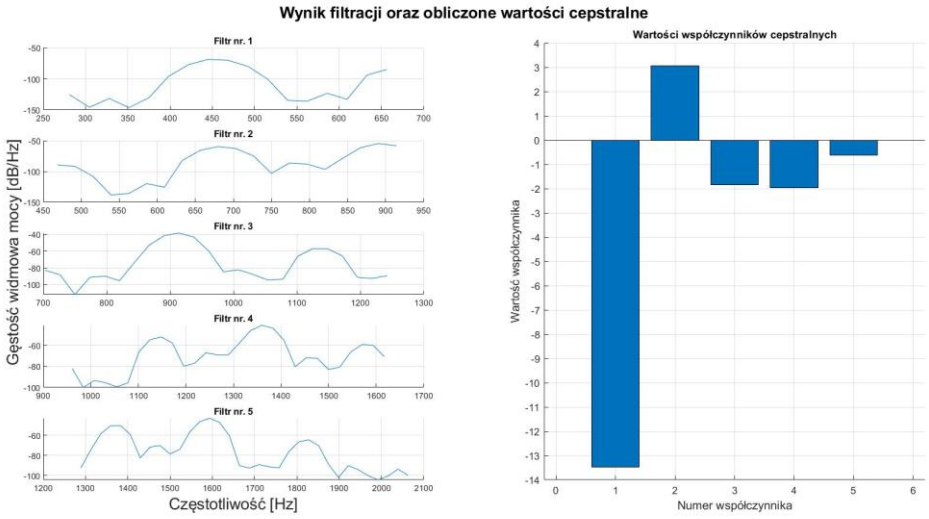
$$H_m(k) = \begin{cases} 0 & dla\ k<f_r(m-1) \\ \frac{k-f_r(m-1)}{f_r(m)-f_r(m-1)} & dla\ f_r(m-1)\leq k\leq f_r(m) \\ \frac{f_r(m+1)-k}{f_r(m+1)-f_r(m)} & dla\ f_r(m)\leq k\leq f_r(m+1) \\ 0 & dla\ k>f_r(m-1) \end{cases}, \tag{25}$$

Gdzie m to numer filtra, a k to wszystkie liczby całkowite z przedziału $\langle 0, NFFT\rangle$.



Rysunek 22. Przebiegi sygnału wykorzystywanego na różnych etapach algorytmu.

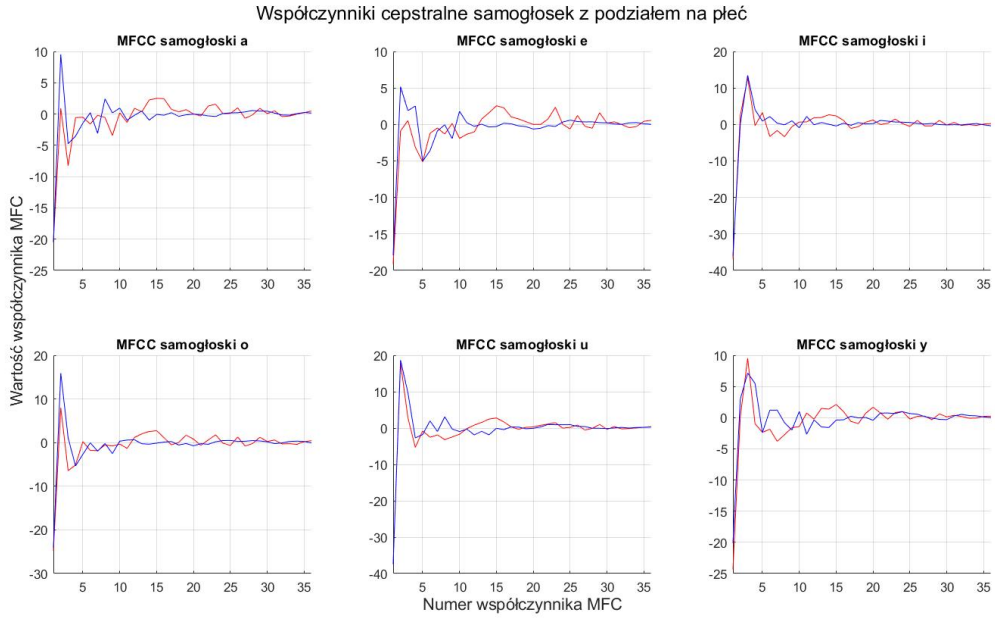
Po obliczeniu banku filtrów melowych, algorytm zaczyna od filtracji badanego sygnału z zastosowaniem filtra preemfazy. Następnie sygnał jest dzielony na fragmenty zwykle o długości 25 – 40 ms, zbierane tak by fragmenty nachodziły na siebie w 60%. Każdy z fragmentów zostaje przemnożony przez okno czasowe o odpowiedniej długości. Dla tak przetworzonego fragmentu oblicza się spektrum gęstości widmowej mocy. Spektrum to następnie zostaje przemnożone przez każdy z filtrów melowych oraz utworzony zostaje wektor obliczonych logarytmów sum energii każdego filtra w banku. Współczynniki Melowe Cepstrum otrzymuje się po poddaniu uzyskanego wektora odwrotnej transformacji kosinusowej.



