POLITECHNIKA GDAŃSKA
Wydział Elektroniki, Telekomunikacji i Informatyki

Adam Kupryjanow

Metoda i algorytmy modyfikacji sygnału do celu wspomagania rozumienia mowy przez osoby z pogorszoną rozdzielczością czasową słuchu

Rozprawa doktorska

Promotor:
prof. dr hab. inż. Andrzej Czyżewski,
prof. zw. Politechniki Gdańskiej
Wydział Elektroniki, Telekomunikacji i Informatyki
Politechnika Gdańska

Gdańsk 2012

Spis treści

Oznaczenia i skróty

symbol	znaczenie
Δt_{vowel}	czas trwania samogłosek w przedziale czasu Δt
$\mu(\text{ROS})_{\text{fast}}$	wartość średnia tempa mowy szybkiej
$\mu(\text{ROS})_{\text{slow}}$	wartość średnia tempa mowy wolnej
$\mu(W[m])_{szum}$	zaktualizowana wartość średnia parametru odpowiadająca sygnałowi szumowemu
bStart	dolna granica indeksu b wykorzystywana podczas wyznaczania E_m^b
bStop	górna granica indeksu b wykorzystywana podczas wyznaczania E_m^b
C	wartość współczynnika wykorzystywanego w inicjalizacji progu w algorytmie VAD
$c[j]$	j-ty współczynnik MFCC
$c_N[j]$	wartość j-tego uśrednionego współczynnika MFCC wyznaczonego dla szumu
$D(t)$	funkcja odkształcająca czas (ang. *Time-Scale Warping Function*)
$d(t)$	różnica obwiedni amplitudowej wyznaczona w niskich i wysokich pasmach krytycznych
d_t	zakres czasu, wewnątrz którego wyszukiwany jest obszar samogłosek w algorytmie detekcji samogłosek wykorzystującym zmodyfikowaną obwiednię amplitudową
$e[n]$	sygnał pobudzający trakt głosowy
$e_h[n]$	transformata Hilberta sygnału $e[n]$.
E_{LF}^m	energia sygnału w niskich pasmach melowych (100–1000 Hz)
E_m	energia m-tej ramki sygnału $x[n]$
E_m^b	energia sygnału wyznaczona dla sygnału $x_b[n]$
E_m^{bf}	energia widma amplitudowego sygnału wyznaczona w podpasmach
E_m^f	energia widma amplitudowego m-tej ramki sygnału
E_m^j	wartość energii sygnału w j-tym filtrze melowym
f	wartość częstotliwości wyrażona w skali liniowej w Hz

Fs	szybkość próbkowania
$h[n]$	obwiednia Hilberta
$H_c^m(p_m)$	entropia widmowa w m-tej ramce sygnału
j	numer współczynnika MFCC
K	liczba prążków widma
k	numer prążka w widmie DFT
$K(\mathbf{x},\mathbf{z})$	funkcja jądra
k_m	miejsce położenia maksymalnej wartość funkcji korelacji skrośnej w m-tej ramce analizy
k_{max}	górna granica przedziału określającego obszar analizy zmienności funkcji korelacji wzajemnej
k_{min}	dolna granica przedziału określającego obszar analizy zmienności funkcji korelacji wzajemnej
L	długość ramki sygnału w próbkach
l	numer próbki przebiegu korelacji wzajemnej
L_k	długości zachodzących na siebie przedziałów w kroku syntezy w algorytmie TSM
m	numer ramki sygnału
$M(f)$	funkcja mapująca częstotliwości do skali melowej
metoda R	metoda modyfikacji czasu trwania sygnału opracowana przez Nejime *et al.*
m_{szum}	numer aktualnej ramki szumowej
n	numer próbki sygnału
N	liczba próbek sygnału
n^b_{cisza}	liczba ramek sygnału błędnie oznaczonych przez algorytm VAD jako cisza
n^b_{mowa}	liczba ramek sygnału błędnie oznaczonych przez algorytm VAD jako mowa
n^b_{sam}	liczba samogłosek wykrytych w miejscu spółgłosek
$n^b_{\text{spół}}$	liczba niewykrytych samogłosek
n^d_{cisza}	liczba ramek sygnału oznaczonych przez algorytm VAD jako cisza
n^d_{mowa}	liczba ramek sygnału oznaczonych przez algorytm VAD jako

	mowa
N_m	ang. *modified loudness*
$N_m(t)$	wygładzona obwiednia amplitudowa
n^o_{cisza}	liczba ramek sygnału oznaczonych ręcznie jako cisza
n^o_{mowa}	liczba ramek sygnału oznaczonych ręcznie jako mowa
n^o_{sam}	całkowitą liczbę samogłosek w analizowanym zbiorze
$N_v(t)$	obwiednia amplitudowa *v*-tego pasma krytycznego
n_{vr}	liczba wystąpień VR w przedziale czasu Δt
O	współczynnik kosztu (ang. *cost*)
o_t	wartość opóźnienia pojawiającego się pomiędzy sygnałem wejściowym a sygnałem spowolnionym
P	okres podstawowy sygnału
p	poziom istotności statystycznej
p_i	założony poziom istotności statystycznej
P_m	moc *m*-tej ramki sygnału $x[n]$
p_m^k	prawdopodobieństwo wystąpienia *k*-tego prążka w widmie *m*-tej ramki sygnału
PR_{det}	algorytm VRD bazujący na analizie parametru PR
PR_{p1}	wartość pierwszego progu w algorytmu VRD opartego na analizę wartości PR
PR_{p2}	wartość drugiego progu w algorytmu VRD opartego na analizę wartości PR
PR_{p3}	wartość trzeciego progu w algorytmu VRD opartego na analizę wartości PR
PVD_{det}	algorytm VRD bazujący na analizie parametru PVD
REC_{det}	algorytm VRD bazujący na analizie parametru REC
$R_m[k]$	funkcja korelacji wzajemnej wyznaczoną dla *m*-tej ramki analizy
$ROS[m]$	wartość tempa mowy wyznaczona dla *m*-tej ramki analizy
ROS_o	oczekiwana wartość tempa mowy spowolnionej
$ROS^{PP}_{\text{szybkie}}$	średnia wartość tempa mowy wypowiadanej w tempie szybkim wykorzystywanej w teście PPTM
$ROS^{PP}_{\text{średnie}}$	średnia wartość tempa mowy wypowiadanej w tempie średnim wykorzystywanej w teście PPTM

ROS^{PP}_{wolne}	średnia wartość tempa mowy wypowiadanej w tempie wolnym wykorzystywanej w teście PPTM
ROS_{PR}	algorytm estymacji tempa mowy oparty na analizie parametru PR
$ROS^{P}_{szybkie}$	średnia wartość tempa mowy wypowiadanej w tempie szybkim wykorzystywanej w teście PTM
$ROS^{P}_{średnie}$	średnia wartość tempa mowy wypowiadanej w tempie średnim wykorzystywanej w teście PTM
ROS_{PVD}	algorytm estymacji tempa mowy oparty na analizie parametru PVD
ROS^{P}_{wolne}	średnia wartość tempa mowy wypowiadanej w tempie wolnym wykorzystywanej w teście PTM
ROS_{REC}	algorytm estymacji tempa mowy oparty na analizie parametru REC
ROS_{th}	próg dzielący tempo mowy na szybkie i wolne
Sa	wartości przesunięcia ramki sygnału w kroku analizy
Ss	wartości przesunięcia ramki sygnału w kroku syntezy
$t_a[m]$	punkt analizy m-tej ramki w algorytmie PSOLA
t_p	czas potrzebny do przetworzenia określonego fragmentu sygnału
T_r	próg stosowany w algorytmie detekcji samogłosek wykorzystującym zmodyfikowaną obwiednię amplitudową
$t_s[m]$	punkt syntezy m-tej ramki w algorytmie PSOLA
t_{wej}	czas trwania nagrania oryginalnego
$t_{zdarzenia}$	czas trwania przetwarzanego fragmentu sygnału
U	rząd analizy LPC
v	numer pasma krytycznego
V_1	parametr opisujący zmienność cepstrum
V_2	parametr opisujący zmienność cepstrum
V_{2N}	parametr opisujący zmienność cepstrum
VAD_{czysty}	opracowany algorytm detekcji mowy niewykorzystujący wygładzania decyzji
VAD_p	wartość progu wykorzystywanego w algorytmie VAD
$VAD_{wygładzony}$	opracowany algorytm detekcji mowy wykorzystujący wygładzanie decyzji

$W[m]$	wartość parametru użytego w procesie detekcji mowy wyznaczona dla m-tej ramki analizy
$w[n]$	funkcja okna
$wh[j]$	waga j-tego współczynnika MFCC
W_{noise}	wartość parametru użytego w procesie detekcji mowy wyznaczona dla ostatniej ramki szumowej
$W_{\text{noise-1}}$	wartości parametru użytego w procesie detekcji mowy dla ramki zawierającej szum i znajdującej się jedną ramkę wcześniejszej od obecnej ramki zawierającej szum
$W_{\text{noise-2}}$	wartości parametru użytego w procesie detekcji mowy dla ramki zawierającej szum i znajdującej się dwie ramki wcześniejszej od obecnej ramki zawierającej szum
W^n_{th}	nowa wartość progu W_{th}
W_{th}	wartość progu wykorzystywana w procesie detekcji mowy
$X(j,k)$	widmo odpowiadające j-temu filtrowi
$X(k)$	widmo amplitudowe sygnału $x[n]$ wyznaczone dla ramki o długości L
$x[n]$	sygnał wejściowy
$x_b[n]$	sygnał przefiltrowanym za pomocą filtra pasmowoprzepustowego
y[n]	sygnał wyjściowy
$Z(k)$	widmo amplitudowe sygnału $z[n]$
$z[n]$	sygnał jednopołówkowy – ang. *half-wave rectified*
ZCR_m	liczba przejść przez zero sygnału w m-tej ramce
α	współczynnik skali
$\alpha(\tau)$	wartość współczynnika skali w chwili τ
α_{brutto}	stosunek czasu trwania nagrania zmodyfikowanego do czasu trwania nagrania wejściowego
α_{cons}	współczynnik skali stosowany dla m-tej ramki sygnału w przypadku wykrycia spółgłoski
α_{de}	współczynnik oczekiwanej skali
α_{ne}	współczynnik naturalnej zmiany skali
α_{netto}	Stosunek czasu trwania sygnału mowy w nagraniu

	zmodyfikowanym do czasu trwania mowy w sygnale oryginalnym
α_o	wartość współczynnika skali ustawiana przez użytkownika
α_{vowel}	współczynnik skali stosowany dla *m*-tej ramki sygnału w przypadku wykrycia samogłoski
β_1	górna granica częstotliwości wykorzystywana podczas wyznaczania wartości p_m^k
β_2	dolna granica częstotliwości wykorzystywana podczas wyznaczania wartości p_m^k
Δt	przedział czasu, dla którego zliczana jest liczba wystąpień VR
Δt_{vowel}	czas trwania samogłosek w przedziale czasu Δt
ζ_i	parametr „zwisu” (ang. *slack variable*)
η	współczynnik określający relację pomiędzy wartościami współczynników skali wykorzystywanych do spowalniania samogłosek i spółgłosek
$\sigma^2(W[m])_{szum}$	zaktualizowana wartość wariancji parametru odpowiadająca sygnałowi szumowemu
σ^2_{nowa}	nowa wartość wariancji szumu
σ^2_{stara}	stara wartość wariancji szumu
$\phi(\mathbf{x})$	funkcja mapująca dane z przestrzeni R^n do przestrzeni R^m
$\chi^2(\cdot)_{cv}$	wartość krytyczna statystyki testu Friedmana

skrót	znaczenie
(C)APD	ośrodkowe zaburzenia słuchu – ang. (*Central*) *Auditory Processing Disorders*
ACR	ang. *Absolute Category Rating*
ALI	system automatycznego rozpoznawania języka – ang. *Automatic Language Identification*
ANOVA	analiza wariancji – ang. *Analysis of Variance*
ANR	sztuczna sieć neuronowa – ang. *Artificial Neural Network*
AOLA	ang. *Adaptive Overlap-Add*
ASHA	Amerykańskie Stowarzyszenie Słuchu i Mowy – ang. *American Speech-Language Hearing Assiociation*
ASR	system automatycznego rozpoznawania mowy – ang. *Automatic Speech Recognition*
b.d.	brak danych
CLT	centralne twierdzenie graniczne – ang. *Central Limit Theorem*
CNG	generator szumu komfortowego – ang. *Comfort Noise Generation*
DCR	ang. *Degradation Category Rating*
DMOS	ang. *Degradation Mean Opinion Score*
enrate	ang. *energy rate*
FBD	ang. *Forward-Backward Divergence*
FM	ang. *Frequency Modulation*
HMM	ukryte modele Markova – ang. *Hidden Markov Model*
HOS	statystyki wyższego rzędu – ang. *High Order Statistics*
HR0	ang. *Non-speech hit ratio*
HR1	ang. *Speech hit ratio*
KSM	Katedra Systemów Multimedialnych
LLI	zaburzenia w nauce języka – ang. *Language Learning Impairement*
LPC	liniowe kodowanie predykcyjne – ang. *Linear Predicitve Coding*
LRT	test ilorazu wiarygodności – ang. *Likelihood Ratio Test*
LSD	ang. *Least Significant Difference*
MD	„średnia delta” – ang. *Mean Delta*
MFCC	współczynniki mel-cepstralne – ang. *Mel-frequency Cepstral Coefficents*

MOS	ang. *Mean Opition Score*
mrate	ang. *multiple rate estimator*
MSE	błąd średniokwadratowy – ang. *Mean Squere Error*
OLA	ang. *Overlap and Add*
PDF	funkcja gęstości prawdopodobieństwa – ang. *Probability Density Function*
PPS	liczba głosek na sekundę – ang. *Phones Per Second*
PPTM	Polski Pediatryczny Test Macierzowy
PR	parametr, którego nazwa powstała od pierwszych liter skrótów PVD + REC
PRM	poprawa rozumienia mowy
PSM	modyfikacja wysokości – ang. *Pitch Scale Modification*
PSOLA	ang. *Pitch Synchronous Overlap and Add*
PTM	Polski Test Macierzowy
PVD	ang. *Peak Valley Difference*
REC	ang. *Reduced Energy Cumulating*
RGDT	test detekcji przerw losowych – ang. *Random Gap Detection Test*
RM ANOVA	Test ANOVA z powtórzeniami – ang. *Repeated Measures Analysis of Variance*
RMS	wartość skuteczna – ang. *Root Mean Square*
ROC	charakterystyka odbiornika – ang. *Receiver Operating Characteristics*
ROS	tempo mowy – ang. *Rate of Speech*
RTF	ang. *Real-Time Factor*
SACF	funkcja autokorelacji widma – ang. *Spectrum Autocorrelation Function*
SAPVR	ang. *Spectral Autocorrelation Peak Valley Ratio*
SBEC	ang. *Spectral Band Energy Cumulating*
SNR	stosunek sygnału do szumu – ang. *Signal to Noise Ratio*
SOLA	ang. *Synchronous Overlap and Add*
SPS	liczba sylab na sekundę – ang. *Syllables Per Second*
SRT_{50}	ang. *Speech Reception Threshold*
SVM	metoda wektorów nośnych – ang. *Support Vector Machine*
TCST	test rozumienia mowy przyspieszonej – ang. *Time Compressed Speech Test*

TCT_{50}	ang. *50% Time-Compressed Speech Threshold*
TDHS	ang. *Time-Domain Harmonic Scaling*
TD-PSOLA	ang. *Time-Domain PSOLA*
TEO	ang. *Teager Energy Operator*
TRMS	Test rozumienia mowy spowolnionej
TSM	modyfikacja czasu trwania – ang. *Time Scale Modification*
TTS	ang. *Text To Speech*
VAD	detekcja aktywności głosowej – ang. *Voice Activity Detection*
VD	detekcja samogłosek – ang. *Vowel Detection*
VER	wskaźnik liczby błędów występujących podczas detekcji samogłosek – ang. *Vowel Error Rate*
VLD	detekcja obszarów samogłosek – ang. *Vowel Landmark Detection*
VM	model samogłoski – ang. *Vowel Model*
VoIP	telefonia internetow – ang. *Voice over Internet Protocol*
VOP	detekcja początku samogłoski – ang. *Vowel Onset Point*
VPS	liczba samogłosek na sekundę – ang. *Vowels per Second*
VR	przedział samogłoski – ang. *Vowel Region*
VRD	segmentacja przedziału samogłoski – ang. *Vowel Region Detector*
WER	wskaźnik liczby błędów słownych – ang. *Word Error Rate*
WPS	liczba słów na sekundę – ang. *Words Per Second*
WSOLA	ang. *Wave Similarity Overlap and Add*
ZCR	liczba przejść przez zero – ang. *Zero Crossing Rate*

1 Wprowadzenie

Problemy w rozumieniu mowy mogą powodować zaburzenia w procesie rozwoju intelektualnego oraz procesu uczenia się dzieci. Często trudności w rozumieniu mowy związane są z głuchotą obwodową. Zastosowanie tradycyjnego aparatu słuchowego, działającego na zasadzie kompandera, pozwala na niemal całkowite wyeliminowanie problemu. Nie zawsze jednak rozumienie mowy musi łączyć się z głuchotą obwodową. W swoich pracach Chermak i Musiek szacują, iż od 10% do 20 % osób powyżej 65 roku życia cierpi na ośrodkowe zaburzenia słuchu (ang. (*Central*) *Auditory Processing Disorders* – (C)APD) [15]. W grupie dzieci w wieku od 6 do 10 lat, tego typu zaburzeniami dotknięte jest od 2% do 5% populacji [19]. Na przestrzeni ostatnich 20 lat powstało wiele definicji (C)APD [6] [3] [62] [11]. Jedną z najbardziej popularnych jest ta opracowana w 1996 (i uaktualniona w 2005 roku) przez Amerykańskie Stowarzyszenie Słuchu i Mowy (ang. *American Speech-Language Hearing Assiociation* – ASHA). Według tej definicji (C)APD to zaburzenie charakteryzujące się niedostatecznymi wynikami w jednej lub wielu funkcjach słuchowych [6]. W definicji ujęto następujące mechanizmy słuchowe będące podstawą takich zdolności i umiejętności słuchowych jak:

- lokalizacja źródła dźwięku,
- rozróżnianie cech dźwięku,
- czasowe aspekty słyszenia takie jak: rozdzielczość czasowa, maskowanie, integracja, porządkowanie czasowe,
- zdolność rozumienia sygnałów mowy przy występowaniu sygnałów „konkurencyjnych”,
- zdolność rozumieniu sygnałów zniekształconych.

