POLITECHNIKA WROCŁAWSKA
WYDZIAŁ ELEKTRONIKI

KIERUNEK: Automatyka i Robotyka
SPECJALNOŚĆ: Technologie informacyjne w systemach automatyki (ART)

PRACA DYPLOMOWA INŻYNIERSKA

Detekcja aktywności mówcy w systemach automatycznego rozpoznawania mowy

Voice activity detection in automatic speech recognition systems

AUTOR:
Paulina Szczerbak

PROWADZĄCY PRACĘ:
Prof. dr hab. inż. Ryszard Makowski

OCENA PRACY:

WROCŁAW 2017

Spis treści

Rozdział 1

Wstęp

Celem niniejszej pracy jest zaprezentowanie wybranych metod detekcji aktywności mówcy (VAD) w systemach automatycznego rozpoznawania mowy w oparciu o napisany program w języku C++. Kolejnym etapem jest porównanie zaimplementowanych metod pod względem dokładności detekcji w separowanych wyrazach oraz w dłuższych ciągach słów.

Rozdział 2. opisuje w uproszczony sposób proces wytwarzania mowy przez człowieka. Prezentuje, w jaki sposób działa aparat mowy oraz z jakich narządów się składa. Na koniec pokazany jest matematyczny model jaki można stworzyć wzorując się na naturalnym systemie generowania mowy.

Rozdział 3. zawiera wyjaśnienie na temat detekcji aktywności mówcy - czym jest oraz gdzie jest wykorzystywana. W tym rodziale opisane są również wybrane algorytmy, które zostały zestawione w dalszej części pracy. Przedstawiona jest zasada ich działania oraz pokrótce wyjaśniona kwestia implementacyjna każdego z nich.

Rozdział 4. objaśnia niektóre aspekty implementacyjne programu.

Rozdział 5. przedstawia sposób oceny wyników. Zawiera wyniki detekcji dla pojedynczych słów oraz całych ciągów. Na końcu tego rozdziału znajduje się również ostateczne porównanie w działaniu wybranych algorytmów i ich ocena.

Rozdział 2

Generowanie sygnału mowy

2.1 Mowa w życiu człowieka

Mowa w życiu większości ludzi stanowi podstawę komunikacji interpersonalnej. Jest sygnałem akustycznym, czyli rozważany jest zakres częstotliwości słyszanych przez człowieka, to jest od 20Hz do 16kHz. Zatem mowa to nic innego jak system artykułowanych dźwięków, które układają się zgodnie z konwencją wybranego języka. Pełni ona funkcję nie tylko komunikacyjną (przekazywanie informacji drugiej osobie o tym, co doświadczyliśmy, czy czego się dowiedzieliśmy), ale również ekspresyjną (można w niej zawrzeć informacje o emocjach nadawcy) oraz regulacyjną (wydawanie i przyjmowanie dyspozycji).

2.2 Biologiczny proces generowania mowy

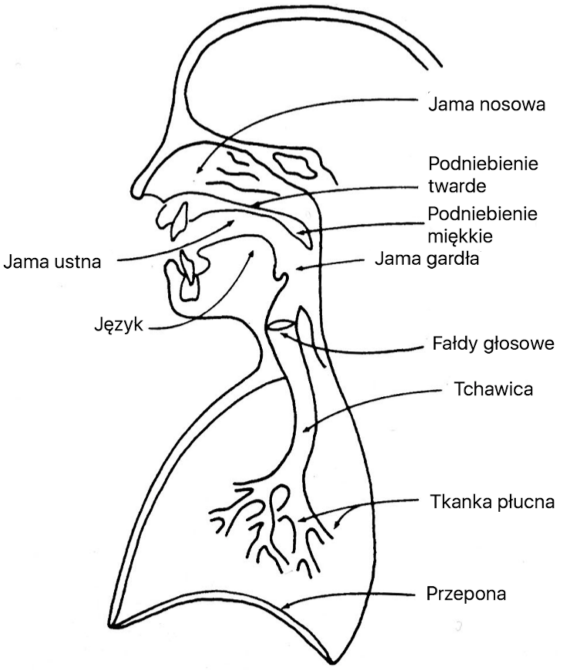
Wszelkie metody przetwarzania sygnału mowy muszą bazować na strukturze sygnału, a ta jest niewątpliwie uzależniona od sposobu, w jaki jest on wytwarzany. Niegdyś generowanie sygnałów mowy było domeną jedynie organizmu człowieka, czyli systemu naturalnego. W celu stworzenia systemu, który w jakiś sposób operuje na sygnałach mowy, czyli np. syntezatora mowy, systemu generującego sygnały mowopodobne, systemu automatycznego rozpoznawania mowy czy detekcji aktywności mówcy, należy mieć przynajmniej podstawową wiedzę na temat systemu naturalnego - tego, w jaki sposób działa aparat mowy człowieka.