Rysunek 23. Wartości energii w bankach melowych dla 5 z 10 filtrów oraz obliczone z nich współczynniki cepstralne

6.2. Analiza współczynników MFC

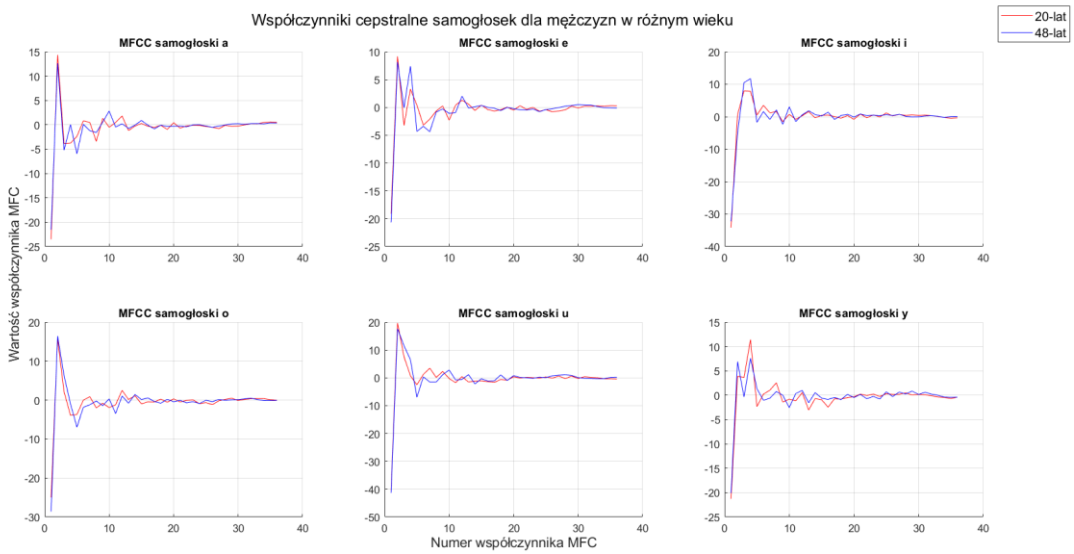
Wykorzystując algorytm opisany w rozdziale 6.1 przebadane zostały nagrania wypowiadanych samogłosek. Dla każdego nagrania obliczono 36 współczynników MFCC z przedziału częstotliwości od 300 Hz do 4000 Hz. Wyniki uśredniono z podziałem na płeć oraz wypowiadane samogłoski, następnie stworzono na ich podstawie wykresy (Rys. 24).