Typowymi symptomami (C)APD są m.in. problemy z rozumieniem mowy w trudnych warunkach akustycznych, trudności w koncentracji, trudności w odbiorze szybko wypowiadanej mowy, trudności w nauce mówienia i czytania. Objawy zaburzeń słuchu związanych z centralnym układem nerwowym mogą być podobne do tych wynikających z ubytku słuchu polegającego na uszkodzeniu układu obwodowego (tzw. głuchota). Jednak sposób wspomagania procesu słyszenia, jak i proces diagnozy obu schorzeń jest, różny. Diagnoza obwodowych zaburzeń słuchu polega na wykonaniu np. audiometrii tonalnej i w

sytuacji wykrycia zaburzenia zastosowaniu aparatu słuchowego, który ma za zadanie dokonać korekcji amplitudy sygnału w pasmach częstotliwości w taki sposób, by skompensować ubytki wykryte w charakterystyce częstotliwościowej słuchu. Zazwyczaj w tym celu stosuje się wielopasmowe kompandery oraz algorytmy redukujące sprzężenia sygnału.

Diagnostyka (C)APD jest złożona, ponieważ musi ona zawierać testy oceniające wszystkie funkcje słuchowe określone w definicji ASHA (ang. *American Speech-Language Hearing Assiociation*) [98]. Dodatkowo zaburzenia te mogą występować wspólnie z zaburzeniami obwodowymi, co może prowadzić do błędnej klasyfikacji (C)APD jako wyłącznie zaburzeń obwodowych. Niewykryte u dzieci (C)APD, mogą prowadzić do sytuacji, w której będą one miały problemy m.in. w nauce i rozwoju mowy. W ostatnich latach opracowano szereg testów pozwalających na wczesne wykrycie tego typu schorzeń zarówno u dzieci, jak i u osób dorosłych [66] [128] [46] [34] [162] [181] [196].

Zalecenia dotyczące postępowania przy rozpoznaniu (C)APD odnoszą się do strategii związanych z poprawą środowiska akustycznego, sposobami kompensacji zaburzeń oraz metodami treningu. Najpopularniejszymi metodami kompensacji zaburzeń oraz poprawy środowiska akustycznego są tzw. systemy FM (ang. *Frequency Modulation*) [144] [185]. Pod tą nazwą rozumie się układ składający się z bezprzewodowego mikrofonu transmitujący sygnał do bezprzewodowego odbiornika połączonego ze słuchawkami. Takie systemy najczęściej stosuje się podczas zajęć lekcyjnych lub w salach seminaryjnych. Pozwalają one na eliminację zakłóceń pochodzących od otoczenia (np. gwaru) i odbić występujących wewnątrz pomieszczenia, dzięki czemu zwiększa się odstęp sygnału od szumu (ang. *Signal to Noise Ratio* – SNR). Jest to osiągane dzięki bezpośredniej transmisji mowy rejestrowanej przez mikrofon umieszczony blisko ust mówcy, do obuusznych słuchawek noszonych przez osobę używającą systemu.

Innego rodzaju rozwiązaniami są metody oparte na algorytmach modyfikacji czasu trwania sygnału (ang. *Time Scale Modification* – TSM). Zasada ich działania polega na wydłużeniu czasu trwania mowy. Metody te bazują na założeniu, iż dodatkowy czas, uzyskany poprzez spowolnienie wypowiedzi, pozwala osobom z (C)APD, na dokładniejsze przyswojenie informacji do nich docierających. Prowadzi to do poprawy rozumienia mowy. Dwa główne rozwiązania pozwalające na spowolnienie sygnału mowy w czasie rzeczywistym, znane z literatury, zostały opracowane przez Nakamura *et al.* [117]

oraz Nejime *et al.* [121]. Czas rzeczywisty w tym przypadku i w odniesieniu do jego wzmiankowania również w dalszej części tej pracy jest pojęciem przyjętym umownie, bowiem stosowanie algorytmów TSM mowy zawsze jest związane z mniejszym lub większym opóźnieniem sygnału, które niekiedy jest na tyle duże, że bywa zauważalne słuchowo. Czas rzeczywisty jest w tym przypadku zatem rozumiany jako działanie na bieżącym sygnale odbieranym poprzez mikrofon dla odróżnienia od przypadku przetwarzania zarejestrowanych nagrań. Oba wspomniane wcześniej urządzenia opracowano z myślą o wspieraniu osób powyżej 65 roku życia. Taki dobór grupy docelowej związany był z założeniem, iż osoby te często mają problemy ze rozumieniem szybko wypowiadanej mowy. Pierwsze urządzenie ma za zadanie redukować tempo mowy odtwarzanej przez odbiornik telewizyjny. Prędkość wypowiedzi dobierana jest tu przez użytkownika systemu, a sygnał jest spowalniany w sposób ciągły z wykorzystaniem metod nierównomiernej TSM (ang. *Time Scale Modification*) sygnału. Drugie urządzanie jest rozwiązaniem przenośnym i zostało stworzone w celu wspierania osób starszych podczas prowadzenia rozmowy. Użytkownik za pomocą przycisku ma możliwość włączenia/wyłączenia spowalniania sygnału. Takie rozwiązanie pozwala na spowalnianie tylko istotnych dla słuchacza fragmentów sygnału oraz zapewnia mechanizm zabezpieczający urządzenie przed wprowadzeniem zbyt dużego przesunięcia czasowego pomiędzy sygnałem wejściowym a sygnałem wyjściowym[1]. Oba urządzenia zostały poddane testom z udziałem grupy osób powyżej 65 roku życia. Jednak Nakamura *et al.* zbadali jedynie subiektywną ocenę „łatwości" rozumienia mowy oraz relatywną jakość mowy spowolnionej. Natomiast Nejime *et al.* dokonali analizy skuteczności rozpoznawania mowy spowolnionej w zależności od zaburzeń rozdzielczości czasowej słuchu. Nie udało im się wyznaczyć zależności pomiędzy zaburzeniami a stopniem poprawy rozumienia mowy [122].

Ostatnią grupą rozwiązań są systemy treningu przeznaczone do wspierania procesu nauki mówienia i czytania (ang. *Language Learning Impairement* – LLI). Za pionierskie uznaje się tu badania Tallala *et al.* [172] [115]. W swoich pracach proponują oni trening polegający na odtwarzaniu pacjentom mowy o wydłużonym czasie trwania oraz o zwiększonej amplitudzie sygnału (w paśmie 3–30 Hz). Badania pokazały, iż czterotygodniowy trening z wykorzystaniem tej metody pozwala na poprawę

[1] Stale rosnące opóźnienie pojawiłoby się w sytuacji ciągłego wydłużania czasu trwania sygnału wejściowego

rozpoznawania mowy, zdolności przetwarzania językowego oraz rozumienia gramatycznego średnio o 1–2 lat w skali równoważnego wieku językowego.

Jak wynika z przedstawionego powyżej przeglądu rozwiązań przeznaczonych osobom z (C)APD, nie opracowano do tej pory rozwiązania uniwersalnego, czyli takiego, które byłoby w stanie wspierać proces rozumienia mowy zarówno u dzieci jak i osób starszych. Ponadto konieczne jest także opracowanie metod działających w pełni automatycznie i niezależnie od warunków akustycznych. Stąd motywacja do opracowania metody modyfikującej czas trwania sygnału mowy działającej w czasie rzeczywistym, która mogłaby zostać zastosowana w wielu scenariuszach użycia (m.in. tych opisanych powyżej). W ramach rozprawy opracowano metodę bazującą na założeniach Nakamura *et al.* oraz Nejime *et al.*. Grupa docelowa, dla której rozwiązanie to zostało opracowane, jest jednak szersza i zawiera osoby z (C)APD, u których występuje pogorszona rozdzielczość czasowa słuchu. Jednym z celów cząstkowych rozprawy było wyznaczenie relacji pomiędzy wpływem spowalniania sygnału mowy na stopień jej rozumienia, a poziomem zaburzeń rozdzielczości słuchu. Osiągnięcie tego celu pozwoli na określenie tego czy dla danego pacjenta opracowana metoda modyfikacji sygnału przyniesie wymierne skutki w postaci poprawy rozumienia mowy.

Nowością w opracowanej metodzie jest wprowadzenie relacji pomiędzy stopniem spowolnienia sygnału mowy a tempem mowy wejściowej. Dzięki tej zależności fragmenty mowy wolnej spowalniane są z wykorzystaniem innych wartości współczynników skali niż fragmenty mowy szybkiej. W celu umożliwienia ciągłej pracy algorytmu, zastosowano szereg rozwiązań pozwalających na utrzymywanie synchronizacji sygnału wejściowego z sygnałem o zmodyfikowanej strukturze czasowej. Mechanizmy synchronizacji oparto na założeniu mówiącym, iż przetwarzany sygnał jest redundantny. Dlatego możliwe jest usunięcie zbędnych, z punktu widzenia rozumienia mowy, fragmentów sygnału i zastąpienie ich fragmentami mowy spowolnionej. Również zastosowanie algorytmu nierównomiernej i zależnej o tempa wypowiedzi modyfikacji czasu trwania sygnału pozwala zmniejszyć rzeczywistą długość sygnału spowolnionego. Istotnym pytaniem jest czy opracowana metoda poprawia rozumienie wypowiedzi. Dlatego postawiono następującą tezę rozprawy:

1. **Zastosowanie nierównomiernej i zależnej od tempa wypowiedzi, modyfikacji czasu trwania mowy, powoduje wzrost zrozumiałości mowy u osób o pogorszonej rozdzielczości czasowej słuchu.**

Wprowadzone przez autora modyfikacje o charakterze nowatorskim wymagały również zbadania jakości mowy spowolnionej oraz oceny skuteczności opracowanych algorytmów wspierających proces modyfikacji struktury czasowej sygnału. Istotny jest tu także fakt, iż w związku z koniecznością przetwarzania w czasie rzeczywistym sygnału rejestrowanego przez mikrofon, niemożliwe było zastosowanie niektórych metod przetwarzania sygnału ze względu na wprowadzane przez nie opóźnienie. Stąd wynika druga teza rozprawy:

2. **Opracowana metoda modyfikacji tempa mowy w czasie rzeczywistym, zapewnia wysoką jakość i naturalność subiektywnie odbieranej wypowiedzi.**

Ważną z punktu widzenia zastosowania opracowanej metody jest jej uniwersalność, która prowadzi do wielu scenariuszy użycia oraz różnorodnych implementacji sprzętowych. Możliwe jest na przykład wykorzystanie tej metody do celów modyfikacji mowy odtwarzanej przez odbiornik telewizyjny. Jednak w odróżnieniu od rozwiązania sprzętowego przedstawionego przez Nakamura et al, możliwa jest tu implementacja algorytmu np. w tunerze telewizji cyfrowej czy bezpośrednio w odbiorniku telewizyjnym. Innym zastosowaniem jest urządzenie przenośne podobne do rozwiązania zaproponowanego przez Nejime *et al.*. W takim rozwiązaniu dedykowany sprzęt zastąpić można za pomocą telefonu komórkowego typu smartfon (*ang. Smartphone*). Kolejnym przeznaczeniem jest wspieranie osób z pogorszoną rozdzielczością czasową słuchu podczas rozmów telefonicznych. W tym zastosowaniu algorytm modyfikacji może być zaimplementowany zarówno po stronie aparatu telefonicznego jak i po stronie centrali telefonicznej. Idealnym przeznaczeniem wydaje się także aparat słuchowy przeznaczony do celów wspomagania rozumienia mowy przez osoby z (C)APD. O konieczności opracowania takiego aparatu mówił Spitzer [171] i podkreślał[2], że takie urządzenie jest „muzyką przyszłości".

Autor pragnąłby nadmienić, iż opracowana w ramach tej rozprawy metoda modyfikacji sygnału mowy została zaimplementowana w wielu różnych wariantach uwzględniających wybrane zastosowania wymienione powyżej [83]. Prace implementacyjne wykonane zostały przez zespół osób pracujących w ramach projektu TYPOSZEREG prowadzonego w Katedrze Systemów Multimedialnych (KSM) [25]. Dodatkowo zarówno opracowana metoda modyfikacji sygnału jak i system ją wykorzystujący zostały zgłoszone w Urzędzie Patentowym Rzeczypospolitej Polskiej [79] [80] w celu przyznania patentu.

[2] W 2002 roku.

2 Wybrane metody modyfikacji czasu trwania i analizy sygnału mowy

W niniejszym rozdziale przedstawiono przegląd wybranych metod modyfikacji czasu trwania sygnału oraz algorytmów detekcji mowy (ang. *Voice Activity Detection* – VAD), detekcji samogłosek oraz estymacji tempa wypowiedzi (ang. *Rate Of Speech* – ROS). W części poświęconej metodom TSM sygnału skupiono się na algorytmach operujących w dziedzinie czasu, pomijając te operujące w dziedzinie częstotliwości [33] [86] [147], bazujące na modelowaniu przetwarzanego sygnału [51] [2] oraz tzw. metody hybrydowe [160] [37] [39]. Zawężenie rozważań dotyczących algorytmów TSM sygnału związane jest z faktem, iż tematyka rozprawy dotyczy metod modyfikacji sygnału mowy, a metody operujące w dziedzinie czasu są projektowane właśnie do tych celów i pozwalają uzyskać wysokiej jakość mowę o zmodyfikowanej strukturze czasowej. Dodatkowo są one mniej złożone obliczeniowo niż metody operujące w dziedzinie częstotliwości (np. wokoder fazowy [147]) czy metody analizujące model sygnału, co jest istotne z punktu widzenia przetwarzania sygnału w czasie rzeczywistym. W podrozdziałach 2.2, 2.3 oraz 2.4 opisano wybrane algorytmy segmentacji sygnału mowy oraz estymacji tempa wypowiedzi. Metody te są często wykorzystywane w procesie nierównomiernej TSM sygnału mowy. Informacje dotyczące zawartości aktualnie modyfikowanego fragmentu sygnału pozwalają na wybór odpowiedniej strategii związanej modyfikacją struktury czasowej sygnału np. poprzez dobór różnych wartości współczynnika skali zależnie od rodzaju przetwarzanego segmentu sygnału (samogłoska/spółgłoska/cisza).

2.1 Metody modyfikacji czasu trwania sygnału mowy

Modyfikacja czasu trwania sygnału polega na wydłużeniu (bądź skróceniu) czasu trwania przetwarzanego sygnału przy jednoczesnym zachowaniu jego oryginalnej wysokości oraz naturalności brzmienia. Zagadnienie TSM sygnału jest dualne do problemu modyfikacji wysokości sygnału (ang. *Pitch Scale Modification* – PSM). W procesie PSM sygnału zmieniana jest jego wysokość przy zachowaniu oryginalnego czasu trwania nagrania. Algorytmy TSM sygnału stosowane są m.in. w: systemach syntezy mowy (ang. *Text To Speech* – TTS), procesie wspierania nauki języków obcych, postprodukcji filmowej (podczas synchronizacji ścieżki dźwiękowej z obrazem), postprodukcji

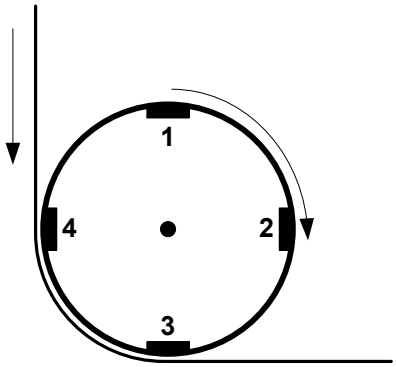
muzycznej, systemach wspomagających osoby niewidome oraz książkach mówionych (ang. *audiobook*) [85] [30] [203] [109] [108] [5].