Wytwarzanie mowy przez człowieka jest procesem niezwykle skomplikowanym, który ma swój początek w mózgu, gdzie następuje konstrukcja wypowiedzi. Później następuje sformułowanie fonetyki i artykulacja poprzez aparat mowy. Ponadto, w procesie generowania mowy można wyróżnić cztery pomniejsze etapy:

- proces psychologiczny - wymyślenie i skonstruowanie wypowiedzi,
- proces neurologiczny - pobudzenie przez układ nerwowy mięśni, które biorą udział w wytwarzaniu mowy,
- proces fizjologiczny - proces kształtowania dźwięków mowy ludzkiej,
- proces aerodynamiczny - drgania i przepływ powietrza przez aparat mowy.

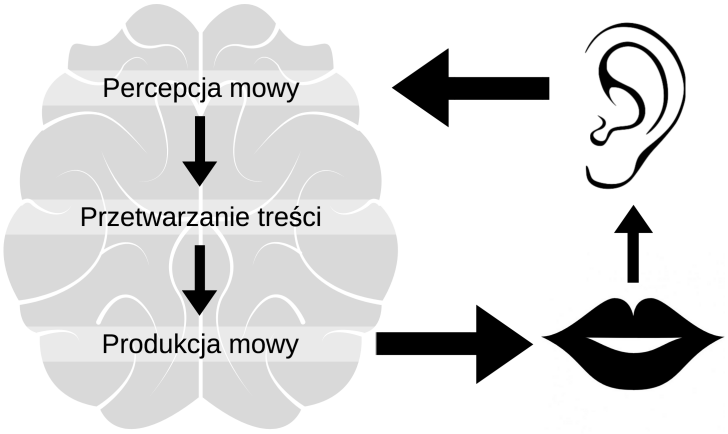
Pierwszym narządem wchodzącym w skład traktu głosowego człowieka są płuca - dostarczają one powietrze do procesu artykulacji, są źródłem zmian ciśnienia akustycznego. Organ mowy człowieka jest napędzany przez wydychane powietrze. Powietrze to, jest prowadzone przez oskrzela i tchawicę do krtani, a drgające w niej struny głosowe modyfikują ciśnienie i wytwarzają dźwięczne fragmenty mowy. Następnie, dzięki wnękom rezonansowym, tworzonym przez język, podniebienie, zęby oraz wargi, dźwięk ten jest modulowany.

Niezwykle ważną rolę przy formowaniu tych wnęk, odgrywają ruchy żuchwy i policzków. Podczas generowania głosek nosowych zamknięta jama ustna spełnia rolę bocznika akustycznego, a dzięki odpowiedniemu ustawieniu języczka podniebienia miękkiego, fala dźwiękowa jest emitowana przez jamę nosową i nozdrza. Struktura traktu głosowego jest przedstawiona schematycznie na rysunku 2.1.



Rysunek 2.1 Aparat mowy człowieka

Ponadto, sterowanie całym systemem generowania mowy jest bardzo złożone i w dużej mierze opiera się na licznych sprzężeniach zwrotnych. Główną rolę odgrywa tutaj sprzężenie zwrotne, które poddaje jakość wydawanych dźwięków bezpośredniej ocenie poprzez analizator słuchowy. Dzięki temu proces artykulacji jest odpowiednio sterowany. Istotę tego sprzężenia zwrotnego potwierdzają trudności z mową wśród ludzi głuchych oraz ludzi słyszących, którzy tymczasowo przebywają w trudnych warunkach środowiskowych, które uniemożliwiają słyszenie własnego głosu.