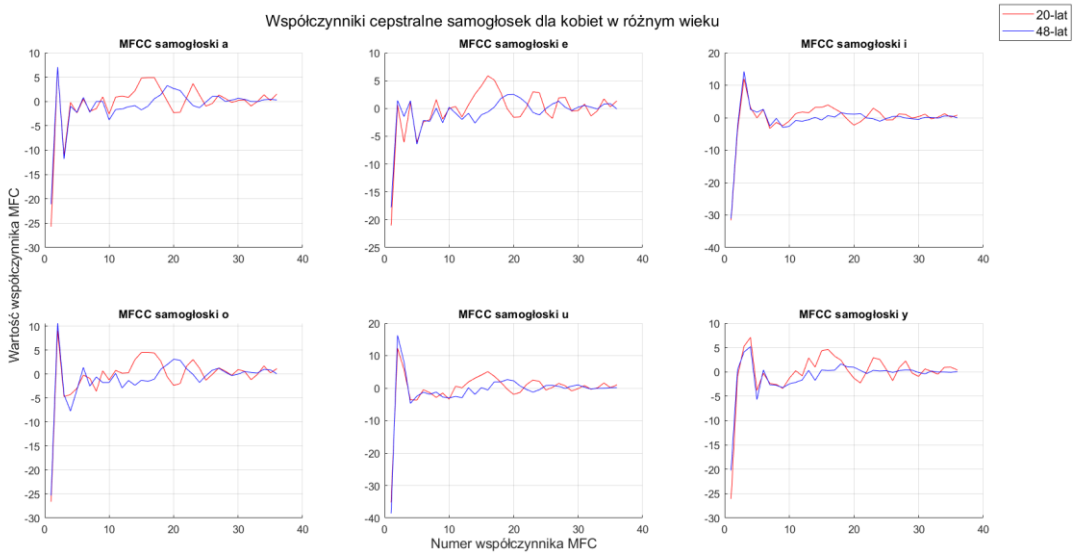
Rysunek 24. Wartości współczynników MFCC z podziałem na płeć oraz samogłoskę. Kolorem niebieskim zaznaczone zostały wyniki mężczyzn natomiast czerwonym- kobiet.

Wyniki pokazują, że wartości współczynników cepstralnych dla mężczyzn zwykle dominują w początkowych 10 MFCC. Wygładzają się one natomiast szybciej oraz posiadają mniej energii dla filtrów melowych wyższych częstotliwości od współczynników należących do głosów żeńskich. Można także zaważyć, że różne samogłoski posiadają wyróżniające je rozkłady energii.

Następnie wykorzystując obliczone wartości stworzono wykresy z podziałem na wiek oraz wypowiadaną samogłoskę osobno dla kobiet oraz mężczyzn.



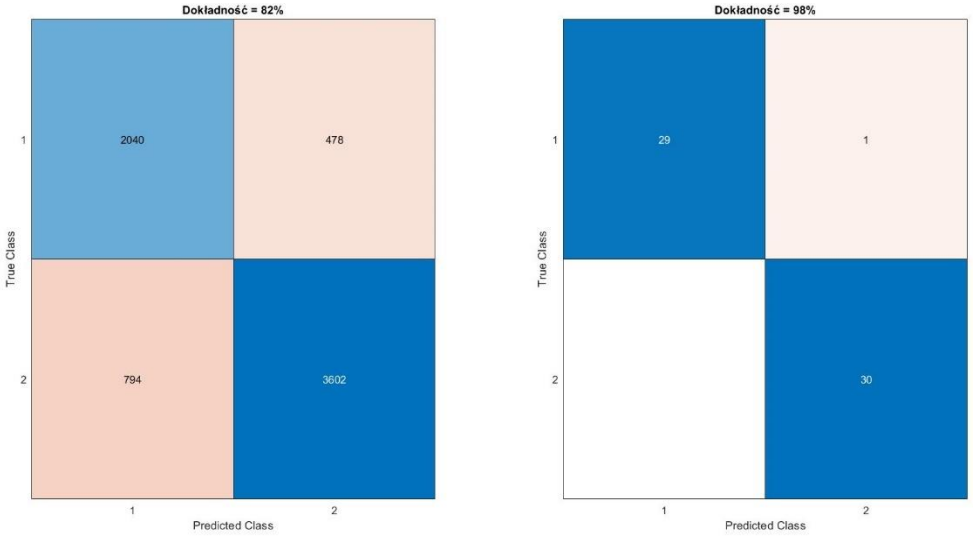
Rysunek 25. Współczynniki cepstralne samogłosek dla mężczyzn w różnym wieku.