Głównym założeniem w procesie TSM sygnału jest uzyskanie jak najwyższego podobieństwa sygnału wejściowego $x[n]$ i sygnału o zmodyfikowanej strukturze czasowej $y[n]$. Funkcją opisującą relację pomiędzy czasem w sygnale oryginalnym a czasem w sygnale zmodyfikowanym jest funkcja mapująca $t \rightarrow t' = D(t)$. Jest ona nazywana funkcją odkształcenia czasu (ang. *Time-Scale Warping Function*). Symbolem t oznacza się czas w nagraniu oryginalnym, a symbol t' czas w sygnale o zmodyfikowanej strukturze czasowej. Funkcję mapującą można zdefiniować za pomocą wzoru:

$$t \rightarrow t' = D(t) = \int_0^t \alpha(\tau) d\tau \tag{2.1}$$

gdzie $\alpha(\tau) > 0$ oznacza zmienny w czasie współczynnik skali (ang. *Time-Modification Rate*). W przypadku równomiernej modyfikacji czasu $\alpha(\tau) = \alpha$, a $D(t) = \alpha t$. Jeżeli α przyjmuje wartości większe od 1 wtedy czas trwania oryginalnego sygnału zostaje wydłużony. W przeciwnym razie ($\alpha < 1$) czas trwania sygnału jest skracany. Należy zauważyć, iż dla algorytmów operujących na sygnale cyfrowym funkcja $\alpha(\tau)$ nie jest funkcją ciągłą. W dalszej części rozprawy, poczyniono pewnego rodzaju uproszczenie polegające na uniezależnieniu opisu współczynnika skali od zmiennej związanej z czasem i oznaczeniu go symbolem α.

Jedną z pierwszych znanych metod modyfikacji czasu trwania sygnału była analogowa metoda opracowana przez Fairbanksa *et al.* [45]. Bazowała ona na specjalnie zaprojektowanym magnetofonie wyposażonym w układ składający się z głowicy zapisującej oraz czterech głowic odczytujących. Głowice odczytujące umieszczono na obracającym się cylindrze. Obrót cylindra odbywał się w kierunku przeciwnym do kierunku przesuwu taśmy. W efekcie odtwarzany sygnał dzielony był na krótkie fragmenty, które w zależności od relacji pomiędzy prędkością obrotową cylindra a prędkością przesuwu taśmy, były duplikowane powodując wydłużenie czasu trwania sygnału albo były usuwane powodując jego skrócenie. Na rys. 2.1 przedstawiono schemat ilustrujący zasadę działania urządzenia opracowanego przez Fairbanksa *et al.*



Rys. 2.1 Schemat ilustrujący zasadę działania urządzenia opracowanego przez Fairbanksa *et al.* [45]. Na rysunku znajduje się wirujący cylinder z czterema głowicami odczytującymi oraz poruszająca się taśma magnetyczna. Strzałki wskazują kierunek obrotu cylindra oraz kierunek przesuwu taśmy.

Zasada TSM sygnału opracowana przez Fairbanksa *et al.*, stała się podstawą większości cyfrowych metod TSM sygnału operujących w dziedzinie czasu. Inspiracją stał się pomysł by proces ten oprzeć na zasadzie duplikowania albo usuwania fragmentów sygnału oryginalnego.

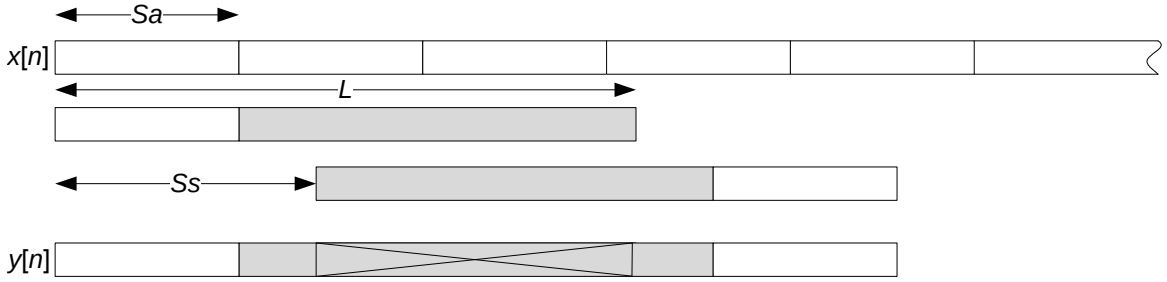
2.1.1. Metoda OLA (ang. *Overlap and Add*)

Zasada modyfikacji czasu trwania sygnału za pomocą algorytmu OLA jest następująca. Przetwarzany sygnał dzielony jest na zachodzące na siebie ramki czasowe o stałej długości L. Przed wykonaniem syntezy sygnału zmieniana jest wartość rozsunięcia ramek, co w konsekwencji prowadzi do modyfikacji struktury czasowej sygnału. Cały proces składa się więc z dwóch kroków: analizy i syntezy. W kroku analizy ramki o długości L próbek pobierane są z sygnału wejściowego ze stałym krokiem Sa. Podczas syntezy długość kroku jest zmieniana i wynosi Ss. Zachodzące na siebie obszary sąsiadujących ramek są ze sobą sumowane z zastosowaniem np. liniowej funkcji zmniejszającej amplitudę sygnału ramki wcześniejszej i zwiększającej amplitudę ramki kolejnej (operacja ang. *cross-fade*). Stosunek długości kroku syntezy i analizy wyznacza wartość współczynnika skali zgodnie ze wzorem:

$$\alpha = \frac{Ss}{Sa} \tag{2.2}$$

Jak można zauważyć, pomimo tego, że opisana metoda w pełni zachowuje oryginalną wysokość sygnału, wprowadza ona zniekształcenia w miejscach łączenia kolejnych ramek. Powstałe zniekształcenia wynikają z braku ciągłości fazy oraz amplitudy fragmentów sygnału znajdujących się w sumowanych przedziałach. Rys. 2.2 ilustruje sposób

modyfikacji czasu trwania sygnału za pomocą algorytmu OLA. W przykładzie przedstawiono proces wydłużania sygnału w czasie.



Rys. 2.2 Sposób modyfikacji czasu trwania sygnału z wykorzystaniem metody OLA.

Operację modyfikacji TSM sygnału wykonywaną zgodnie z algorytmem OLA można zapisać w następujący sposób [179]:

$$y[n] = \frac{\sum_{m=1}^{M} w^2[n - m \cdot Sa]x[n - m \cdot Sa + Sa]}{\sum_{m=1}^{M} w^2[n - m \cdot Sa]} \tag{2.3}$$

gdzie $w[n]$ jest funkcją okna wykorzystywana podczas analizy i syntezy sygnału, a m jest numerem kroku.

2.1.2. Algorytm SOLA (ang. *Synchronized Overlap and Add*)

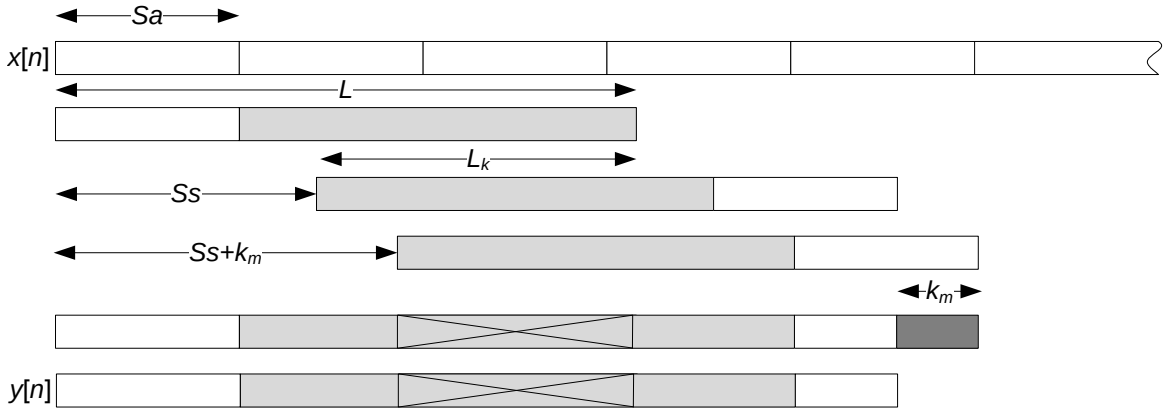
W algorytmie OLA brakuje mechanizmu synchronizującego pozycję ramki sygnału sumowanej w kroku syntezy. Taki mechanizm zaproponowali Roucos i Wilgus [157] w algorytmie nazwanym SOLA. Synchronizacja odbywa się tu poprzez wyznaczenie podobieństwa ramki przesuniętej o krok syntezy z końcem sygnału syntezowanego. Zaproponowane rozwiązanie miało na celu poprawić jakość algorytmu TSM sygnału opartego na wokoderze fazowym [147]. Jak pokazały badania przeprowadzone przez Roucos, jakość mowy modyfikowanej za pomocą opracowanego algorytmu była na tyle wysoka, iż nie wymagane było wykonywanie dodatkowych operacji na widmie przetwarzanego sygnału. Podobieństwo zachodzących na siebie obszarów sygnału może być wyznaczone np. za pomocą funkcji korelacji zdefiniowanej za pomocą wzoru:

$$R_m[l] = \frac{\sum_{j=0}^{L_k-1} y[mSs + l + j]x[mSa + j]}{\sqrt{\sum_{j=0}^{L_k-1} x^2[mSa + j] \sum_{j=0}^{L_k-1} y^2[mSs + l + j]}} \tag{2.4}$$

gdzie $R_m[l]$ jest wartością l-tej próbki tej funkcji korelacji wyznaczoną dla m-tej ramki analizy, a L_k przedziałem zachodzących na siebie części sygnału należących do końca sygnału modyfikowanego i początku aktualnej ramki sygnału. W literaturze można znaleźć także propozycję zastosowania innych miar podobieństwa np. AMDF (ang. *Average Magnitude Difference Function*) [36] czy uproszczonej funkcja korelacji [87]. W celu uzyskania jak najwyższej jakości sygnału zmodyfikowanego w procesie synchronizacji, znajdowane jest miejsce położenia maksymalnej wartość funkcji korelacji k_m dla $k_{min} \le k \le k_{max}$. Łączenie sygnału odbywa się z przesunięciem $m \cdot Ss + k_m$. Zmodyfikowany sygnał można zapisać za pomocą wzoru [179]:

$$y[n] = \frac{\sum_{m=1}^{M} w^2[n - mSa + k_m] x[n + mSa - mSs + k_m]}{\sum_{m=1}^{M} w^2[n - mSa + k_m]} \tag{2.5}$$

Na rys. 2.3 przedstawiono sposób modyfikacji czasu trwania sygnału z wykorzystaniem algorytmu SOLA. Rysunek przedstawia proces wydłużania czasu trwania sygnału. Dzięki zastosowanemu mechanizmowi synchronizacji ramek sygnału, częściowo wyeliminowano problem braku ciągłości fazy i prawie całkowicie problem braku ciągłości amplitudy sygnału zmodyfikowanego.



Rys. 2.3 Sposób modyfikacji czasu trwania sygnału z wykorzystaniem metody SOLA.

Jak można zauważyć po zsumowaniu aktualnie analizowanej ramki sygnału z sygnałem wyjściowym, sygnał $y[n]$ jest dłuższy o k_m próbek od oczekiwanej długość wynikającej z wartości współczynnika skali α. Dlatego przed kolejnym krokiem syntezy sygnału konieczne jest usunięcie z sygnału $y[n]$, k_m ostatnich próbek. Dzięki temu długość sygnału wyjściowego nie przekracza długości wynikającej z wartości współczynnika skali.

Należy tu zauważyć, iż jakość sygnału zmodyfikowane mocno zależy od właściwego doboru długości ramki L, wielkości kroku analizy Sa oraz przedziału poszukiwań wykorzystywanego podczas synchronizacji. W literaturze istnieje wiele propozycji strategii doboru tych parametrów. Najpopularniejszym rozwiązaniem jest zastosowanie ramki o długości 30 ms (długość trzech okresów podstawowych najniższej harmonicznej przetwarzanego sygnału), kroku analizy równego $L/2$ i przedziału poszukiwań mieszczącego się pomiędzy $-L/2$ do $L/2$ [36]. Autor rozprawy także prowadził pomiary mające na celu wyznaczenie optymalnych wartości tych parametrów. Badania oparto na serii testów subiektywnych, podczas których określono wartości L i Sa, dla których ocena jakości mowy spowolnionej jest najwyższa. Wyniki testów pokazały, iż dla mowy najwyżej oceniano jakość sygnału spowolnionego, gdy L/Fs = 46,33 ms, a Sa = L/2 [78] [81] [82]. Innym sposobem jest uzależnienie wartości parametrów algorytmu od współczynnika skali. Jak wykazał Dorran *et al.* [38] uwzględnienie tej relacji pozwala na znaczną poprawę jakości mowy zmodyfikowanej. W swojej pracy zaproponował on następujący sposób doboru parametrów algorytmu [38]:

$$L = L_k + \alpha \cdot \left(\frac{L_k - P}{|1-\alpha|} \right) \tag{2.6}$$

$$L_k = L - S_s \tag{2.7}$$

$$k_{\max} - k_{\min} = P \tag{2.8}$$

gdzie L_k zgodnie ze wzorem (2.7) jest długością zachodzących na siebie obszarów w kroku syntezy, a P jest okresem podstawowym modyfikowanego sygnału.

2.1.3. Algorytm WSOLA (ang. *Waveform Similarity Overlap and Add*)

Algorytm WSOLA został zaproponowany przez Verhelsta i Roelandsa [179] [180] jako pewnego rodzaju odwrotność algorytmu SOLA. Różni się on sposobem znajdowania podobieństwa sygnałów sumowanych w kroku syntezy. Algorytm SOLA dąży do uzyskania możliwie największego podobieństwa po stronie sygnału wyjściowego $y[n]$, a ramki analizy pobierane są w równomiernych odstępach $m \cdot Sa$. W algorytmie WSOLA maksymalizowane jest podobieństwo po stronie sygnału wejściowego $x[n]$, przez co korygowana jest wartość $m \cdot Sa$, w taki sposób, by uzyskać maksymalne podobieństwo pomiędzy ramką analizy a fragmentem sygnału wejściowego, który w sposób naturalny mógłby zostać połączony z sygnałem wyjściowym. Jako miara podobieństwa

wykorzystana może być np. funkcja korelacji skrośnej. Ramki w kroku syntezy, po ich przesunięciu o wartość *Ss*, łączone są ze sobą poprzez zastosowanie operacji *cross-fade* w obszarze zachodzących na siebie fragmentów końca sygnału $y[n]$ i początku aktualnie analizowanej ramki sygnału wejściowego. Ponieważ podobieństwo wyznaczane jest w kroku analizy, w kroku syntezy wykorzystywana jest stała wartość przesunięcia *Ss*. Stały krok syntezy eliminuje problem z różną (od wynikającej z wartości współczynnika skali) długością sygnału $y[n]$ występującą w kolejnych krokach przetwarzania.

Modyfikację czasu trwania sygnału za pomocą algorytmu WSOLA można zapisać za pomocą wzoru [180]:

$$y[n] = \sum_{m=1}^{M} w^2[n - mSs]x[n + mSa - mSs + k_m] \tag{2.9}$$

gdzie k_m jest przesunięciem kroku analizy wyznaczonym dla m-tej ramki sygnału. Jak można zauważyć, z powodu użycia stałej wartości kroku syntezy mianownik równości (2.5) odpowiadający za normalizację sygnału, wymaganą w przypadku wykorzystania okna analizy innego niż kwadratowe (np. okna Hanninga), jest równy 1.

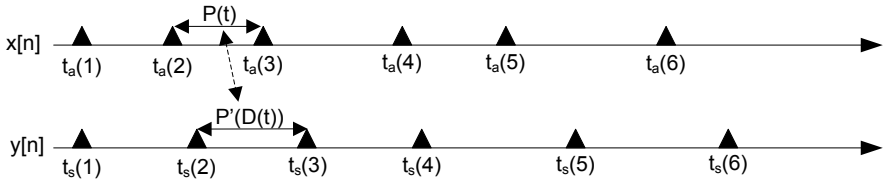
2.1.4. Algorytm PSOLA (ang. *Pitch Synchronous Overlap Add*)

Omówione do tej pory metody TSM sygnału mowy bazowały na zasadzie maksymalizacji podobieństwa łączonych fragmentów przetwarzanego sygnału (po stronie sygnału wyjściowego lub wejściowego). Moulines i Charpentier [108] zaproponowali metodę przeznaczoną do modyfikacji czasu trwania sygnału mowy nazwaną PSOLA lub TD-PSOLA (ang. *Time-Domain* PSOLA). Jest ona oparta na założeniu, że zmiany struktury czasowej sygnału należy dokonywać z uwzględnieniem lokalnego okresu podstawowego sygnału mowy.

W odróżnieniu od algorytmów omówionych powyżej, algorytm PSOLA wykorzystuje zmienną długość ramki sygnału *L* oraz zmienną wartość kroku analizy *Sa*. Długość ramki zależna jest tu od wartości chwilowego okresu podstawowego sygnału mowy wyznaczonego dla każdej ramki analizy. Najczęściej stosuje się *L* dwukrotnie większe od lokalnej wartości okresu podstawowego występującej w analizowanym sygnale. Krok analizy jest zsynchronizowany ze zmiennością okresu podstawowego sygnału. Dla mowy bezdźwięcznej długość ramki oraz krok pobierania ramek sygnału jest stały. W metodzie PSOLA ramki analizy są zsynchronizowane z okresem podstawowym modyfikowanej

mowy, przez co w kroku syntezy możliwe jest bezpośrednie łączenie sąsiadujących ze sobą ramek.