Rysunek 2.2 Sprzężenie zwrotne w wytwarzaniu mowy

Rozdział 3

Wybrane metody detekcji aktywności mówcy

3.1 Czym jest detekcja aktywności mówcy oraz gdzie się ją wykorzystuje

Detekcja aktywności mówcy (Voice Activity Detection - VAD) jest powszechnie stosowana w systemach automatycznego rozpoznawania mowy. Podczas rejestrowania wypowiedzi do późniejszego przetwarzania jej przez system ARM, zostaje zarejestrowana cała wypowiedź mówcy, włącznie z częścią, która nie zawiera mowy. Jeżeli we fragmencie jest zawarty sygnał mowy, mówimy, że mówca jest aktywny. Aktywnością mówcy nazywa się emitowny przez niego dźwięk. Zawartość semantyczna wypowiedzi jest zawarta w głównej mierze we fragmentach, kiedy mówca jest aktywny. Analizowanie całego zarejestrowanego sygnału mowy, bez wykorzystania systemu VAD, jest oczywiście możliwe, aczkolwiek niepotrzebnie zwiększa czas obliczeń oraz istnieje prawdopodobieństwo, że fragment, gdy mówca nie jest aktywny, zostanie błędnie zaklasyfikowany jako jakiś konkretny fonem - zatem w dużej mierze może popsuć jakość rozpoznania. Detekcja aktywności mówcy w ogólnym przypadku zakłada, że sygnał może występować w dwóch stanach: tylko szum (brak sygnału mowy), szum + sygnał mowy. Korzystając z zagadnienia hipotez ze statystyki, możemy pierwszy stan oznaczyć jako hipotezę H_0, a drugi jako H_1, dzięki czemu możemy przedstawić to w następujący sposób:

$$\begin{aligned} H_0 &: f(n) = x(n) \\ H_1 &: f(n) = v(n) + x(n) \end{aligned} \tag{3.1}$$

Przy takim rozumowaniu konieczne jest określenie statystyki $S(n)$ sygnału, dzięki czemu możliwe będzie dokonywanie detekcji, a w dalszej kolejności zastosowanie kryterium decyzyjnego. Kryterium decyzyjne zwykle polega na porównaniu wartości $S(n)$ z progiem detekcji, który w mniej skomplikowanych algorytmach przyjmuje stałą wartość. Natomiast w tych bardziej złożonych, może występować np. jako funkcja czasu. Wartość stałej wartości progu jest ustalana w wyniku teoretycznych rozważań lub empirycznie. Zatem detekcja $\gamma(n)$, w ogólnej postaci, będzie prezentować się następująco:

$$\begin{aligned} S(n) &\geqslant \gamma(n) \rightarrow H_1 \\ S(n) &< \gamma(n) \rightarrow H_0 \end{aligned} \tag{3.2}$$

3.2 Metoda bazująca na energii sygnału z adaptacyjnym współczynnikiem skalującym

Najbardziej powszechną metodą do obliczenia energii dla całego pasma w sygnale mowy jest:

$$E_j = \frac{1}{N} \sum_{i=(j-1)N+1}^{jN} x^2(i)$$

gdzie: E_j - energia j-tej ramki

3.2.1 Początkowy próg detekcji

Początkowa wartość progu jest ważna dla jego dalszego rozwoju - ponieważ będzie się zmieniał zgodnie ze śledzonym poziomem szumu w sygnale. Przyjęto, że początkowe 100 ms nagrania nie zawiera mowy. Jest to podyktowane tym, że mówca potrzebuje czasu na rekację, nabranie powietrza, aktywację strun głosowych. Te 100 ms są uznawane za przebieg pozbawiony sygnału mowy i ich średnia wartość obliczana jest zgodnie z powyższym wzorem.

3.2.2 Próg detekcji zmieniający się dynamicznie w czasie

Główną ideą tego algorytmu jest możliwośc obilczenia progu detekcji bez potrzeby korzystania z obszarów niezawierajacych mowy. Wykorzystywana jest minimalna oraz maksymalna energia sygnału mowy.

Innym popularnym sposobem na obliczenie energii sygnału mowy jest pierwiastek ze średniokwadratowej wartości energii (root mean square energy - RMSE), dany jako:

$$E_j = \sqrt{\frac{1}{N} \sum_{i=(j-1)N+1}^{jN} x^2(i)}$$

'Dynamiczny' VAD jest oparty o obserwację, że estymata mocy sygnału mowy pokazuje wyraźne szczyty i doliny. Podczas gdy szczyty odpowiadają aktywności mowy, doliny mogą zostać wykorzystane do uzyskania estymaty mocy szumu. Ponadto, RMSE jest bardziej odpowiedni.

3.2.3 Wyznaczanie progu

Estymacja progu bazuje na poziomach energii E_{min} oraz E_{max} otrzymane z ciągu nadchodzących ramek. Te wartości są trzymane w pamięci i próg θ jest obliczany jako:

$$\theta = k_1 E_{max} + k_2 E_{min}$$

gdzie k_1 i k_2 to współczynniki wykorzystane do interpolacji wartości progu dla optymalnych wyników.

Jeżeli energia bieżęcej ramki jest mniejsza niż wartość progu, ramka zostaje oznaczana jako niezawierająca mowy.