Rysunek 26. Współczynniki cepstralne samogłosek dla kobiet w różnym wieku.

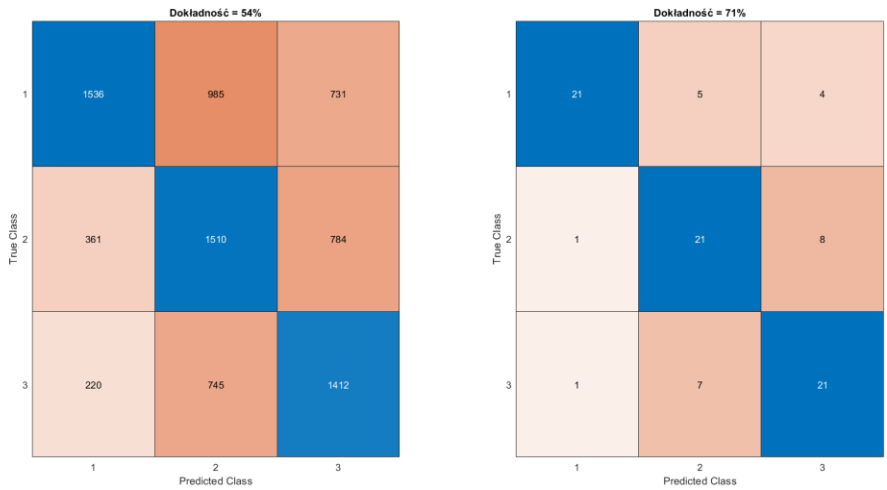
Wpływ wieku na współczynniki MFC jest o wiele bardziej widoczny dla kobiet. Największą różnicą jest zmniejszenie się wartości MFC dla filtrów badających wyższe częstotliwości. Na wykresie samogłosek dla kobiety w wieku 20 lat można zaobserwować widoczny wzrost wartości współczynników w okolicach 23 filtru, dla kobiety w wieku 48 lat wzrost ten jest dużo mniejszy oraz przesunięty w okolice 25-26 filtru.

Do badań możliwości współczynników MFC dla bazy danych Mozilla Common Voice wybrano po 120 nagrań męskich oraz żeńskich. Postępując podobnie jak podczas analizy formantów z wykorzystaniem modeli GMM, stworzono bazę fragmentów uprzednio wybranych nagrań dla których utworzono wektor cech złożony z współczynników MFCC. Fragmenty dla 90 nagrań zostały wykorzystane do trenowania poszczególnych modeli GMM natomiast pozostałe 30 zostały wykorzystane do testowania ich skuteczności. Najlepsze wyniki udało nam się uzyskać dla 32 filtrów melowych w zakresie od 300 do 4000Hz oraz 16 komponentów gaussowskich dla modelu GMM. Wynik wynosił odpowiednio 82% skuteczności wykrywania płci dla ramek oraz 98% wykrywania dla całych nagrań.



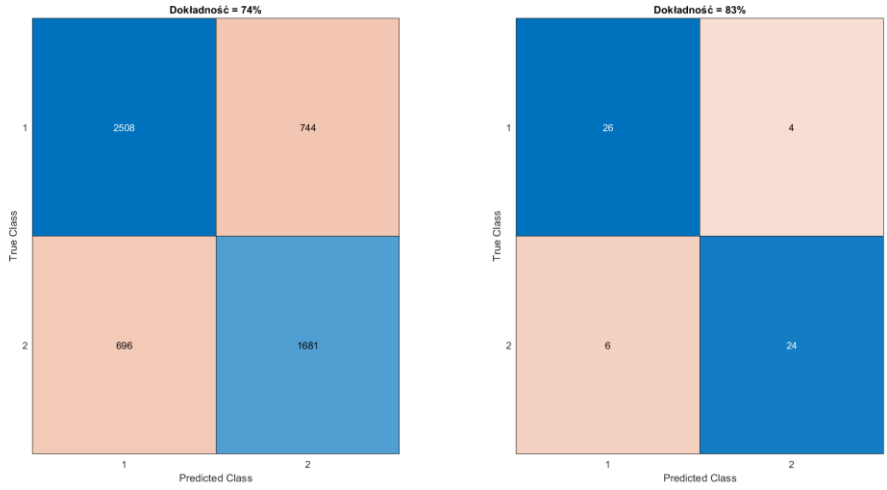
Rysunek 27. Macierze pomyłek rozpoznawania płci dla ramek (lewy) oraz nagrań (prawy). 1-mężczyźni, 2-kobiety.