Przetwarzanie sygnału z wykorzystaniem algorytmu PSOLA wykonywane jest w dwóch krokach. Pierwszy krok algorytmu przedstawiono symbolicznie na rys. 2.4. Na podstawie położenia chwil analizy $t_a[m]$ wyznaczane są tu chwile syntezy sygnału $t_s[m]$. Chwile analizy odpowiadają środkom ramek wykorzystywanych podczas analizy sygnału. Operacja przypisania odbywa się w taki sposób by zachowana została wartość chwilowa okresu podstawowego sygnału oryginalnego $P(t)$, oraz uwzględniona zostało wartość użytego współczynnika skali. Oznacza to, iż odległość pomiędzy sąsiednimi znacznikami $t_s[m-1]$ i $t_s[m]$ musi być równa okresowi podstawowemu sygnału oryginalnego wyznaczonego w okolicy chwili $t_a = D^{-1}(t_s[m])$.



Rys. 2.4 Wyznaczenie chwil syntezy

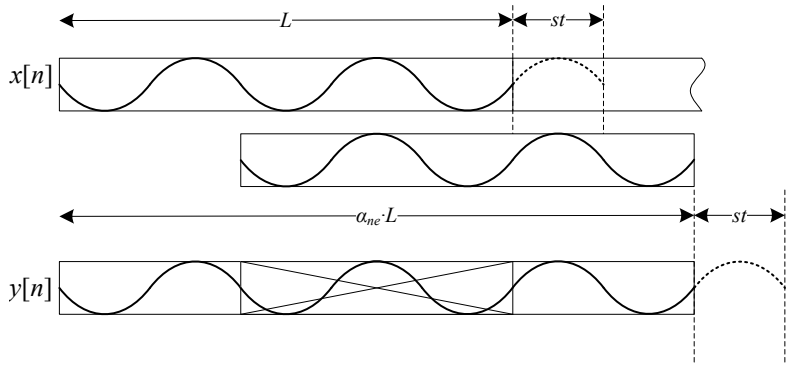
W drugim kroku każdej chwili syntezy przypisywany jest jedna chwila analizy. Operacja ta wykonywana jest w taki sposób, by chwili t_s odpowiadała taka chwila t_a, której środek znajduje się możliwie blisko chwili $D^{-1}(t_s)$. Przy zwiększaniu czasu trwania sygnału niektóre ramki sygnału wejściowego są duplikowane, a przy skracaniu czasu trwania sygnału niektóre ramki są usuwane. Sygnał o zmienionej strukturze czasowej powstaje poprzez połączenie ramek odpowiadających punktom syntezy.

Moulines i Charpentier nie opracowali metody wyznaczania okresu podstawowego sygnału. Zaleca się jednak [109] zastosowanie jednej z istniejących metod np. metody autokorelacyjnej. W literaturze można znaleźć także metody detekcji okresu podstawowego opracowane specjalnie na potrzeby algorytmu PSOLA. Za przykład mogą posłużyć tu algorytmy proponowane przez Lina [89] czy Chalamandarisa [13].

2.1.5. Algorytm AOLA (*ang. Adaptive Overlap and Add*)

Algorytm AOLA został zaproponowany przez Lawlora i Fagana [87] jako metoda pozwalająca, przy zachowaniu bardzo niewielkiej złożoności obliczeniowej (około 90% obliczeń mniej niż dla algorytmu SOLA), uzyskać zadawalającą jakość mowy zmodyfikowanej. Sygnał przetwarzany jest tu stosując ramki o stałej długości L. Zakłada się, iż długość ramki powinna by większą lub równa dwóm okresom najniższej składowej

częstotliwościowej występującej w sygnale przetwarzanym (zazwyczaj L = 50 ms). Skok ramki podczas analizy jest zmienny. Dodatkowo zdefiniowano dwie wartości współczynnika skali: α_{ne} – współczynnik naturalnej zmiany skali oraz α_{de} współczynnik oczekiwanej skali. Wartość α_{de} – jest stosunkiem długości sygnału zmodyfikowanego przez algorytm do długości sygnału wejściowego. Wartość ta będzie wykorzystywana także w dalszej części rozprawy. Na rys. 2.5 przedstawiono sposób wydłużania czasu trwania sygnału mowy z wykorzystaniem metody AOLA.



Rys. 2.5 Sposób modyfikacji czasu trwania sygnału z wykorzystaniem metody AOLA.

Przetwarzanie sygnału odbywa się w następujących krokach:

a. Z sygnału wejściowego ($x[n]$) pobierana jest ramka sygnału o długości L,

b. tworzona jest kopia ramki,

c. skopiowana ramka sygnału przesuwana jest tak, by szczyty lub doliny sygnału po przesunięciu były ze sobą zsynchronizowane (operacja ta jest wykonywana poprzez znalezienie dwóch maksimów w analizowanej ramce sygnału),

d. połączenie ramki analizy oraz jej przesuniętej kopii prowadzi do uzyskania sygnału wydłużonego zgodnie z naturalnym współczynnikiem skali α_{ne},

e. wartość oczekiwanego współczynnika skali osiąga się poprzez dodanie na końcu połączonych ramek fragmentu sygnału $x[n]$. Długość dodanego fragmentu jest równa st,

f. w następnym kroku przetwarzania koniec ramki analizy znajduje się w chwili $L+st$, a procedura przetwarzania powtarzana jest od początku.

Zaletą opisanej powyżej metody jest to, iż zachowuje ona ciągłość sygnału. Właściwość ta wynika z faktu, iż w każdym kroku przetwarzania sygnał wydłużany jest o naturalnie łączący się z nim fragmentem sygnału wejściowego st. Długość dodawanego fragmentu obliczona może być za pomocą wzoru:

$$st = L\frac{(1-\alpha_{ne})}{(1-\alpha_{de})} \tag{2.10}$$

2.1.6. Metody nierównomiernej modyfikacji czasu trwania sygnału mowy

Użycie liniowej funkcji odkształcającej czas jest jednoznaczne z równomiernym wydłużeniem albo skróceniem sygnału oryginalnego. Jednak nie zawsze równomierna modyfikacja czasu trwania sygnału pozwala zapewnić pełną naturalność sygnału zmodyfikowanego. O ile dla nagrań muzycznych liniowy przebieg funkcji $D(t)$ jest zazwyczaj pożądany, o tyle dla sygnału mowy zastosowanie równomiernej modyfikacji różnych segmentów sygnału, takich jak cisza, samogłoski, spółgłoski czy głoski wybuchowe, może powodować powstanie nienaturalnie brzmiącej mowy zmodyfikowanej. Jest to związane ze strategią wytwarzania mowy, w której człowiek w sposób naturalny zmieniając tempo wypowiedzi wydłuża (albo skraca) różne głoski w sposób nierównomierny. Opisane powyżej metody TSM sygnału umożliwiają wykorzystanie nieliniowej funkcji $D(t)$, jednak zasady określające sposób jej tworzenia muszą zostać określone niezależnie od zastosowanego algorytmu.

Dla sygnału mowy funkcja $D(t)$ powinna być zależna od zawartości sygnału wejściowego (mowa/ brak mowy, głoski dźwięczne/bezdźwięczne, samogłoski, spółgłoski) oraz od wartości docelowego współczynnika skali α_{de}. W praktyce funkcja $D(t)$ jest tworzona poprzez stosowanie różnych chwilowych wartości współczynnika skali α zależnych od zawartości sygnału wejściowego.

W literaturze istnieje wiele propozycji analizy zawartości sygnału wejściowego stosowanych do celów tworzenia nieliniowej funkcji $D(t)$ [30] [31] [21] [116] [118] [121] [123] [22] [16]. Dwa najpopularniejsze rozwiązania opierają się na podziale sygnału wejściowego na następujące kategorie:

I. mowa/brak mowy, samogłoski, dźwięczne/bezdźwięczne spółgłoski, głoski wybuchowe, transjenty (przejścia pomiędzy różnymi kategoriami)

II. głoski akcentowane/nieakcentowane, mowa/brak mowy.

Jak można zauważyć II sposób segmentacji jest podziałem ogólnym zawierającym w sobie szczegółowe kategorie uwzględnione w sposobie I. Na przykład, do głosek akcentowanych można zaliczyć samogłoski, dźwięczne spółgłoski oraz głoski wybuchowe, a do głosek nieakcentowanych spółgłoski.

Oprócz sposobu analizy sygnału wejściowego na naturalność mowy przetworzonej duży wpływ ma strategia doboru chwilowej wartości α. W zależności od tego, czy sygnał ma zostać wydłużony czy skrócony, sposób doboru wartości α powinien być inny. Z perspektywy tej rozprawy istotna jest jedynie sytuacja, w której czas trwania sygnału jest zwiększany. Dlatego w dalszej części tego rozdziału skupiono się na opisie strategii doboru wartości α w sytuacji wydłużania czasu trwania sygnału.

Jedną z prostszych, pod względem analizy zawartości sygnału, jest metoda zaproponowana przez Nejime *et al.* [121] [122]. Prostota tej metody wynika z założenia, iż musi ona być w stanie modyfikować czas trwania sygnału w czasie rzeczywistym. Należy tu przypomnieć, że w rozprawie jako metody TSM sygnału mowy operujące w czasie rzeczywistym rozumie się takie algorytmy które pozwalają na przetworzenie sygnału mowy rejestrowanej przez mikrofon i jednoczesne jej odtworzenie oraz wprowadzają opóźnieniu analizy nie większym niż to powodowane przez cyfrowe karty dźwiękowe (około 30–50 ms). Dobór współczynników skali odbywa się tu z wykorzystaniem segmentacji sygnału na trzy kategorie: cisza, mowa akcentowana, mowa nieakcentowana. Mowa akcentowana spowalniana jest z wykorzystaniem większych wartości α, niż ma to miejsca przy mowie nieakcentowanej. Dodatkowo, fragmenty ciszy nie podlegają modyfikacji, a jeżeli trwają dłużej niż jedną sekundę są skracane. Takie rozwiązanie pozwala na wydłużanie jedynie najistotniejszych (z punktu widzenia rozumienia wypowiedzi) segmentów mowy i redukcję redundantnego sygnału. Modyfikacja struktury czasowej sygnału wykonywana jest za pomocą algorytmu TDHS (ang. *Time-Domain Harmonic Scaling*). Jest to algorytm opracowany przez Malaha [96]. Zasada modyfikacji czasu trwania sygnału jest tu podobna do tej stosowanej w algorytmie PSOLA. Głównym założeniem metody jest podział sygnału mowy na ramki czasowe z uwzględnieniem chwilowej wartości okresu podstawowego modyfikowanego sygnału. Długość ramki analizy jest tu wielokrotnością okresu podstawowego analizowanego sygnału.

Nejime *et al.* [122] [121] nie przeprowadzili testów subiektywnych pozwalających na ocenę jakości mowy spowolnionej z wykorzystaniem ich algorytmu. Zbadali jednak skuteczność metody w zastosowaniu, polegającym na wspomaganiu procesu rozumienia mowy przez osoby z pogorszoną rozdzielczością czasową słuchu. Wyniki tych badań pokazały, iż spowolniona w ten sposób mowa jest lepiej rozumiana przez niektórych słuchaczy. W ich badaniach nie pokazano jednak zależności pomiędzy zaburzeniami słuchu a stopniem poprawy rozumienia mowy.

Nakamura *et al.* [118] [116], także zaproponowali metodę nierównomiernej modyfikacji czasu trwania sygnału działającą w czasie rzeczywistym. Ich rozwiązanie przeznaczone było do celów modyfikacji sygnału mowy odtwarzanej przez odbiornik telewizyjny. Użytkownik mógł wybierać niezależnie wartości współczynników skali dla głosek dźwięcznych w zakresie od 1,0 do 1,6 i dla ciszy w zakresie od 1,0 do 3,0. Głoski bezdźwięczne nie były poddawane modyfikacji. W celu uzyskania synchronizacji pomiędzy sygnałem wejściowym i sygnałem spowolnionym, wartości współczynników skali były automatycznie zmniejszane przez algorytm tak by w ramach jednej frazy, czas sygnału na wejściu i na wyjściu algorytmu był zbliżony. Takie rozwiązanie skutkowało tym, iż na początku frazy mowa była spowalniana, a na końcu przyspieszana. Podobnie jak Nejime *et al.*, Nakamura *et al.* do modyfikacji struktury czasowej sygnału wykorzystali algorytm TDHS.

Coyle *et al.* [23] [35] w swojej metodzie zaproponowali wykorzystanie analizy sygnału mowy poprzez jej segmentację. Użyto tu podziału na kategorie należące do grupy I z pominięciem transjentów. Przedstawione reguły doboru α, zostały oparte na badaniach Ebihara *et al.* [40] i Kuwabara [84]. W pierwszych badaniach pokazano, iż dla zmiennego tempa wypowiadanej mowy, czas trwania bezdźwięcznych fragmentów jest mniej zmienny niż czas trwania fragmentów dźwięcznych. Dlatego w procesie nierównomiernej TSM sygnału sugerowano stosowanie modyfikacji czasu trwania fragmentów mowy dźwięcznej lub miejsc występowania samogłosek. Dodatkowo Kuwabara zauważył, iż czas trwania dźwięcznych spółgłosek, niezależnie od tempa wypowiedzi, zmienia się bardziej niż czas trwania bezdźwięcznych spółgłosek.

W opisanej przez badaczy metodzie, Coyle zakładał zastosowanie trzech różnych wartości współczynnika skali: $\alpha_1 > \alpha_2 > \alpha_3 > 1$. Wartość α_3 używano dla samogłosek, α_2 dla dźwięcznych spółgłosek oraz α_1 w obszarach występowania bezdźwięcznych spółgłosek. Ponadto uznano, iż w obszarach ciszy współczynnik skali powinien być taki sam jak w miejscach występowania bezdźwięcznych spółgłosek, a struktura czasowa głosek wybuchowych musi pozostać nienaruszona ($\alpha = 1$). Jak pokazały porównawcze badania subiektywne przeprowadzone przez autorów tej metody, 88% słuchaczy preferowało mowę zmodyfikowaną za pomocą zaproponowanego sposób. Algorytm został porównany z metodą równomiernej modyfikacji oraz z dwiema metodami, w których modyfikowano jedynie czas trwania samogłosek lub czas trwania głosek dźwięcznych. Do zmiany struktury czasowej sygnału wykorzystano algorytm AOLA. Niestety autorzy nie

opracowali algorytmów pozwalających na automatyczny podział mowy na segmenty. Do celów eksperymentalnych wykorzystali oni ręczną segmentację wypowiedzi.

Innym przykładem bazującym na tych samych kategoriach segmentów co metoda Coyle *et al.* jest metoda opracowana przez Demola *et al.* [30] [31]. Użyto tu jednak innej strategii doboru chwilowych wartości współczynnika skali. W swoich pracach Demol *et al.* przedstawili sposób adaptacji wartości współczynnika skali zarówno w sytuacji, gdy $\alpha_{de} > 1$, jaki i w sytuacji odwrotnej (gdy $\alpha_{de} < 1$). Niestety opis doboru parametrów jest tu niespójny, co prowadzi do problemu w określeniu wartości użytych współczynników skali. Chodzi tu o sytuację, w której autorzy referatu piszą [30], iż wartość α w obszarze samogłosek jest wyższa niż w przypadku spółgłosek, a liczbowo zdefiniują to w następujący sposób: chwilowa wartość współczynnika skali jest równa iloczynowi $\delta_n \cdot \alpha$, gdzie α wyznaczana jest na podstawie wartości α_{de}, δ_n to waga zależna od rodzaju segmentu sygnału. Relacja pomiędzy wagami jest następująca $\delta_1 < \delta_2 < \delta_3 < \delta_4 < 1$, wagi przypisano do segmentów mowy tak, że δ_1 jest wagą przypisaną do segmentów ciszy, δ_2 – wagą samogłosek, δ_3 – wagą spółgłosek, δ_4 to waga transjentów. Z powyższej opisanej strategii doboru wartości współczynnika skali wynika, iż samogłoski powinny zostać wydłużone ze współczynnikiem mniejszym niż spółgłoski. Innym wytłumaczeniem tej nieścisłości może być odwrotna, niż tradycyjnie przyjęta w literaturze, definicja współczynnika skali α, mianowicie wartość powyżej 1 oznacza skrócenie, a poniżej 1 wydłużenie sygnału. Nieścisłość ta jest istotna, ponieważ wyniki testów subiektywnych przedstawione przez Demola *et al.* pokazują, iż statystycznie nie można zauważyć różnicy pomiędzy oceną jakości mowy spowolnionej z wykorzystaniem metody równomiernej i nierównomiernej modyfikacji czasu trwania sygnału.

Demol *et al.* [30] [31] w swojej metodzie do modyfikacji struktury czasowej sygnału użyli algorytmu WSOLA. Segmentacja sygnału odbywała się tu z wykorzystaniem trzech algorytmów: detekcji mowy (progowa analiza energii), detekcji transjentów (analiza różnic energii) oraz detekcji dźwięczności. Segmentację wykonywano w ramkach czasowych przesuwanych z krokiem odpowiadającym 5 ms. Przebiegała ona niezależnie od TSM sygnału. Modyfikację czasu trwania sygnału wykonywano zgodnie z wynikami dostarczonymi przez algorytmy segmentacji.

Istotną innowację w strategii doboru wartości współczynników skali, wprowadzili Covell *et al.* [21] w metodzie nazwanej MACH1. Jest to metoda opracowana głównie do celu zwiększania tempa wypowiedzi, ale niektóre reguły tam przedstawione, mogą zostać

wykorzystane także podczas spowalniania mowy. W systemie MACH1 zaproponowano rozszerzenie tradycyjnych metod analizy o analizę tempa wypowiadanej mowy. Zależnie od estymowanej wartości tempa dokonywano innego rodzaju adaptacji współczynnika skali. Przy szybkiej wypowiedzi była on przyspieszana z wykorzystaniem mniejszej wartości α niż wypowiedź wolna. Taka strategia jest celowa, gdyż prowadzi do zrównania (na wyjściu algorytmu) temp mowy oryginalnie szybkiej i wolnej.