Jako, że mogą się pojawić pewne anomalie spowodowane zbyt niską energią, wprowadzono odpowiednią prewencję. Parametr E_{min} jest nieznacznie zwiększany dla każdej ramki, zdefiniowane jako:

$$E_{min}(j) = E_{min}(j-1)\Delta(j)$$

Parametr Δ dla każdej ramki jest zdefiniowany jako:

$$\Delta(j) = \Delta(j-1) \cdot 1.0001$$

3.2.4 Rozszerzenie algorytmu

Możliwe jest przedstawienie równania na wyznaczenie dynamicznie zmieniającego się progu przy pomocy jednego parametru λ (np. $\lambda = k_2$).

$$\theta = (1-\lambda)E_{max} + \lambda E_{min}$$

gdzie λ to parametr skalujący , który kontroluje proces estymacji.

Detektor mowy działa wiarygodnie, gdy λ należy do przedziału [0.950, ..., 0.999]. Jednakże, wartości dla różnych typów sygnałów mogą nie być takie same i informacja a priori wciąż wymaga poprawnego ustalenia wartości λ. Równanie poniżej pokazuje jak sprawić, żeby współczynnik skalujący λ był niezależny i odporny na zmieniające się warunki środowiska.

$$\lambda = \frac{E_{max} - E_{min}}{E_{max}}$$

3.3 Metoda bazująca na obwiedni sygnału podzielonego na pasma z filtracją pojedynczych częstotliwości

Do dokonania detekcji, należy wcześniej policzyć obwiednię dla sygnału podzielonego na 185 pasm - od 30Hz do 4000Hz, co 20Hz. Wybrany przedział częstotliwości pokrywa się z użytecznym pasmem, który jest wykorzystywany przez mowę. Poniżej zostały przedstawione kolejne kroki potrzebne do policzenia 185 obwiedni i odpowiedniego przekształcenia ich w funkcję czasu, na której będzie można dokonać detekcji.

Sygnał mowy ma zależności zarówno w dziedzinie czasu, jak i w dziedzinie częstotliwości. Skutkuje to tym, że stosunek sygnał-szum (SNR) jest funkcją czasu, jak i częstotliwości. DLa idealnego szumu o danej całkowitej mocy, moc jest równo rozdzielona na wszystkie częstotliwości, podcza

3.3.1 Obwiednie sygnału dla każdej częstotliwości

Sygnał mowy w zdyskretyzowanej dziedzinie czasu $s(n)$ jest różniczkowany i jest rozumiany jako $x(n) = s(n) - s(n-1)$. Częstotliwość próbkowania to fs. Sygnał $x(n)$ jest przemnażany przez zespoloną sinusoidę o danej znormalizowanej częstotliwości $\bar{\omega}_k$. Wynik tej operacji w dziedzinie czasu jest dany jako:

$$x_k(n) = x(n)e^{j\bar{\omega}_k n},$$

$$\text{gdzie } \bar{\omega}_k = \frac{2\pi\bar{f}_k}{f_s}$$

Kiedy pomnożymy $x(n)$ przez $e^{j\bar{\omega}_k n}$, wynikowe widmo $x_k(n)$ będzie przesuniętym widmem $x(n)$. Czyli:

$$X_k(\omega) = X(\omega - \bar{\omega}_k),$$

gdzie $X_k(\omega)$ i $X(\omega)$ to odpowiednio widma $x_k(n)$ i $x(n)$.

Sygnał $x_k(n)$ jest przepuszczany przez jednobiegunowy filtr, którego transmitancja jest dana jako:

$$H(z) = \frac{1}{1 + rz^{-1}}$$

Jednobiegunowy filtr ma biegun na osi liczb rzeczywistych w odległości r od początku układu współrzędnych. Lokalizacja pierwiastka jest w $z = -r$ na płaszczyśnie liczb zespolonych, co odpowiada połowie częstotliwości próbkowania, np. $fs/2$. Wyjście filtra $y_k(n)$ jest dane jako:

$$y_k(n) = -ry_k(n-1) + x_k(n)$$

Obwiednia sygnału $y_k(n)$ jest dana jako:

$$e_k(n) = \sqrt{y_{kr}^2(n) + y_{ki}^2(n)},$$

gdzie $y_k r(n)$ i $y_{ki}(n)$ są odpowiednio częścią rzeczywista i urojoną $y_k(n)$.