Do badania wykorzystano po 130 nagrań dla trzech najliczniejszych grup wiekowych w bazie. Podobnie jak w przypadku wykrywania płci, po 100 nagrań dla każdego zbioru wykorzystano do trenowania modeli GMM natomiast pozostałe do testowania jego jakości. Najlepsze wyniki uzyskano dla ilości współczynników równej 32-36 zebranych z przedziału 300-7000Hz oraz ilości komponentów gaussowskich z przedziału 3-7. Uzyskana dokładność dla ramek wahała się między 52-56% co pozwoliło na uzyskanie poprawności dla nagrań na poziomie 68-74%.



Rysunek 28. Macierze pomyłek rozpoznawania wieku dla ramek (lewy) oraz nagrań (prawy). 1=13-19 lat, 2=20-29 lat, 3=30-39 lat

Obserwując macierze pomyłek przedstawione na rysunku (Rys. 28) można zauważyć, że najczęściej mylone są próbki pochodzące z przedziałów sąsiadujących ze sobą wiekowo. Powodem tej sytuacji mogą być bardzo małe różnice wiekowe na granicach przedziałów. Przykładowo osoba 29 letnia będąc jeszcze w przedziale drugim przedziale wiekowym, jest tylko rok młodsza od osoby w wieku 30 lat będącej w trzecim przedziale. Mając to na uwadze sprawdzono, jak przedstawia się dokładność rozpoznawania wieku mówców należących tylko do pierwszego oraz trzeciego przedziału wiekowego. Najlepsze wyniki uzyskano dla 32 współczynników MFC oraz 16 komponentów gaussowskich modelu GMM. Dokładność wzrosła do ~80% dla osoby oraz ~70% dla ramek.



Rysunek 29 Macierze pomyłek rozpoznawania wieku dla ramek (lewy) oraz nagrań (prawy). 1- 13-19 lat, 2 – 30-39 lat

Wnioski

Wyciąganie cech osobniczych mówcy z nagrań jego głosu jest złożonym, wieloetapowym procesem, wykorzystującym wiele różnych metod oraz algorytmów. W pracy przedstawione zostały sposoby na poprawę jakości nagrań, estymację parametrów głosu oraz ich powiązanie z cechami mówcy, takimi jak płeć czy wiek. Pokazany został także przykład wykorzystywania otrzymanych parametrów celem automatycznego rozpoznawania mówcy z wykorzystaniem modeli GMM.

Wszystkie analizowane w pracy cechy głosu, czyli ton podstawowy, formanty oraz współczynniki MFC pozwoliły na uzyskanie wysokiej skuteczności podczas rozpoznawania płci mówcy dla nagrań z bazy danych Mozilla Common Voice. Najlepiej sprawdziły się współczynniki MFC oraz ton podstawowy, dając ponad 95% skuteczność rozpoznawania z wykorzystaniem modeli GMM. Gorzej, bo w granicach 85%, plasowały się formanty.

Rozpoznawanie wieku, zwłaszcza w bliskich przedziałach jest zagadnieniem trudniejszym niż rozpoznawanie płci. Dla nagrań dostępnych w wykorzystywanej bazie danych ton podstawowy nie wykazał zależności między badanymi przedziałami wiekowymi. Formanty pozwoliły na ~50% dokładność w rozpoznawaniu wieku między przedziałami 13-19, 20-29 oraz 30-38 lat, natomiast współczynniki MFC na ~70% dokładność.

Istnieje dalsza możliwość rozwoju pracy poprzez zbadanie dokładności rozpoznawania wieku oraz płci dla próbek zawierających nagrania osób z innych przedziałów wiekowych, przykładowo osób starszych oraz dzieci. Możliwe jest także przeprowadzenie analizy dla baz nagrań zawierających obcojęzyczne nagrania oraz porównanie wyników z bazą polskojęzyczną, celem sprawdzenia możliwości wykazania różnic wynikających z pochodzenia.