W tab. 2.1 przedstawiono zestawienie znanych w literaturze metod nierównomiernej modyfikacji czasu trwania sygnału mowy. Jak można zauważyć, wszystkie z nich opierają się na segmentacji sygnału wejściowego, a wykrycie sygnału użytecznego (mowy) stanowi tu jeden z głównych elementów. Dlatego w dalszej części tego rozdziału przedstawiono przegląd algorytmów, które mogą zostać wykorzystane do celów detekcji sygnału mowy. Algorytmy segmentacji używane przez autorów przedstawionych metod nierównomiernej TSM sygnału zazwyczaj były dobierane w taki sposób by zbadać strategię doboru chwilowych wartości α, a nie stworzyć algorytm pozwalający na pracę w warunkach rzeczywistych. Dlatego w dalszej części rozdziału przedstawiono przegląd metod estymacji tempa wypowiedzi, które wydaje się być kluczowym parametrem pozwalającym na zwiększenie jakości mowy zmodyfikowanej, a zarazem zminimalizowanie różnicy w czasie trwania sygnału oryginalnego i spowolnionego. Założenie to potwierdzają m.in. wyniki przedstawione przez Covella *et al.* [21] w eksperymencie polegającym na przyspieszeniu wypowiedzi algorytmem, w którym chwilowa wartość α zależała od tempa wypowiedzi. Dodatkowo wyniki przedstawione w tab. 2.1 sugerują, że nie tyle detekcja dźwięczności co detekcja samogłosek ma istotny wpływ na jakość sygnału wydłużonego. W związku z tym faktem w następnym podrozdziale przedstawiono także znane z literatury algorytmy detekcji samogłosek. Należy także zauważyć, iż prawie zawsze zastosowanie nierównomiernej TSM sygnału mowy pozwalało uzyskać wyższą jakość sygnału zmodyfikowanego niż modyfikacja za pomocą algorytmu stosującego równomierną TSM sygnału.

Tab. 2.1 Porównanie metod nierównomiernej modyfikacji czasu trwania sygnału

Autor/Nazwa metody[3]	**Algorytm TSM**[4]	**Detekcja segmentów**	**Sposób tworzenie funkcji *D*(*t*)**	**Jakości**[5]
Coyle [23] [35] (O)	AOLA (E)	dźwięcznych, samogłosek, mowy/ciszy, głosek wybuchowych, transjentów[6]	$1 < \alpha_3 < \alpha_2 < \alpha_1$ cisza→ α_1 samogłoski→ α_2 dźwięczne spółgłoski→ α_1 bezdźwięczne→ α_3 głoski wybuchowe→ $\alpha = 1$	T
Demol [31] [30] (O)	WSOLA (E+C)	dźwięcznych, samogłosek, mowy/ciszy, głosek wybuchowych, transjentów	$\delta_1 < \delta_2 < \delta_3 < \delta_4 < 1$ cisza→$\delta_1 \cdot \alpha$ samogłoski→$\delta_2 \cdot \alpha$ spółgłoski→$\delta_3 \cdot \alpha$ transjenty→$\delta_4 \cdot \alpha$ głoski wybuchowe→ $\alpha = 1$	N
W. Chu [16] (O)	WSOLA (C)	analiza energii sygnału	$\alpha_2 < \alpha_1 < 1$ wysoka energia→ α_1 niska energia→ α_2	T $\alpha < 0{,}4$
MACH1 [21] (O)	SOLA (C)	analiza energii sygnału, estymacja tempa wypowiedzi	$\alpha_2 < \alpha_1 < 1$; $\alpha_3 < \alpha_4 < 1$ mowa szybka→ α_1 mowa wolna→ α_2 sylaby akcentowane→ α_3 sylaby nieakcentowane i cisza→ α_4	T
Nakamura [118] (R)	Zmodyfikowany TDHS (E)	dźwięczność, częstotliwość podstawowa	głoski dźwięczne→ $\alpha \epsilon <1{,}0;1{,}6>$ głoski bezdźwięczne→$\alpha = 1$ cisza→ $\alpha \epsilon <1{,}0; 3{,}0>$	b.d.
Nejime [123] [122] [121] (R)	Zmodyfikowany TDHS (E)	mowa/cisza, sylaby akcentowane,	$\alpha_2 > \alpha_3 > 1$ cisza→ $\alpha_1 = 1$ sylaby akcentowane→ α_2 sylaby nieakcentowane i cisza→ α_3	b.d.

2.2 Detekcja mowy

Zadaniem algorytmów VAD jest wykrycie w sygnale fonicznym przedziałów czasu występowania mowy i odróżnienie ich od fragmentów zawierających jedynie szum otoczenia. Za szum otoczenia uznawany jest każdy sygnał, który nie jest interesujący z punktu widzenia analizy sygnału mowy np. mowa innych osób znajdujących się w otoczeniu osoby, której głos jest analizowany. Takie postawienie problemu klasyfikacji wiąże się z koniecznością opracowania algorytmów odpornych na zakłócenia będące zarówno sygnałami stacjonarnym (np. szum biały), jak również sygnałami o charakterze

[3]R – przetwarzanie sygnału w czasie rzeczywistym, O – przetwarzanie sygnału w trybie *offline*
[4]E- Możliwość wydłużenia czasu trwania sygnału, C – możliwość skrócenia czasu trwania sygnału
[5]Subiektywna ocena jakości – T jakość wyższa niż dla równomiernej modyfikacji, N nie zauważono poprawy jakości
[6]Segmentacja sygnału wykonywana ręcznie

niestacjonarnym: hałas silnika samochodowego (ang. *volvo noise*), hałas prac w fabryce (ang. *factory noise*), sygnały powstałe poprzez sumowanie sygnałów mowy pochodzących od wielu mówców (ang. *babble noise*).

Zakłada się, iż idealny VAD powinien cechować się następującymi własnościami [161]:

- odpornością na szum – niezależnie od SNR oraz charakteru sygnału zakłócającego, algorytm powinien osiągać wysoki współczynnik poprawnie rozpoznany ramek sygnału mowy oraz ramek szumowych.
- dokładnością – zarówno w detekcji wystąpień mowy jak i w detekcji fragmentów szumowych,
- adaptacyjnością do różnych warunków pracy – algorytm powinien być w stanie dopasowywać się do różnej charakterystyki widmowej i poziomu szumu,
- prostotą – zbyt duża złożoność obliczeniowa może wprowadzać ograniczenia związane z możliwością wykorzystania algorytmu do detekcji mowy w czasie rzeczywistym,
- pracą w czasie rzeczywistym – w większości zastosowań konieczna jest detekcja mowy wykonywana w czasie rzeczywistym, jednak np. w niektórych systemach ASR warunek ten nie musi być spełniony,
- brakiem założeń związanych z charakterystyką szumu – właściwość ta związana jest z adaptacyjnością i odpornością na szum, czego konsekwencją jest to, iż detektor nie może być odporny jedynie na wybrane rodzaje szumu.

Algorytmy VAD znajdują zastosowanie w wielu dziedzinach związanych z analizą i przetwarzaniem sygnału mowy m.in. w telekomunikacji [61] [43], w algorytmach automatycznego rozpoznawania mowy (ASR) [65] [93], w algorytmach redukcji echa [49], algorytmach nierównomiernej modyfikacji czasu trwania sygnału (p. 2.1.6), w celu szacowania poziomu SNR w sygnale zawierającym mowę [182], w algorytmach automatycznej redukcji szumu [97] [32].

W telekomunikacji algorytmy VAD wykorzystuje się podczas kodowania sygnału podczas nadawania. Analiza zawartości sygnału fonicznego pozwala na kodowanie i transmisję wyłącznie energetycznych/aktywnych fragmentów sygnału zawierającego sygnał mowy, a na etapie zdekodowania sztuczne generowanie i dodawanie szumu (ang.

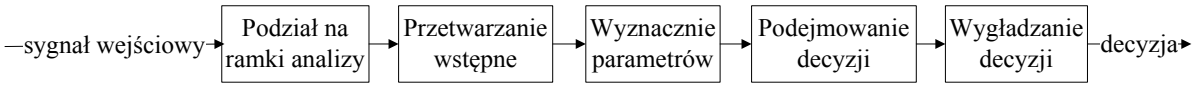
Comfort Noise Generation – CNG) [61] [43] w miejscach występowania ciszy w sygnale oryginalnym. Dzięki takiemu rozwiązaniu uzyskuje się znaczące ograniczenie objętości danych transmitowanych podczas rozmowy. Wynika to ze znaczącej redundancji sygnału wejściowego. Szacuje się, że podczas typowej rozmowy telefonicznej udział każdego z mówców wynosi mniej niż 50% [120], a w rozmowie z wykorzystaniem sieci IP (ang. *Voice over Internet Protocol* – VoIP) około 40% [41]. Dodatkowo w każdej wypowiedzi występują pauzy pomiędzy słowami oraz pomiędzy zdaniami [69]. W systemach ASR zastosowanie detekcji mowy pozwala na zmniejszenie złożoności obliczeniowej procesu rozpoznawania mowy oraz na zwiększenie jej skuteczności np. podczas pracy w niekorzystnych warunkach akustycznych (niski SNR tła) [140].

Problem detekcji mowy jest w literaturze dobrze rozpoznany. Istnieje wiele norm specyfikujących algorytmy przeznaczone do stosowania w telekomunikacji m. in. normy ITU G.729 definiujące algorytm VAD w koderze mowy G.729 [61], ETSI AMR wersje 1 i 2 [43], ETSI AFE [44]. Opracowano także wiele rozwiązań, których dokładność przewyższa detektory zalecane w normach [153] [170], oraz takie, które przeznaczone są do specjalnych zastosowań [182] [188].

Na rys. 2.6 umieszczono schemat blokowy ilustrujący sposób przetwarzania sygnału w typowym algorytmie detekcji mowy. W pierwszym kroku sygnał wejściowy dzielony jest na ramki czasowe o stałej długości L. Długość ramki analizy dobierana jest tak, aby zapewnić możliwość pracy algorytmu w czasie rzeczywistym oraz uwzględnić charakter sygnału mowy (m.in. maksymalny okres podstawowy sygnału mowy). Długość skoku ramki Sa zazwyczaj równa jest L, jednak czasami stosuje się nakładkowanie ramek.

Przetwarzanie wstępne ma za zadanie uwypuklić charakterystyczne dla mowy cechy sygnału. W tym kroku najczęściej stosuje się jedną lub wiele z wymienionych poniżej operacji:

- filtrację za pomocą filtru dolnoprzepustowego [61],
- redukcję szumu [48] [58] [174],
- preemfazę,
- estymatę energii chwilowej – przebiegu TEO (ang. *Teager Energy Operator*) [183].



Rys. 2.6 Schemat blokowy algorytmu detekcji mowy.

W kolejnym kroku, niezależnie od rodzaju zastosowanej metody podejmowania decyzji, ramki sygnału zostają sparametryzowane. W zależności od zdolności danego parametru do podziału danych, zastosowania oraz od algorytmu decyzyjnego, ramki sygnału parametryzowane są z wykorzystaniem jednego lub wielu parametrów. Szczegółowy opis parametrów stosowanych w procesie detekcji mowy został umieszczony w dalszej części tego rozdziału.

Na podstawie wyznaczonych parametrów podejmowana jest decyzja dotycząca przynależności ramki sygnału do jednej z kategorii (mowa/brak mowy). Sposób podejmowania decyzji zależny jest od rodzaju zastosowanego algorytmu. W literaturze można znaleźć podział na trzy podstawowe grupy algorytmów detekcji mowy w zależności od sposobu analizy sygnału: adaptacyjne algorytmy progowe analizujące krótkoczasowe parametry sygnału [103] [104] [197], algorytmy statystyczne bazujące na modelach prawdopodobieństw wyznaczonych dla sygnału mowy i sygnałów szumu [112] [170] [28] [153], algorytmy oparte na metodach sztucznej inteligencji, np. na metodzie wektorów nośnych (ang. *Support Vector Machine* – SVM) [63] [132] [42] [193] [8] [93], sztucznych sieciach neuronowych (ang. *Artificial Neural Network* – ANR) [70], oraz algorytmy będące połączeniem powyższych metod [189].

W ostatnim kroku detekcji, wykonuje się operację wygładzania decyzji mającą na celu minimalizację błędu polegającego na klasyfikacji sygnału mowy, jako jej brak. Wygładzenie uzyskiwane jest poprzez zastosowanie m.in. ukrytych modeli Markova (ang. *Hidden Markov Model* – HMM) [170] lub maszyny stanów [28]. W tym kroku zakłada się, iż decyzja podjętą przez klasyfikator dla *n*-tej ramki analizy zależy wyłącznie od poprzednich decyzji oraz, że prawdopodobieństwo warunkowe bycia w stanie „*brak mowy*” pod warunkiem, że w poprzednim kroku znajdowało się w stanie „*mowa*” jest wyższe niż odwrotne prawdopodobieństwo warunkowe [170].

W dalszej części podrozdziału opisano wybrane metody: parametryzacji sygnału stosowane w algorytmach VAD oraz metody podejmowania decyzji. Wybór opisanych metod ograniczył się do tych, które mogą zostać wykorzystane do detekcji sygnału mowy w czasie rzeczywistym.

2.2.1. Wybrane metody parametryzacji sygnału

Zastosowanie energii

Jako podstawowy parametr w procesie detekcji sygnału mowy wykorzystuje się krótkookresową energię sygnału [150] [88] [159] [61] [197] [103] [192] [158] [58] [149] lub moc sygnału [146] wyznaczane zgodnie ze wzorami:

$$E_m = \sum_{n=1}^{L} x[n]^2 \tag{2.11}$$

$$P_m = \frac{1}{L}\sum_{n=1}^{L} x[n]^2 \tag{2.12}$$

gdzie odpowiednio E_m i P_m oznaczają energię i moc sygnału $x[n]$ wyznaczoną dla m-tej ramki sygnału.

Zakłada się, iż sygnał mowy charakteryzuje się wyższymi wartościami energii niż szum. Założenie to jest prawdziwe jedynie przy danych wartościach SNR oraz dla szumu o charakterze stacjonarnym. W literaturze można znaleźć propozycje zastosowania wielu różnych innych sposób wyznaczenia energii oraz metod opartych na analizie dodatkowych parametrów. Pierwsze rozwiązania bazowały m.in. na analizie częstości przejść przez zero (ang. *Zero Crossing Rate* – ZCR) [150] [88] [61] [192] [149] oraz parametrów związanych z energią sygnału, energią widma amplitudowego sygnału [149], pierwiastkiem średniokwadratowym energii sygnału (ang. *Root Mean Square* – RMS) [194] [90] [158] oraz energią sygnału i widma amplitudowego sygnału wyznaczaną w podpasmach [188] [61] [91] [43] [149] [90]. Podczas wyznaczania energii typowo stosuje się dwa podziały: podpasma rozłożone równomiernie w paśmie mowy 250 Hz–3500 kHz [64] [43] lub podział na pasma w skali melowej [90] (skala melowa została omówiona w dalszej części tego podrozdziału). Parametry, o których była mowa można wyznaczyć korzystając ze wzorów:

$$\mathrm{ZCR}_m = \sum_{n=1}^{L} \left|\mathrm{sgn}(x[n]) - \mathrm{sgn}(x[n-1])\right| \tag{2.13}$$

$$E_m^f = \sum_{k=1}^{K} \left|X(k)\right|^2 \tag{2.14}$$

$$E_m^{rms} = \sqrt{\frac{\sum_{n=1}^{L} x[n]^2}{L}} \tag{2.15}$$

$$E_m^{\mathrm{bi}} = \sum_{k=bStart}^{bStop} \left|X(k)\right|^2 \tag{2.16}$$

$$E_i^b = \sum_{n=1}^{L} x_b[n]^2 \tag{2.17}$$

gdzie ZCR_m oznacza częstość przejść przez zero sygnału w m-tej ramce, E_m^f – energię widma amplitudowego m-tej ramki sygnału, $X(k)$ – dyskretna transformata Fouriera wyznaczona dla m-tej ramki sygnału $x[n]$, E_m^{bf} – energię widma amplitudowego sygnału wyznaczoną w podpasmach, których granice oznaczone są przez wartości *bStart* i *bStop*, E_m^b – energię sygnału sygnału $x_b[n]$, a $x_b[n]$ – sygnał o ograniczonym paśmie.

Metody oparte na analizie energii sygnału oraz ZCR, z powodu braku odporności na szum, stosowane są jedynie w środowiskach o niewielkim poziomie szumu lub w połączniu z algorytmami redukcji szumu. W literaturze [102] zaproponowano wiele różnych parametrów o wyższej odporności na szum niż te opisane powyżej. Poniżej wymieniono oraz podano definicje najważniejszych parametrów, które mogą zostać wykorzystane podczas detekcji sygnału mowy w czasie rzeczywistym.