Kiedy filtrowanie $x_k(n)$ będzie zrobione dla $\frac{f_s}{2}$, powyższa obwiednia $e_k(n)$ będzie odpowiadać obwiedni sygnału $x_k(n)$ przefiltrowanego w pożądanej częstotliwości

$$f_k = \frac{f_s}{2} - \bar{f}_k$$

.

Powyższa metoda estymowania obwiedni składowej dla częstotliwości f_k jest określana jako podejście filtracji pojedynczych częstotliwości (Single Frequency Filtering). Wybór filtra z biegunem w $z = -r$ do estymacji obwiedni przefiltrownego sygnału wydaje się być bardziej odpowiedni, jako że obwiednie są obliczane w możliwie najwyższych częstotliwościach ($f_s/2$). Ponadto, wybór filtru w stałej częstotliwości dla jakiejkolwiek pożądaniej częstotliwości f_k zapobiega efektu przeskalowania w związku z różnymi wzmocnieniami filtrów w różnych częstotliwościach. Jeżeli biegun zostanie wybrany z obszaru na kole jednostkowym, np $z = r = -1$, może to skutkować niestabilnością wyjścia filtru. Stabilność filtru jest zapewniona dzięki przesunięciu bieguna nieco bardziej wewnątrz koła jednostkowego. Z tego powodu r zostało dobrane jako 0.99.

W tym badaniu obwiednia została obliczona dla każdych 20Hz w przedziale od 300Hz do 3000Hz jako funkcja w dziedzinie czasu. Wybrany został przedział częstotliwości 300-4000Hz, ponieważ pokrywa się z użytecznym pasmem wykorzystywanym przez mowę. Zatem mamy obwiednie dla 185 częstotliwości jako funkcja w dziedzinie czasu. Zasadniczo obwiednia może zostać obliczona dla każdej pożądanej częstotliwości.

3.3.2 Ważone składowe obwiedni sygnału mowy

Kiedy sygnał mowy ma bardzo dużą rozpiętość tonalną w dziedzinie częstotliwości, sygnał może mieć wysoką wartość mocy w niektórych częstotliwościach w każdej chwili czasowej. W tych częstotliwościach SNR będzie miał większą wartość, jako, że moc szumu będzie prawdopodobnie mniejsza w związku z większym rozkładem jednostajnym mocy. Nawet dla szumów z nierównomiernym rozkładem mocy, niższe korelacje próbek szumu skutkują w niższej rozpiętości tonalnej w rozpiętości mocy szumu przez częstotliwość, w porównaniu z sygnałem mowy. Zauważmy, że widmowa rozpiętość tonalna daje przejaw korelacji próbek w dziedzinie czasu.

Moc szumu tworzy funkcję cechy (podłogi) dla obwiedni dla każdej częstotliwości i poziom cechy zależy od rozkładu mocy szumu wobec częstotliwości. Podłoga jest bardziej jednorodna wobec czasu, jeżeli szum jest niemalże stacjonarny. Nawet jeżeli szum jest niestacjonarny, jest względnie stacjonarny ponad większymi przerwami w czasie niż sygnał mowy. W takich przypadkach, poziom cechy może zostać obliczony ponad długimi przerwami w dziedzinie czasu dla każdej częstotliwości, jeżeli jest to potrzebne.

Żeby zrekompensować efekt szumu, wartość wagi dla każdej częstotliwości jest obliczana używając wartości funkcji cechy. Dla każdego wyrażenia, średnia (μ_k) z 20 % najmniejszych wartości, wartości obwiedni dla każdej częstotliwości f_k jest wykorzystywana do obliczenia znormalizowanej wagi wartości ω_k dla danej częstotliwości. Wybór akurat 20% wartości jest oparty o założenie, że w jest przynajmniej 20% ciszy w każdym wyrażeniu mowy. Znormalizowana waga wartości w każdej częstotiwości jest dana jako:

$$\omega_k = \frac{\frac{1}{\mu_k}}{\sum_{l=1}^{N} \frac{1}{\mu_t}},$$

gdzie N to liczba kanałów.

Obwiednia $e_k(n)$ dla każdej częstotliwości f_k jest przemnażana przez wartość wagi w_k w celu zrekompensowania poziomu szumu w tej częstotiwości. Wynikowa obwiednia jest określana jako obwiednia z ważonymi komponentami. Zauważmy, że przez to ważenie, obwiednia dla każdej częstotliwości jest dzielona przez estymatę cechy szumu u_k.

Do wszystkich sygnałów została dodana mała ilośc białego szumu, żeby mieć pewność, że wartość funkcji cechy nie jest zerem. Dla obliczeń w_k, wartości w dodanych obszarach ciszy nie będą rozważane.