Innym kierunkiem rozwoju pracy może być próba wykorzystania dodatkowych parametrów takich jak ΔMFCC lub ΔΔMFCC oraz łączenie kilku parametrów celem poszukiwania najlepszej możliwej dokładności dla rozpoznawania płci oraz wieku osób będących w polskiej bazie nagrań Mozilla Common Voice.

Bibliografia

[1] Mozilla Common Voice, https://commonvoice.mozilla.org

[2] J. Picone, *Signal Modeling Techniques in Speech Recognition*, Proceedings of the IEEE 81, pp.1215–1247, 1993.

[3] S. Vihari, A. Murthy, P. Soni oraz D. Naik, *Comparison of speech enhancement algorithms*, Procedia computer science, vol. 89, pp. 666–676, 2016.

[4] C. Plapous, C. Marro, L. Mauuary and P. Scalart, *A two-step noise reduction technique*, Proc. IEEE Int. Conf. Acoust. Speech Signal Process, pp. 289-292, 2004.

[5] S. Gannot, *Speech Enhancement - Audio Samples*, http://www.eng.biu.ac.il/~gannot/examples1.html

[6] D. Reynolds, *Gaussian Mixture Models*, Encyclopedia of Biometrics, Springer, Boston, MA 2009

[7] Scikit-learn, *Gaussian mixture models*, https://scikit-learn.org/stable/modules/mixture.html

[8] P. Smyth, *Notes on the EM Algorithm for Gaussian Mixtures: CS 274A, Probabilistic Learning*

[9] I. Titze, *Physiologic and acoustic differences between male and female voices*, Journal of the Acoustical Society of America 85, 1998

[10] P. Keating, G. Kuo, *Comparison of speaking fundamental frequency in English and Mandarin*, PubMed, 2012

[11] M. Ayoub, P. Larrouy-Maestri, D. Morsomme, *The Effect of Smoking on the Fundamental Frequency of the Speaking Voice*, Journal of Voice, Volume 33, Issue 5, 2019

[12] S. Xue, D. Deliyski, *Effects of aging on selected acoustic voice parameters: Preliminary normative data and educational implications*, Educ. Gerontol, 2001

[13] S. Harris, *The Voice & Ageing*, The British Voice Association

[14] D. Jouvet, Y. Laprie, *Performance Analysis of Several Pitch DetectionAlgorithms on Simulated and Real Noisy Speech Data*, 25th European Signal Processing Conference, 2017

[15] L. Tan, M. Karnjanadecha, *Pitch detection algorithm: autocorrelation method and AMDF*, Proceedings of the 3rd International Symposium on Communications and Information Technology, Vol. II, pp. 541-546, 2003

[16] B. Bogert, M. Healy, J. Tukey, *The Quefrency Alanysis of Time series for Echoes: Cepstrum, Pseudo-Autocovariance, Cross-Cepstrum, and Saphe Cracking*, Wiley, 1963

[17] R. Randall, *A History of Cepstrum Analysis and its Application to Mechanical Problem*, 2017

[18] F. Le Bourdias, *A Short Tutorial on Cepstral Analysis for Pitch-tracking, http://flothesof.github.io/cepstrum-pitch-tracking.html*

[19] P. Cyril, B. Pascal, *The Role of Pitch and Timbre in Voice Gender Categorization*, Frontiers in Psychology Vol. 3, 2012

[20] J. Bachorowski, M. Orwen, *Acoustic correlates of talker sex and individual talker identity are present in a short vowel segment produced in running speech*, The Journal of the Acoustical Society of America 106, 1999.

[21] M. Drobner, S. Golachowski, *Akustyka muzyczna*, Polskie Wydawnictwo Muzyczne, Kraków 1953
[22] R. Snell, *Formant Location From LPC Analysis Data*, IEEE, 1993
[23] S. Davis; P. Mermelstein, *Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences*, IEEE Transactions on Acoustics, Speech, and Signal Processing, 1980
[24] James Lyons, *Mel Frequency Cepstral Coefficient (MFCC) tutorial*, http://practicalcryptography.com