Wykorzystanie entropii widmowej

Zastosowanie entropii widmowej w celu detekcji sygnału mowy zostało zaproponowane przez Shena *et al.* [167]. Zaobserwował on, iż stosując pewne heurystyczne reguły podczas wyznaczania funkcji gęstości prawdopodobieństwa (ang. *Probability Density Function* – PDF) widma amplitudowego, a następnie obliczając (na jego podstawie) entropie zgodnie z definicją Shannona [164] uzyskuje się parametr odporny na rożnego rodzaju szum. Rozkład prawdopodobieństwa kwadratu widma amplitudowego dla m-tej ramki estymowana jest zgodnie ze wzorem:

$$p_m^k = \frac{\left|X(k)\right|^2}{\sum_{k=1}^{K} \left|X(k)\right|^2} \tag{2.18}$$

gdzie p_m^k oznacza prawdopodobieństwo wystąpienia k-tego prążka w widmie amplitudowym. Dodatkowo podczas obliczania tej wartości jeżeli k-ty prążek odpowiada

częstotliwościom poniżej 250 Hz albo powyżej 6 kHz to $|X(k)|^2 = 0$, a jeżeli $p_m^k < \beta_2$ i $p_m^k > \beta_1$ to $p_m^k = 0$. Dolna granica β_2 ma na celu usunięcie z analizy sygnału szumu o stałym rozkładzie widma, a górna granica β_1 pozwala na pominięcie składowych skoncentrowanych w danym paśmie. Na podstawie wyznaczonych wartości p_m obliczana jest entropia $H_c^m(p_m)$ zgodnie ze wzorem:

$$H_c^m(p_m) = -\sum_{k=1}^{K} p_m^k \cdot \log(p_m^k) \tag{2.19}$$

Jak pokazały badania przeprowadzone przez Shena *et al.* parametr ten, w przypadku różnego rodzaju szumów, pozwala na uzyskanie średniej skuteczności wyższej, od standardowego algorytmu opartego na analizie energii sygnału, o 16%. Ponieważ wartości entropii widma sygnału zaszumionego szumem kolorowym są zbliżone do wartości uzyskiwanych dla fragmentów szumu, Ouzounov [127] zaproponował zastosowanie normalizacji widma poprzez dzielenie widma ramek przez uśredniony kwadrat widma amplitudowego sygnału mowy wyznaczane w pewnym przedziale czasu.

Z kolei Haung i Yang [59] sugerowali połączenie analizy opartej na krótkookresowej energii oraz entropii widma amplitudowego poprzez obliczenie parametru zgodnie ze wzorami:

$$M_m = \left(E_m - \overline{E}\right) \cdot \left(H_c^m(p_m) - \overline{H}_c\right) \tag{2.20}$$

$$EH_c^m(p_m) = \sqrt{1 + |M_m|} \tag{2.21}$$

gdzie $EH_c^m(p^m)$ jest parametrem zaproponowanym przez Haunga i Yanga wyznaczonym dla m-tej ramki sygnału, $\overline{E}$ i $\overline{H}_c$ to średnie wartości energii i entropii widma amplitudowego sygnału wyznaczone dla pierwszym 10 ramek analizy. Połączenie tych dwóch parametrów pozwala na korzystanie z zalet obu wartości jednocześnie.

Wu i Wang. [188], w celu uodpornienia parametru na szum, zaproponowali wyznaczanie entropii sygnału w podpasmach i adaptacyjną analizę istotności poszczególnych podpasm sygnału. Wynikowy parametr jest sumą ważoną wartości entropii sygnału odpowiadającej poszczególnym pasmom, a wagi dobierane są tak by eliminować wpływ pasm nie zawierających sygnału mowy.

Użycie funkcji autokorelacji

Kolejną grupą parametrów są parametry oparte na analizie funkcji autokorelacji wyznaczanej na podstawie widma amplitudowego sygnału (ang. *Spectrum Autocorrelation Function* – SACF). Analiza funkcji autokorelacji pozwala na ocenę okresowości sygnału. Jak pokazują badania prowadzone przez Ouzounova [126], większą separowalność klas (mowa/szum) można uzyskać obliczając funkcję autokorelacji dla kwadratu widma amplitudowego. Dodatkowo zarówno dla głosek dźwięcznych jak i bezdźwięcznych funkcja autokorelacji jest okresowa, a dla szumu jej rozkład jest losowy. Funkcja autokorelacji energii widma wykorzystywana przez Ouzounova zdefiniowana jest następująco [55]:

$$R_p[l] = \sum_{k=1}^{K} \left|X(k)\right|^2 \cdot \left|X(k+l)\right|^2 \tag{2.22}$$

gdzie $R_p[l]$ jest wartością funkcji autokorelacji dla l-tego przesunięcia, a $X(k)$ jest widmem amplitudowym m-tej ramki.

W literaturze można spotkać wiele propozycji opisania funkcji SACF poprzez użycie jednego parametru. Poniżej wymieniono najważniejsze parametry: liczba szczytów funkcji autokorelacji sygnału, stosunek wartości szczytów do wartości dolin wyznaczony dla funkcji SACF (ang. *Spectral Autocorrelation Peak Valley Ratio* – SAPVR) oraz wartość maksymalnego szczytu funkcji autokorelacji sygnału [20] [71]. Ouzounov zaproponował parametryzację bazującą na analizie „funkcji delta" zdefiniowanej wzorem:

$$\Delta R_p[l] = \frac{\sum_{q=-Q}^{Q} qR_p[l+q]}{\sum_{q=-Q}^{Q} q^2} \tag{2.23}$$

gdzie $l = 0, 1, \ldots, L$, Q zazwyczaj przyjmuje wartości z przedziału od 2 do 5. Na podstawie funkcji delta wyznaczonej dla SACF obliczany jest parametr nazwany „średnią deltą" (ang. *Mean Delta* – MD):

$$\mathrm{MD} = \frac{1}{\Delta L} \sum_{l=L_1}^{L_2} \left|\Delta R_p[l]\right| \tag{2.24}$$

gdzie L_1 i L_2 są granicami przesunięcia, a $\Delta L = |L_2 - L_1|$ jest różnicą przesunięć.

Analiza tonalności

Detekcja sygnału mowy na podstawie analizy wysokości dźwięku była proponowana przez wielu autorów [57] [126] [71]. Analiza wysokości pozwala na detekcję głosek dźwięcznych poprzez m. in. śledzenie formantów z wykorzystanie kodowania LPC (ang. *Linear Predicitve Coding*), zastosowanie funkcji autokorelacji [68], czy stosując analizę cepstralną [150].

Yoo i Yook [195] jako pewnego rodzaju odmianę analizy tonalności sygnału, zaproponowali parametr PVD(*VM, A*) (ang. *Peak Valley Difference*) bazujący na analizie podobieństwa widma amplitudowego sygnału do widma amplitudowego samogłosek. Parametr ten zdefiniowali w następujący sposób (2.25):

$$\mathrm{PVD}(VM, X) = \frac{\sum_{k=0}^{K-1}(X(k)\cdot \mathrm{VM}(k))}{\sum_{k=0}^{K-1}\mathrm{VM}(k)} - \frac{\sum_{k=0}^{K-1}(X(k)\cdot(1-\mathrm{VM}(k)))}{\sum_{k=0}^{K-1}(1-\mathrm{VM}(k))} \tag{2.25}$$

gdzie PVD(*VM, A*) jest wartością parametru dla jednej ramki sygnału wejściowego, *X*(*k*) jest wartością *k*-tego prążka w widmie amplitudowym sygnału, a VM(*k*) jest wartością *k*-tego elementu w wektorze opisującym model samogłoski (ang. *Vowel Model* – VM).

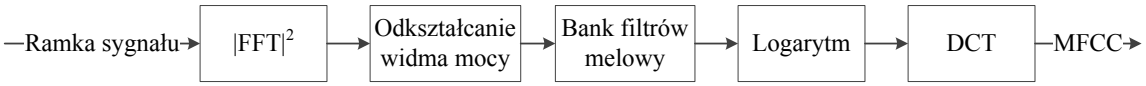
Wektor VM wyznaczany jest na podstawie analizy wartości uśrednionych widm amplitudowych sygnałów zawierających samogłoski wypowiadane przez wielu mówców. Uśrednione widma amplitudowe samogłosek dzielone są na zbiory z wykorzystaniem algorytmu *k*-średnich (ang. *k-means*) [53]. Dla każdego reprezentanta *k*-tego zbioru tworzony jest wektor binarny. Zawiera on wartości 1 w miejscach odpowiadających wystąpieniom szczytów w widmie amplitudowym i 0 w pozostałych miejscach. Podczas tworzenia modelu, analizowane są tylko te szczyty, których wartość przekracza założony próg. Podczas detekcji mowy wybierany jest wektor z modelu VM dla którego wartość parametru PVD jest największa. Jeżeli widmo amplitudowe analizowanego sygnału jest silnie skorelowane z widmem samogłoski, wtedy PVD przyjmuje wartości wyższe niż dla spółgłosek czy braku mowy.

Zastosowanie współczynników mel-cepstralnych

Cepstralna analiza sygnału mowy wykorzystująca parametry mel-cepstralne (ang. *Mel-Frequency Cepstral Coefficents* – MFCC) typowo stosowana jest w systemach automatycznego rozpoznawania mowy oraz mówcy [67] [152]. Parametry MFCC są odporne na różnice poziomów sygnałów wejściowych, dzięki czemu możliwe jest ich

stosowanie w systemach, z których korzystają różni mówcy, a mowa rejestrowana jest w różnych odległościach od mikrofonu. W literaturze można znaleźć postulaty stosowania parametrów MFCC w procesie detekcji sygnału mowy. Kinnunen *et al.* [67] zaproponowali użycie 12 współczynników MFCC oraz ich pierwszej i drugiej różnicy symetrycznej aproksymującej pochodną, w połączeniu z klasyfikatorem SVM. Stosując swoją metodę uzyskali skuteczności detekcji zbliżone lub wyższe od standardowych algorytmów VAD.

Na rys 2.7 przedstawiono schemat blokowy algorytmu obliczanie MFCC [27] [105].

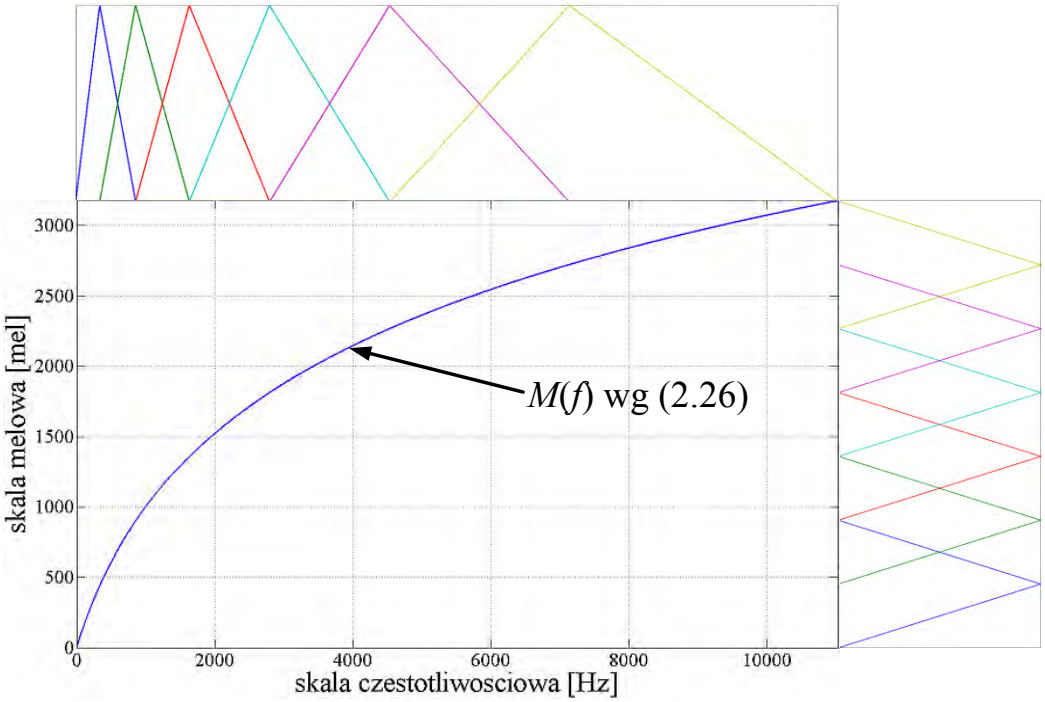


Rys. 2.7 Schemat blokowy algorytmu obliczanie MFCC.

W pierwszym kroku algorytmu obliczany kwadrat widma amplitudowego sygnału. Typowa długość stosowanej ramki wynosi 25 ms a przesunięcie 10 ms. W celu adaptacji rozdzielczości częstotliwości do charakterystyki ludzkiego słuchu w kolejnym kroku przetwarzania wykonuje się odkształcaniem widma amplitudowego z liniowego do rozkładu zgodnego ze skalą melową. Wykonywane jest to zgodnie ze wzorem:

$$M(f) = 2595 \log_{10}(1 + \frac{f}{700}) \tag{2.26}$$

gdzie $M(f)$ oznacza wartość w skali melowej, a f częstotliwość w skali liniowej. W kolejnym kroku obliczany jest splot kwadratu widma amplitudowego z charakterystykami częstotliwościowymi banku filtrów melowych. Typowo stosuje się bank filtrów składający się z 20 do 50 filtrów. Filtry te są o trójkątnej charakterystyce amplitudowej. Częstotliwości środkowe filtrów są równomiernie rozłożone w skali melowej (szerokości pasm wszystkich filtrów są takie same). W celu uproszczenia obliczeń zamiast odkształcenia widma amplitudowego, wykonywanego przed splotem z bankiem filtrów, stosuje się alternatywnie odkształcenie banku filtrów, co pozwala na pominięcie drugiego kroku algorytmu. Na rys. 2.8 przedstawiono bank filtrów w skali melowej i skali liniowej.



Rys. 2.8 Bank filtrów w skali melowej i w skali liniowej.

W kolejnym kroku obliczany jest logarytm widm odpowiadających kolejnym filtrom banku. W ostatnim kroku wyznaczana jest transformacja kosinusowa każdego z pasm zgodnie ze wzorem:

$$c[j] = \sum_{k=1}^{K} |X(j,k)|^2 \cos(j\frac{\pi}{K}(k-\frac{1}{2})) \tag{2.27}$$

gdzie $X(j,k)$ jest widmem odpowiadającym j-temu filtrowi, $c[j]$ jest j-tym współczynnikiem MFCC, a $j = 1, 2, \ldots, J$.

W celu opisu współczynników MFCC za pomocą jednego parametru, który mógłby być wykorzystany w algorytmie VAD, Skorik i Berthommier [169] zaproponowali trzy parametry opisujące zmienność cepstrum:

$$V_1 = \sum_{j=1}^{8} wh[j]|c[j]| \tag{2.28}$$

$$V_2 = \sqrt{\sum_{j=1}^{8} wh[j]^2 c[j]^2} \tag{2.29}$$

$$V_{2N} = \sqrt{\sum_{j=1}^{8} wh[j]^2 (c[j] - \overline{c_N[j]})^2} \tag{2.30}$$

gdzie V_1, V_2, V_{2N} są parametrami opisującymi zmienność cepstrum wewnątrz ramki sygnału, $wh[j]$ to waga j-tego współczynnika MFCC, a $c_N[j]$ jest wartością j-tego uśrednionego współczynnika MFCC samego szumu. Ponieważ parametry V_{2N} uwzględniają uśrednione wartości MFCC dla szumu, może on być stosowany jedynie dla szumu stacjonarnego.

2.2.2. Metody podejmowania decyzji

Adaptacyjne algorytmy progowe

Algorytmy progowe, ze względu na niewielką złożoność obliczeniową oraz wynikający bezpośrednio z wizualnej oceny zmienności analizowanych parametrów sposób podejmowania decyzji, są często stosowane w procesie detekcji sygnału mowy. Uzupełnienie algorytmu podejmowanie decyzji o adaptację progu pozwala (w ograniczonym zakresie) na pracę algorytmu VAD w zmiennych warunkach akustycznych. Podstawowym założeniem detekcji sygnału mowy z wykorzystaniem wartości progowej jest to, iż fragmenty sygnału, dla których wartość analizowanego parametru W przekracza wartość progu W_{th}, uznawane są za fragmenty zawierające mowę, a pozostałe ramki sygnału uznaje się za ramki zawierające szum. Typową wartością początkową progu W_{th} jest średnia obliczona dla pierwszych N ramek sygnału. Wtedy zakłada się, iż początek sygnału nie zawiera mowy, i jest w pełni reprezentatywny (jeżeli chodzi o charakterystykę szumu). Dodatkowo do wartości średniej dodaje się pewną stałą wartość, która przesuwa progu tak by niewielkie odchylenia wartości parametru od wartości progu nie powodowały fałszywych alarmów. Typowo przesunięcia progu obliczane jest jako k-krotnie zwiększona wartość odchylenia standardowego parametru lub wariancja obliczona dla N pierwszych ramek sygnału [149].