W każdej chwili czasu średnia ($\mu(n)$) kwadratu ważonych obwiedni obliczonych wobec częstotliwości odpowiada w przybliżeniu energii sygnału w danej chwili (rys 2(c)) Oczekuje się, że $\mu(n)$ będzie wyższe dla mowy, niż dla szumu w obszarach, gdzie występuje sygnał mowy, ponieważ wartosci szumu są o obniżonej wadze. W każdej chwili czasu, odchylenie standardowe ($\sigma(n)$) kwadratu ważonych obwiedni rownież będzie względnie wyższe dla mowy niż dla szumu w obszarach mowy - związane jest to ze strukturą formantu. Dlatego $\sigma(n) + \mu(n)$ jest na ogół wyższe w obszarach mowy i niższe w regionach pozbawionych mowy. Ponieważ oczekuje się, że rozpiętość szumu (po kompensacji) będzie niższa, zaobserwowano, że wartości $\sigma(n) - \mu(n)$ są zwykle niższe w obszarach pozbawionych mowy, w porównaniu do obszarów zawierających mowę (rys. 2(e)). Pomnożenie (sigma(n) + u(n)) przez (sigma(n)+ u(n)) daje ($sigma^2(n) - u^2(n)$), co podkreśla kontrast pomiędzy obszarami zawierającymi mowę i tymi, które mowy nie zawierają.

W związku z dużą rozpiętością tonalną wartości ($sigma^2(n) - u^2(n)$), ciężko jest zaobserwować obszary mowy z małymi wartościami ($sigma^2(n) - u^2(n)$). Aby podkreślić kontrast pomiędzy obszarami mowy i obszarami niezawierającymi mowy, rozpiętość tonalna jest redukowana poprzez obliczenie

$$\delta(n) = \sqrt[M]{|(\sigma^2(n) - \mu^2(n))|},$$

gdzie M zostało wybrane jako 64

*** tutaj rysunek Rys.3.1 z przykladami $\mu(n)$, $\sigma(n)$, $\sigma(n) - \mu(n)$, $\delta(n)$ ze znakiem, $\delta(n)$

Wartość M nie jest decydująca. Każda wartość M z przedziału 32-256 wydaje się być dobra, aby zapewnić dobry kontrast pomiędzy obszarami zawierającymi mowę, a tymi, które mowy nie zawierają na wykresie $\delta(n)$. W obliczeniach $\delta(n)$ brana jest po uwagę tylko wartość bezwzględna wartości chwilowej ($\sigma^2(n) - \mu^2(n)$). Jeżeli znak wyrażenia ($\sigma^2(n) - \mu^2(n)$). jest przypisany(?) do delta(n), wartości będą wahać się w okolicach zera w obszarach pozbawionych mowy dla większości typów szumów , ale krótki czas (20-40 msec) tymczasowych średnich wartości będzie mały i będzie się wahał, sprawiając, że cecha szumu będzie nierówna. To powoduje trudności w ustaleniu progu detekcji dla obszarów pozbawionych mowy. Wartości $\delta(n)$ będą miały wysoką średnią w obszarach pozbawionych mowy z małą średnią wariancją. Pomoże to w ustaleniu odpowiedniego progu do odizolowania obszarów pozbawionych mowy od tych, które mowę zawierają. Zakres $\delta(n)$ ze znakiem rys. 3.1(f)) jest inny niż wartości $\delta(n)$ (rys. 3.1(g)). Mały tymczasowy obszar wartości $\delta(n)$ w obszarach niezawierających mowy i jego średnia wartość pomagają w dobraniu pasującego progu. Wartości $\delta(n)$ w obszarach niezawierających mowy są podyktowane poziomem szumu. Zauważmy, że rozważając wartości $\delta(n)$ bez znaku, tracimy trochę zalet w rozróżnialności obszarów niezawierających mowy, które maja zarówno dodatnie, jak i ujemne wartości - natomiast obszary zawierające mowę mają w większości dodatnie wartości. Wartości $\delta(n)$ z $M = 64$ są wykorzystywane do dalszego przetwarzania do podejmowania decyzji. Warto zauważyć zmiany w przeskalowaniu na rys 3.1(f) i 3.1(g), aby zrozumieć istotę używania wartości bezwzględnej, np $\delta(n)$ bez znaku.