Adaptacja progu odbywa się dla ramek niezawierających sygnału mowy. W literaturze można znaleźć kilka propozycji adaptacji progu. Najpopularniejszym rozwiązaniem jest wykorzystanie dolnoprzepustowego filtru pierwszego rzędu zdefiniowanego za pomocą wzoru :

$$W_{\text{th}}^{n} = (1-a)W_{\text{th}} + aW_{\text{noise}} \tag{2.31}$$

gdzie parametr a jest liczbą z przedziału (0,1) i określa, jaki wpływ na nową wartość progu W^{n}_{th} będzie miała nowa wartość parametru W_{noise} wyznaczona dla ramki sygnału zawierającego szum. Typowo a przyjmuje wartość równą 0,2 [159]. W celu większego uodpornienia adaptacji progu na zmiany warunków akustycznych Sangwan *et al.* [159]

zaproponowali zastosowanie dodatkowo adaptacji wartości parametru a w zależności od wariancji wartości parametrów odpowiadających fragmentom sygnału zawierającego jedynie szum. Zasugerowali oni, iż wraz ze wzrostem wariancji wartości W_{noise} (wyznaczonej dla 10 ostatnich ramek szumowych), parametr a powinien przyjmować wyższe wartości mieszczące się w przedziale 0,1 do 0,25.

Innym rozwiązaniem jest wykorzystanie filtrów wyższego rzędu w celu spowolnienia procesu dopasowywania wartości progu do chwilowych wartości analizowanego parametru. Filtry te definiuje się za pomocą wzorów :

$$W_{\text{th}}^{n} = \frac{(1-a^2)W_{\text{th}} + (1-a)W_{\text{noise-1}} + aW_{\text{noise}}}{(2-a^2)} \tag{2.32}$$

$$W_{\text{th}}^{n} = \frac{(1-a^3)W_{\text{th}} + (1-a^2)W_{\text{noise-2}} + (1-a)W_{\text{noise-1}} + aW_{\text{noise}}}{(3-a^3-a^2)} \tag{2.33}$$

gdzie parametry $W_{\text{noise–1}}$ i $W_{\text{noise–2}}$ oznaczają odpowiednio wartości parametru dla ramki zawierającej szum i znajdującej się jedną i dwie ramki wcześniejszej od obecnej ramki zawierającej szum. Jak pokazały badania, opisane przez Sangwan *et al.* [159], zastosowanie filtrów wyższego rzędu poprawia jakość mowy kodowanej z użyciem algorytmu VAD. W badaniach nie oceniono jednak procentowej skuteczności detekcji sygnału mowy.

W swojej pracy Wu i Wang [187] wykorzystują metody adaptacji progu proponowane przez Gerven i Xie [47]. Algorytm oparty na tej metodzie podejmowania decyzji uzyskał wysoką skuteczność detekcji mowy, jak również małą liczbę błędów [187]. Do wyznaczenia wartości progu W_{th} używa się tu funkcji zdefiniowanej za pomocą wzoru (2.34):

$$W_{\text{th}}^{n} = \mu(W[m]) + b \cdot \sigma(W[m]), m = 0,1,...,M \tag{2.34}$$

gdzie $\mu(W[m])$ jest wartością średnią parametru dla pierwszych M ramek sygnału, $\sigma^2(W[m])$ jest wariancją parametru $\underline{W[m]}$, a b współczynnikiem dopasowującym, który jest dobierany eksperymentalnie. Adaptacja progu odbywa się poprzez wykorzystanie filtru dolnoprzepustowego pierwszego rzędu zdefiniowanego za pomocą wzoru (2.31). Adaptacji poddana jest wartości średnia sygnału szumu oraz wartości wariancji. Operację tą można wyrazić za pomocą wzoru (2.35):

$$\sigma(W[m])_{\text{szum}} = \sqrt{\left|\left(a \cdot \mu(W[m])^2 + (1-a) \cdot W^2\right) - \mu(W[m])_{\text{szum}}{}^2\right|} \tag{2.35}$$

gdzie $\mu(W[m])_{\text{szum}}$ oznacza zaktualizowaną wartość średnią parametru odpowiadającego szumowi, a $\sigma^2(W[m])_{\text{szum}}$ zaktualizowaną wartość wariancji parametru odpowiadającego sygnałowi szumowemu.

Algorytmy oparte na metodzie SVM

Podejmowanie decyzji na podstawie analizy jednego parametru, niezależnie od tego jak mocno separuje on dwie rozpoznawane klasy, nie pozwala na pełne uwzględnienie różnic pomiędzy klasami i uzyskanie bardzo wysokiej skuteczności detekcji w dowolnych warunkach akustycznych. Rozwiązania uwzględniające więcej niż jeden parametr, często bazują na metodach sztucznej inteligencji. Najczęściej w literaturze spotyka się algorytmy VAD oparte na metodzie SVM, która zapewnia wysoką skuteczność klasyfikacji.

Metoda SVM została opracowana przez Vapnika [176] w celu statystycznego podziału wektorów danych na dwa zbiory. Zbiór parametrów $\mathbf{x} = [x_1, x_2, \ldots, x_n]$, poddawany jest podziałowi na dwie klasy oznaczane jako **y**, gdzie **y** przyjmuje wartości $\{-1, 1\}$. Podział dokonywany jest za pomocą liniowej funkcji zdefiniowanej za pomocą wzoru:

$$f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + o \tag{2.36}$$

Płaszczyzna wyznaczana przez równanie $f(\boldsymbol{x}) = 0$, nazywana jest hiperpłaszczyzną. Trening klasyfikatora SVM polega na znalezieniu zbioru parametrów **w** minimalizujących margines, rozumiany jako odległość wektorów należących do każdej z klas do powierzchni hiperpłaszczyzny. Znalezienie minimum sprowadza się do znalezienia minimum zdefiniowanego za pomocą wzoru (2.37) z uwzględnieniem warunków (2.38) i (2.39):

$$\min \frac{1}{2}\|\mathbf{w}\|^2 + O\sum_i^l \zeta_i \tag{2.37}$$

$$y_i \cdot \left((\mathbf{w} \cdot x_i) + o\right) \geq 1 - \zeta_i, \forall_i \in U \tag{2.38}$$

$$\zeta_i \geq 0, \forall_i \in U \tag{2.39}$$

gdzie $\|\mathbf{w}\|$ oznacza długość wektora **w**, O jest współczynnikiem kosztu (ang. *cost*), a ζ_i parametrem „zwisu" (ang. *slack variable*) wyznaczanym niezależnie dla każdego wektora x_i. Uwzględnienie parametrów kosztu oraz zwisu pozwala na podział zbiorów nieseparowalnych linowo. Wartość kosztu ustalana jest przez użytkownika podczas treningu klasyfikatora. Im większą wartość przyjmuje parametr O, tym mniejszą wartość

przyjmuje margines, co może powodować zbytnie dopasowanie klasyfikatora do danych treningowych.

W celu rozszerzenia metody SVM, Vapnik zaproponował zastosowanie funkcji mapującej $\phi(\mathbf{x})$ [177]. Po jej uwzględnieniu funkcja separująca przyjmuje następującą postać (2.40):

$$f(\mathbf{x}) = \mathbf{w} \cdot \phi(\mathbf{x}) + b \tag{2.40}$$

gdzie $\phi(\mathbf{x})$ oznacza funkcję mapującą. Funkcja przenosi problem klasyfikacji z przestrzeni R^n do przestrzeni R^m, gdzie $n < m$. Wyznaczenie hiperpłaszczyzny minimalizującej margines wymaga rozwiązanie nierówności (2.37) z uwzględnieniem warunków (2.39) i :

$$y_i \cdot \left(\left(\mathbf{w} \cdot \phi(x_i) \right) + b \right) \geq 1 - \zeta_i, \forall_i \in U \tag{2.41}$$

Dodatkowo, w celu uproszczenia obliczeń nie wyznacza się tu bezpośrednio funkcji mapujących a jedynie ich iloczyn $K(\mathbf{x},\mathbf{z}) = \phi(\mathbf{x}) \cdot \phi(\mathbf{z})$. Rozwiązanie to nazywane jest ang. *kernel trick*, a funkcję $K(\mathbf{x},\mathbf{z})$ nazywa się funkcją jądra.

W procesie treningu konieczne jest wykorzystanie możliwie różnorodnych sygnałów w celu uzyskania klasyfikatora, będącego w stanie dokonywać klasyfikacji w różnorodnych warunkach akustycznych. Typowo podczas treningu klasyfikatora wykorzystuje się nagrania sztucznie zaszumione szumem o różnym poziomie oraz różnej charakterystyce widmowej.

2.2.3. Porównanie wybranych metod detekcji mowy

Ocena skuteczności algorytmów VAD może zostać wykonana w różny sposób. Jedną z popularnych metod jest wykorzystanie krzywej ROC (ang. *Receiver Operating Characteristics*). Krzywa ta przedstawia zależność pomiędzy skutecznością detekcji mowy (HR1) a błędem pierwszego rodzaju. Dzięki temu możliwa jest szczegółowa ocena klasyfikatora i znalezienie kompromisu pomiędzy skutecznością a liczbą fałszywych alarmów. Porównanie różnych metod detekcji opisanych w literaturze z wykorzystaniem krzywej ROC nie jest możliwe, ponieważ autorzy tych metod nie dostarczają wystarczających informacji, które bez implementacji danej metody pozwoliłyby na wykreślenie krzywej. Dlatego w tej części pracy, jako miary oceny jakość detektorów mowy wykorzystano wielkości liczbowe, które są najczęściej dostępne w literaturze. Są to:

- HR0 (ang. *non-speech hit ratio*) – skuteczność detekcji ciszy,
- HR1 (ang. *speech hit ratio*) – skuteczność detekcji mowy,

- FAR (ang. *false alarm rate*) – procentowa liczba błędów pierwszego i drugiego rodzaju.

Wykorzystane miary oceny jakości algorytmów można wyznaczyć korzystając ze wzorów:

$$\mathrm{HR0} = \frac{n^{d}_{\mathrm{cisza}}}{n^{o}_{\mathrm{cisza}}} \cdot 100\% \tag{2.42}$$

gdzie n^{d}_{cisza} jest liczbą ramek oznaczonych przez algorytm VAD jako cisza, n^{o}_{cisza} to liczbą ramek oznaczonych ręcznie jako cisza.

$$\mathrm{HR1} = \frac{n^{d}_{\mathrm{mowa}}}{n^{o}_{\mathrm{mowa}}} \cdot 100\% \tag{2.43}$$

gdzie n^{d}_{mowa} jest liczbą ramek oznaczonych przez algorytm VAD jako mowa, n^{o}_{mowa} to liczbą ramek oznaczonych ręcznie jako mowa.

$$\mathrm{FAR} = \frac{n^{b}_{\mathrm{cisza}} + n^{b}_{\mathrm{mowa}}}{n} \cdot 100\% \tag{2.44}$$

gdzie n^{b}_{cisza} jest liczbą ramek błędnie oznaczonych przez algorytm VAD jako cisza, n^{b}_{mowa} to liczba ramek błędnie oznaczonych przez algorytm VAD jako mowa, a n jest całkowitą liczbą ramek w analizowanym zbiorze.

W celu porównania znanych w literaturze metod VAD w tab. 2.2, 2.3 i 2.4 umieszczono opis sposobu podejmowania decyzji oraz skuteczność uzyskiwaną przez te algorytm. Wybrano algorytmy reprezentujące każdą z grup omówionych w podrozdziale 2.2. Jako że nie wszyscy autorzy oceniają skuteczność opisywanych metod w jednakowo szczegółowy sposób, wyniki podzielono pomiędzy dwie tabele. W tab. 2.3 przedstawiono szczegółowe wyniki skuteczności detekcji w różnych warunkach akustycznych (SNR oraz rodzaj szumu). W tab. 2.4 umieszczano uśrednione, dla różnych wartości SNR i różnych rodzajów szumu, wartości HR0, HR1 i FAR. Porównanie metod detekcji mowy nie jest proste z powodu niepełnych informacji dotyczących skuteczności oraz różnych baz wykorzystanych podczas przeprowadzania testów.

Porównując metody VAD będące częścią standardów, czyli G.729, AMR1 i AMR2, można zauważyć, iż najwyższe wartości HR1 uzyskuje algorytm AMR2. Dodatkowo wartość HR0 dla szumu typu *babble noise* jest bardzo niska dla wszystkich ustandaryzowanych algorytmów. Algorytm zaproponowany przez Sohna [170] uzyskuje wyższe wartości HR0, ale wynik ten został osiągnięty poprzez analizę niewielkiej bazy

zawierającej jedynie 46 s nagrań. Algorytmy bazujące na jednym parametrze (LSED [149], ALED [149], Yoo [195], Shen [167]), pozwalają na uzyskanie niższych skuteczności, niż w metodach analizujących wiele różnych wartości. Wartości HR0 oraz HR1 wyższe niż dla oryginalnego algorytmu uzyskał Chen [14] poprzez zastosowanie klasyfikator SVM i parametrów wykorzystywanych w algorytmach AMR1 i AMR2. Wynik ten wskazuje, iż zastosowanie metod inteligentnego podejmowania decyzji, pozwala klasyfikować sygnał z wyższą skutecznością, niż poprzez zastosowanie metod progowych. Jednak z powodu konieczności treningu klasyfikatora oraz w związku z koniecznością pracy w różnych warunkach akustycznych utworzenie VAD opartego na tych metodach jest kłopotliwe i często niepraktyczne.

Tab. 2.2 Porównanie wykorzystywanych parametrów oraz metod podejmowania decyzji w standardowych algorytmach VAD.

Autor/Nazwa metody	**Użyte parametry**	**Metoda podejmowania decyzji**
G.729 annex B [61]	*ZCR*;*E*;E_b;$R_p(l)$;*parametry LPC* i dyferencjał *E*, E_b, *ZCR* i parametrów *LPC*	wieloprogowa analiza zmienności parametrów, wygładzanie decyzji
ETSI AMR 1 [43]	*P,* E_b (9 pasm), częstotliwość podstawowa, składowe tonalne,	adaptacyjna analiza zmienności parametrów, wygładzanie decyzji
ETSI AMR 2 [43]	estymata *E,* SNR długo i krótkookresowe, dźwięczność,	wieloprogowa, adaptacyjna analiza zmienności parametrów, wygładzanie decyzji
LSED [149]	E_b	adaptacyjna analiza zmienności parametru
ALED [149]	*E*	adaptacyjna analiza zmienności parametru
Yoo [195]	PVD	stała wartość progu
Sohn [170]	widmo amplitudowe	LRT(ang. *Likelihood Ratio Test*)
Nemer	HOS, LPC	metody statystyczne (prawdopodobieństwa)
Shen [167]	*H*	dwuprogowa analiza zmian wartości parametru
Chen [14]	parametry wykorzystywane w ETSI AMR	SVM
Baig [8]	wartości próbek sygnału	SVM

Tab. 2.3 Skuteczność detekcji algorytmów VAD znanych z literatury.

Autor/Nazwa metody	**SNR**	**Rodzaj szumu**						***Rodzaj bazy***
		Biały		***Babble noise***		***Volvo noise***		
		HR0	**HR1**	**HR0**	**HR1**	**HR0**	**HR1**	
G.729 annex B [134]	0	89,54	54,27	65,33	57,36	b.d.	b.d.	TIMIT
	10	89,33	77,47	65,33	77,47	b.d.	b.d.	
	20	94,52	91,28	75,48	92,04	b.d.	b.d.	
ETSI AMR 1 [132]	0	87,53	67,58	47,82	95,75	b.d.	b.d.	TIMIT
	10	92,4	93,45	72,27	97,26	b.d.	b.d.	
	15	95,65	95,78	84,52	97,15	b.d.	b.d.	
ETSI AMR 2 [132]	0	94,76	89,05	44,59	95,66	b.d.	b.d.	TIMIT
	10	93,23	97,95	45,49	99,7	b.d.	b.d.	
	15	93,00	98,79	57,09	99,55	b.d.	b.d.	
Nemer [124]	6	b.d.	b.d.	b.d.	b.d.	85,9	79,9	TIA [173]
	12	b.d.	b.d.	b.d.	b.d.	90,9	88,4	
	18	b.d.	b.d.	b.d.	b.d.	94,3	93,6	
Sohn [170]	5	98,66	84,58	76,82	93,04	95,16	97,3	Nagranie 46 s
	15	96,73	96,93	76,2	98,43	92,81	99,62	
	25	94,83	99,87	75,25	99,75	92,22	99,87	
Autor/Nazwa metody		**FAR**		**FAR**		**FAR**		***Rodzaj bazy***
Yoo [195]		2,5		6		3,3		TIMIT
Shen [195]		8		24,4		18,6		TIMIT

Tab. 2.4 Wartości FAR osiągane przez algorytmy VAD znane z literatury.

Autor/Nazwa metody	**HR0**	**HR1**	**FAR**	**Rodzaj bazy**
ALED [149]	b.d.	b.d.	~60	Dialog
LSED [149]	b.d.	b.d.	~15	Dialog
Chen [14]	96,6	99,4	?	100 zdań
Baig [8]	b.d.	b.d.	15,9	1h wypowiedź

2.3 Estymacja tempa wypowiedzi

Tempo mowy jest istotnym parametrem charakteryzującym daną wypowiedź. Zależny ono m. in. od prozodii, języka wypowiedzi, cech indywidualnych mówcy, płci mówcy, rodzaju mowy (mowa czytana/spontaniczna), wieku mówcy oraz stanu emocjonalnego. Informacja dotyczą tempa analizowanej mowy, często jest wykorzystywana w systemach detekcji i przetwarzania sygnału mowy. Jednym z najczęstszych zastosowań algorytmów estymacji ROS (ang. *Rate of Speech*) są systemy ASR [101] [119] [184] [141] [178] [9]. Tempo mowy wykorzystywanej podczas trenowania modeli HMM jest zazwyczaj stałe. Jak pokazują badania, systemy ASR uzyskują niższe skuteczności rozpoznawania mowy wypowiadanej w szybkim tempie [133] [168] [101]. Zastosowanie algorytmu estymacji ROS działającego niezależnie od procesu rozpoznawania mowy pozwala na modyfikację

parametrów modeli HMM i poprawę skuteczności rozpoznawania mowy [178] [101]. Inną grupą zastosowań algorytmów estymujących ROS są systemy automatycznego rozpoznawania języka (ang. *Automatic Language Identification* – ALI) [139]. Analiza przeprowadzona przez Pellegrino *et al.* [138] pokazała, iż różnice w średnich wartościach ROS pomiędzy językami są znaczące. Dodatkowo badania te dowiodły, iż poza różnicami występującymi pomiędzy różnymi języki, różnym tempem mowy charakteryzują się także mówcy męscy i żeńscy. Obserwacje te pozwoliły na opracowanie systemu automatycznego rozpoznawania płci mówcy oraz języka wypowiedzi. Tempo mowy pozwala także określić płynność wypowiedzi. Zależność ta została wykorzystana przez De Jong [29] podczas oceny mowy mówców wypowiadającego się w języku obcym.