3.3.3 Logika podejmowania decyzji

Logika podejmowanej decyzji opiera się o $\delta(n)$ dla każdego wyrażenia poprzez wyprowadzenie najpierw progu detekcji z przyjętego z założenia obszaru zawierającego szum, a później zastosować ten próg na tymczasowo wygładzonych wartosciach $\delta(n)$. Rozmiar okna l_w wykorzystany do wygładzenia $\delta(n)$ jest zaadoptowany w oparciu o estymatę rozpiętości tonalnej (ρ) energii zaszumionego sygnału dla każdego wyrażenia, zakładając, że jest przynajmniej 20% obszarów zawierających ciszę w każdym wyrażeniu. Binarna decyzja odnośnie mowy i jej braku w każdej chwili czasowej, oznaczana odpowiednio jako 1 i 0, jest dalej wygładzana z wykorzystaniem okna adaptacyjnego, żeby dotrzeć do ostatecznej decyzji detekcji. Następujące 5 kroków opisuje implementację szczegółów w logice podejmowania decyzji:

1) Obliczenie progu (θ):

Obliczyć średnią (μ_θ) i wariancję (σ_θ) dla 20% najmniejszych wartości.

Próg $\theta = \mu_\theta + 3\sigma_\theta$ jest używany we wszystkich przypadkach. Wartość θ zależy od analizownego wyrażenia. Zatem wartość progu, odpowiadająca wartości cechy z $\theta(n)$, jest adaptopwana do konkretnego wyrażenia w zależności od charakterystyki sygnału i szumu w tym wyrażeniu.

2) Wyznaczenie okna wygładzającego l_w:

Energia E_m sygnału $x(n)$ jest obliczana dla ramki 300msec z przesunieciem 10msec, gdzie m to numer ramki. Rozpiętość tonalna (ρ) sygnału jest obliczana jako:

$$\rho = 10 \log_{10} \frac{max_m(Em)}{min_m(Em)}).$$

Parametr opisujący długość okna l_w do wygładzenia sygnału jest uzyskiwany z rozpiętości tonalnej (ρ) sygnału. Wartości ρ różnią się dla różnych szumów przy tym samym SNR, ponieważ charakterystyki szumów się różnią. Wskaźnik SNR dla mowy z dystansu zależy od warunków środowiskowych oraz od odległości, z jakiej mówca mówi do mikrofonu. Zaobserwowano, że wartości ro dla mowy z odległości są rozciągnięte w porównaniu z wartościami ro dla różnych szumów. Jest to głównie spowodowane efektem echa. Rozkład wartości ro zależy również od odległości mówcy od mikrofonu. Wartość ρ dla każdego wyrażenia jest wykorzystywana do określenia wartości niektórych parametrów do dalszego przetwarzania $\delta(n)$ i do otrzymania decyzji o klasyfikacji. W przypadkach, gdzie $\delta(n)$ reprezentuje charakterystyka dyskryminacyjna przedstawiająca zarówno mowę, jak i jej brak, odpowiadające wartości ρ są wysokie, jak zaobserwowano w przypadku szumów volvo, lamparta i karabinu maszynowego. W takich przypadkach używane są małe wartości parametru l_w okna wygładzającego. Następujące wartości l_w zostały wybrane na drodze przeprowadzonych doświadczeń z sygnałem mowy zaszumionym przez różne typy szumów z różnymi poziomami SNR:

$$l_w = 400msec, \quad dla \ \ \rho < 30$$

$$l_w = 300msec, \quad dla \ \ 30 \leqslant \rho < 40$$

$$l_w = 200msec, \quad dla \ \ \rho > 40.$$

3) Logika podejmowania decyzji w każdej chwili czasowej:

Wartości $\delta(n)$ są uśredniane przez okno o rozmiarze l_w, aby otrzymać uśrednione wartości $\bar{\delta}(n)$ w każdej próbce o indeksie n. Decyzja jest podejmowana według następujących zależności:

$$d(n) = 1, \quad dla \ \bar{\delta}(n) > \theta$$

$$d(n) = 0, \quad dla \ \bar{\delta}(n) \leqslant \theta.$$

4) Wygładzenie decyzji na poziomie próbek:

Decyzja $d(n)$ dla każdej próbki jest okienkowana o rozmiarach okienek 300msec, 400msec, 600msec, dla, odpowiednio, $\rho < 30$, $30 \leqslant \rho < 40$, $\rho > 40$. Załóżmy, że η jest progiem (w procentach) w zależności od wartości $d(n)$, które dają 1 w okienku. Jeżeli wartość procentowa wartości $d(n)$, które wynoszą 1 w okienku, jest wyższa niż wartość η, wtedy ostateczna decyzja $d_f(n)$ jest ustawiana na 1 w chwili czasowej n, w przeciwnym wypadku - 0. Wartość przypisana dla η to 60%.

5) Decyzja na poziomie ramek:

Decyzja w metodzie AMR jest podejmowana dla każdej ramki, co 10msec. W celu porównania zaproponowanej metody z metodą AMR, decyzja $d_f(n)$ jest konwertowana na 10msec ramkę w oparciu o podjętą decyzję. Dla każdej 10msec, nienakładającej się ramki, jeżeli przeważająca ilość decyzji $d_f(n)$ wynosi 1, to cała ramka jest oznaczana jako zawierająca mowę, w przeciwnym wypadku jest oznaczana jako niezawierająca mowy. Informacje dotyczące sygnałów mowy pozyskane z empirycznie, są rownież otrzymywane z każdej 10msec ramki.

3.4 Zmieniony algorytm Single Frequency Filtering

W celu zmniejszenia złożoności obliczeniowej oraz zwiększenia dokładności detekcji, zostały zaproponowane pewne zmiany w algorytmie SFF:

1) Zmiana w sposobie obliczania obwiedni sygnału. Wykorzystana rekursywna filtracja kwadraturowa, poniżej fragment kodu przedstawiający liczenie obwiedni.

```
vector<SampleType> wavEnvelope;
const double pi = acos(-1);

double omega = 2*pi*normalizedFrequency/_samplingFrequency;
double module = 0.97;

///filter factors
double a1 = module * cos(omega);
double a2 = module * sin(omega);

///filter initialization
vector<double> XReal(wavDifferenced.getSamplesCount(), 0);
vector<double> XImaginary(wavDifferenced.getSamplesCount(), 0);

///recursive square filter
for (int i = 1; i < wavDifferenced.getSamplesCount(); i++) {
        XReal[i]=wavDifferenced.sample(i)+a1*XReal.at(i-1)-a2*XImaginary.at(i-1);
        XImaginary[i]=a1*XImaginary.at(i-1)+a2*XReal.at(i-1);
}

///envelope
for (int j = 0; j < wavDifferenced.getSamplesCount(); j++) {
        wavEnvelope.push_back(sqrt(XReal.at(j)*XReal.at(j)*XImaginary.at(j)*XImaginary.at(j)));
}
```

2) Metoda SFF zakłada obliczenie 185 obwiedni począwszy od 300Hz do 4000Hz co 20Hz. Doświadczalnie sprawdzono, że wystarczające jest obliczenie co czwartej obwiedni zaczynając od 300Hz, oraz zaczynając od 340Hz, co 20Hz każda. Zatem, zamiast 185 obwiedni, liczonych jest jedynie 92.

3) Zmiana w sposobie liczenia progu detekcji. Liczone jest 15% z najmniejszych wartości. Tworzony jest również histogram dla $\delta(n)$, na podstawie którego zostaje wyliczonych 15% najmniejszych wartości. Znacząco zwiększa to szybkość wykonywanych obliczeń.

//czy tutaj przedstawić proces generowania histogramu i sposób wyznaczania progu?

Rozdział 4

Implementacja programu

Do zaimplementowania algorytmów skorzystano z biblioteki Aquila, która wspiera operacje na sygnałach oraz z standardowych bibliotek C++, takich jak cmath, stl.

Rozdział 5

Wyniki badań

5.1 Sposób oceny

Do przeprowadzenia badań zostało wybranych po 2 sygnały (głos damski oraz głos męski) zawierających pojedyncze słowa oraz po 2 sygnały zawierające ciągi słów. Pierwszy etap badań założył manualne wybranie próbek zawierających początki oraz końce wypowiedzi, rozważone zostały dwa przypadki, gdy: 1. mowa rozpoczyna się w momencie nabrania powietrza przez mówcę i aktywację strun głosowych, 2. mowa rozpoczyna się w momencie generowania słyszalnych dźwięków przez mówcę. Następnie te same sygnały zostały zbadane przez 3 algorytmy detekcji, które również wybrały numery próbek, które miały reprezentować początki i końce mowy. Dla każdego pomiaru wykonano 10 prób i ostateczne wartości poddane badaniom są średnią wartością z tych prób.

Następnie została policzona wartość bezwzględna próbki wybranej manualnie oraz próbki wybranej przez algorytm. Wyniki dla poszczególnych algorytmów, w dwóch rozważanych przypadkach zostały przedstawione w tabeli poniżej.

5.2 Wyniki dla pojedynczych słów

5.3 Wyniki dla ciągów słów

Rozdział 6

Podsumowanie

Bibliografia

Spis rysunków

Spis tabel