Istnieje wiele definicji tempa mowy. Można je podzielić na dwie główne grupy. Pierwsza z nich nawiązuje do naturalnej strategii używanej przez człowieka w celu podziału mowy na różne kategorie ROS: szybka, normalna i wolna. Ten podział nie wynika bezpośrednio z analizy sygnału mowy i jest rozmyty (ang. *fuzzy*). Ludzie przypisują mowie tempo na podstawie analizy czynników takich jak prozodia, akcent czy częstość występowania poszczególnych elementów mowy.

Druga grupa metod opisu ROS opiera się na przypisaniu tempu mowy wartości liczbowej, bezpośrednio skorelowanych z sygnałem mowy. W literaturze można spotkać wiele propozycji ilościowego opisu tempa wypowiedzi. Do tej grupy definicji zaliczają się liczba: słów, sylab, akcentowanych sylab, głosek, samogłosek oraz różnych połączeń głosek (tj. spółgłoska – samogłoska; spółgłoska – spółgłoska – samogłoska, itp.). Wartości te normalizowane są względem długości wypowiedzi lub względem interwału czasowego dla którego zostały wyznaczone. Najczęściej stosowanymi opisami są: liczba sylab na sekundę (ang. *Syllables Per Second* – SPS) [119] [143] oraz liczba głosek na sekundę (ang. *Phones Per Second* – PPS) [143] [142]. Zheng *et al.* [199] [200] [201] postulowali uzupełnienie miary PPS poprzez uwzględnienie modelu prawdopodobieństw długości głosek. Motywacją był fakt, iż głoski mają naturalnie różne rozkłady prawdopodobieństw czasu ich trwania. Dlatego słowa złożone z głosek o naturalnie krótkim czasie trwania oceniane byłyby poprzez miarę SPS, jako szybkie, chociaż wcale nie musiałyby być odczuwane przez człowieka w ten sposób. Uzyskane przez Zhenga *et al.* wyniki świadczą, o tym, iż ich założenie było słuszne. Jednak wyznaczenie tempa mowy z uwzględnieniem czasu trwania głosek wymaga transkrypcji analizowanego nagrania lub informacji z systemu ASR. Dlatego takie rozwiązanie jest trudne do realizacji w systemie działającym

w czasie rzeczywistym. Spotyka się, także pewne uproszczenie miary SPS, w którym zamiast wyznaczania obszaru sylab stosuje się detekcję wystąpień samogłosek. W większości języków wystąpienie samogłoski wiąże się z wystąpieniem jednej sylaby. Dodatkowo każda sylaba składa się tylko z jednej samogłoski. W ten sposób wyznaczanie liczby samogłosek na sekundę (ang. *Vowels Per Second* – VPS) odpowiada wyznaczeniu liczby połączeń głosek, w których centrum znajdują się samogłoski. Tempo opisywane liczbą słów na sekundę (ang. *Words Per Second* – WPS) jest rzadziej stosowane, ponieważ wyznaczenie obszaru słowa wymaga analizy lingwistycznej, co wiąże się z konieczności przeprowadzenia ASR [184].

Z estymacją ROS wiążą się także pojęcia tempa *brutto* oraz *netto* [186] [143] [138] [139]. Wartość *brutto* odnosi się do tempa mowy wyznaczonej w obrębie całego nagrania. Oznacza to, że jeżeli wewnątrz wypowiedzi występują fragmenty ciszy, zająknięć czy głośnego wzdychania, są one także wliczane w czas trwania wypowiedzi. Do wyznaczenia wartości *netto* nie bierze się pod uwagę wyżej wymienionych przedziałów sygnału, wskutek czego obliczone wartości ROS są wyższe niż dla wartości *brutto*. Do estymacji wartości *netto* konieczne jest połączenie algorytmu estymacji ROS z algorytmami VAD oraz algorytmami detekcji zająknięć. Takie połączenie postulował Pellegrino *et al.* [139], uważając iż wartość *netto* ROS mocniej koreluje z odczuwalnym przez człowieka tempem mowy niż wartość *brutto*.

Dodatkowo należy zauważyć, iż zależnie od interwału czasowego wykorzystywanego podczas wyznaczania tempa, uzyskuje się różną rozdzielczość czasową estymacji. W literaturze proponuje się podział na trzy grupy ROS w odniesieniu do rozdzielczości czasowej: globalna (długoczasowa), lokalna (krótkoczasowa) i relatywna (porównawczą) [125] [142]. Estymacja globalna polega na wyznaczeniu tempa mowy w odniesieniu do całego fragmentu wypowiedzi (zdania, frazy, itp.). Globalne tempo nie oddaje charakterystyki tempa mowy związanej z prozodią, akcentami oraz zmiennością wewnątrz zdania. Niesie ono natomiast informację ogólną opisującą całą wypowiedź. Lokalne tempo mowy wyznaczanej jest poprzez estymację tempa wewnątrz krótkich ramek czasowych. Pfitzinger [142] zaobserwował, że optymalne wartości interwału pozwalające na dokładne odzwierciedlenie prozodii mowy, przy detekcji SPS, powinna być większa od długości najdłuższej samogłoski występujących w danym języku. Do ostatniej grupy należy relatywna estymacja tempa, uzyskiwana poprzez porównanie tempa mowy referencyjnej z tempem mowy nagrania analizowanego [125].

W dalszej części tego rozdziału opisano najważniejsze metody automatycznej estymacji ROS skupiając się na tych, które mogą zostać zmodyfikowane, tak by ich praca odbywa się w czasie rzeczywistym. W związku z faktem, iż parametrami najmocniej korelującym z odczuwanym przez człowieka tempem mowy są wartości SPS i VPS, opis metod ograniczono do tych opracowanych w celu estymacji tych parametrów. Opisane algorytmy bazują na kilku założeniach:

- energia wewnątrz sylaby jest znacznie większa niż energia głosek na jej końcu, ponieważ z każdą sylabą związany jest akcent oraz samogłoska,
- z uwagi na występowanie samogłosek sylaby wewnątrz swojego obszaru mają charakter dźwięczny,
- nie jest dostępna transkrypcja wypowiedzi,
- tempo mowy może zmieniać się dowolnie wewnątrz zdania.

2.3.1. Analiza głośności chwilowej sygnału

Jedną z pierwszych prac opisujących sposób pomiaru tempa mowy był referat opracowany przez Mermelsteina [100]. Zaproponowana tam metoda oparta była na analizie głośności chwilowej sygnału (ang. *loudness*), wyznaczanej w paśmie częstotliwości od 500 Hz do 3500 Hz. Dodatkowo amplituda chwilowa sygnału była poddawana dolnoprzepustowej filtracji w celu uzyskania gładkiej obwiedni. Analiza tempa dokonywana była poprzez znalezienie szczytów będących centrami sylab. Weryfikacja tego czy szczyt odpowiada środkowi sylaby, czy jest on tylko chwilową zmianą amplitudy sygnału wewnątrz sylaby, odbywała się z wykorzystaniem analizy powłoki wypukłej (ang. *convex hull*) funkcji reprezentującej obwiednię sygnału. Za powłokę wypukłą uznawana jest minimalna wartość obwiedni, która jest monotonicznie niemalejąca w obszarze od początku do wartości szczytowej występującej w analizowanym fragmencie sygnału, oraz jest monotonicznie nierosnąca w obszarze od szczytu do końca analizowanego fragmentu. Podział sygnału odbywał się poprzez progową analizę różnicy pomiędzy obwiednią sygnału a powłoką wypukłą.

Zmodyfikowana wersja metody zaproponowanej przez Mermelsteina została także wykorzystana przez Xie'a i Niyogia [190]. Wprowadzone zmiany polegały na uzupełnieniu analizy o algorytm wyznaczający okres podstawowy sygnału oraz na zastąpieniu funkcji obwiedni sygnału przebiegiem energii sygnału. Analiza dźwięczności pozwala na dodatkową weryfikację tego czy dany szczyt w przebiegu energii odpowiada

wystąpieniu sylaby w sygnale. Szczyty w przedziale, w którym nie występuje dźwięczny fragment sygnału, nie są oznaczane jako centra sylab. Analiza częstotliwości podstawowej wykorzystywana była także przez innych badaczy i wykazano jej istotny wpływ na poprawę skuteczności estymacji tempa mowy [29] [198].

Głównym ograniczenie metod bazujących na analizie powłoki wypukłej, jest konieczność analizy co najmniej 500 ms fragmentu sygnału. Dlatego nie jest możliwa implementacja estymatora ROS opartego na tej metodzie działającego w czasie rzeczywistym.

2.3.2. Analiza obwiedni energii sygnału

Kolejną grupą algorytmów są te, oparta na analizie obwiedni energii sygnału. Głównym reprezentantem tej grupy jest algorytm opracowany przez Morgana *et al.* [107]. Zasugerował on wykorzystanie parametru enrate (ang. *energy rate*), który reprezentuję obwiednie energii sygnału, a szczyty w przebiegu zmian tego parametru reprezentują centra wystąpień sylab. Dlatego może on zostać wykorzystany bezpośrednio do estymacji ROS. Wartość tego parametru wyznaczana jest zgodnie ze wzorem:

$$\text{enrate} = \frac{\sum_{k=s}^{K} k\left|Z(k)\right|^2}{\sum_{k=s}^{K}\left|Z(k)\right|^2} \tag{2.45}$$

gdzie $Z(k)$ jest widmem amplitudowym sygnału $z[n]$. Sygnał $z[n]$ to sygnał jednopołówkowy (ang. *half-wave rectified*), pasmowo ograniczony do 16 Hz i zdecymowany do prędkości próbkowania 100 Hz. Obliczenie tego parametru odbywa się w ramkach czasowych od długości od 1 do 2 sekund, przesuwanych z krokiem równym 75% długości ramki. Badania przeprowadzone przez Morgana *et al.* pokazały, że korelacji pomiędzy szczytami przebieg enrate i ręcznie wyznaczonych obszarami sylab jest niska i wynosi 0,4. Zaproponowane przez badaczy rozwiązanie mające na celu zwiększenie niezawodności metody, polegało na uzupełnieniu jej o analizę dodatkowych dwóch parametrów [106]. Pierwszym z nich jest liczba szczytów w obwiedni energii szerokopasmowej. Drugi parametr reprezentuje liczbę szczytów w sygnale powstałym poprzez obliczenie funkcji korelacji pomiędzy obwiedniami energii sygnału w podpasmach. Funkcja ta definiowana jest wzorem:

$$R[l] = \frac{1}{G} \sum_{g=1}^{H-1} \sum_{h=g+1}^{H} x_g[n] x_h[n] \tag{2.46}$$

gdzie $x_g[n]$ i $x_h[n]$ oznaczają obwiednie energii w podpasmach g oraz h, a G jest liczbą wszystkich kombinacji bez powtórzeń wyznaczoną dla H analizowanych podpasm. W wyniku uśrednienia wartości trzech wykorzystywanych parametrów uzyskiwana jest miara nazwana mrate (ang. *multiple rate estimator*).

Metody opracowane przez Morgana *et al.* [107] [106] zainspirowały Wanga i Narayanana [184] [119] do przeprowadzanie szczegółowej analizy wpływu doboru zmiennych użytych podczas wyznaczania oraz obliczania parametru mrate. Zaproponowane modyfikacje adresowane były do kluczowych problemów wynikających z wykorzystania miary mrate. Poniżej wymieniono główne zarzuty:

- zniekształcenia wprowadzane przez spółgłoski półotwarte powodujące wyrycie większej liczby sylab,
- rozmazywanie sąsiadujących ze sobą szczytów powodujące błąd wykrycia jednej zamiast dwóch sylab,
- zawyżenie liczby sylab poprzez dwukrotne policzenie tych, które zostały wydłużone przez mówcę,
- niska rozdzielczość czasowa analizy wynikająca z wykorzystywania długich ramek czasowych.

Badania przeprowadzone przez Wanga i Narayanana pozwoliły na opracowanie szeregu rozwiązań adresowanych tym problemom. Są to:

- sugestia analizy wyłącznie wybranych podpasm podczas wyznaczania funkcji korelacji. Dobór właściwych podpasm powinien prowadzić do uwydatnienia struktury formantowej samogłosek.
- zastosowanie badania dźwięczności w celu eliminacji błędu detekcji spółgłosek, jako samogłoski.
- obliczanie korelacji pomiędzy wektorami tworzonymi przez wartości energii w podpasmach uzyskane dla kolejnych ramek analizy (w wyniku tego procesu uzyskuje się wygładzenie sygnału korelacji $R(m)$).
- wykorzystanie okna wygładzającego przebieg analizowanego parametru.

- wygładzanie przebiegu poprzez zastosowanie filtru Gaussa.
- optymalizacja wartości parametrów z wykorzystaniem metody *Monte Carlo*.

Kolejne uzupełnienie opisanych w tej części pracy metod postulował Zhang i Glass [198]. Założył on, iż tempo mowy w obszarze zdania (frazy) nie podlega nagłym zmianom. Możliwe więc jest estymowanie jego wartości na podstawie wyników uzyskanych dla kilku pierwszych sylab wypowiedzi. Miejsce estymowanego wystąpienia kolejnej sylaby może posłużyć jako punkt startowy użyty do wyszukiwania centra kolejnej sylaby.

2.3.3. Metody oparte na algorytmach detekcji samogłosek i sztucznych sieciach neuronowych

Metody estymacji ROS oparte na analizie wystąpień samogłosek były rozwijane przez dwie grupy badaczy. Pfau i Ruske [141] zaproponowali detekcję samogłosek opartą na analizie zmodyfikowanej obwiedni amplitudowej sygnału wspartą parametrem ZCR (ang. *Zero Crossing Rate*). Natomiast Pellegrino *et al.* [139] dokonywał detekcji poprzez analizę energii sygnału w pasmach melowych. Szczegółowy opis tych metod zamieszczono w rozdziale 2.4 opisującym metody detekcji samogłosek. Na podstawie wykrytych miejsc występowania samogłosek w sygnale mowy, wyznaczali oni ROS korzystając z miary VPS.

Innym naturalnym rozwiązaniem wydaje się wykorzystanie metod sztucznej inteligencji w celu estymacji ROS. Jedną z pierwszych prac poświęconych temu podejściu jest referat napisany przez Verhasselta i Martensa [178], w którym przedstawiono metodę estymacji ROS opracowaną z wykorzystaniem sztucznych sieci neuronowych. Zadaniem ANR (ang. *Artificial Neural Network*) było znajdowanie granic przedziałów odpowiadających sylabom. Niestety, pomimo, iż metoda zaproponowana przez Verhasselta i Martensa jest wielokrotnie cytowana w innych pracach i podawana, jako przykład skutecznego estymatora ROS, w referacie Verhasselt i Martens nie opisali sposobu parametryzacji sygnału ani innych szczegółów dotyczących sposobu trenowania klasyfikatora.

2.3.4. Porównanie wybranych metod estymacji tempa wypowiedzi

Bezpośrednie porównanie przedstawionych metod estymacji ROS nie jest proste. Wiąże się to z faktem, że nie istnieje jednorodny system oceny skuteczności tych algorytmów. Każdy z badaczy wykorzystywał inna bazę oraz liczbę nagrań, a mowa wypowiadana jest w różnych językach. W tab. 2.5 zamieszczono wyniki przedstawiające

skuteczności estymacji ROS opublikowane w literaturze. Miary oceniające, jakość tych metod to:

- współczynnik korelacji Pearsona obliczony pomiędzy wartościami ROS uzyskanymi na podstawie ręcznego oznaczenia zawartości nagrania a wartościami estymowanego tempa wypowiedzi.
- skuteczność detekcji sylab/samogłosek, liczba błędów pierwszego i drugiego rodzaju.
- skuteczność zaklasyfikowania zdania do jednej z trzech klas tempa mowy: szybkie, średnie, wolne.

W pierwszej kolumnie tabeli literami od A do D oznaczono algorytmy bazujące na analizie tych samych parametrów. Można zauważyć, iż wewnątrz poszczególnych grup algorytmów wprowadzane przez badaczy modyfikacje powodowały wzrost skuteczność danej metody. Dodatkowo porównując skuteczności pomiędzy grupami algorytmów widać, iż najwyższe skuteczności estymacji uzyskiwano analizując liczbę wystąpień samogłosek oraz estymując liczbę sylab na podstawie obwiedni energii.

Tab. 2.5 Porównanie metod estymacji ROS.

<table>
<tr><th>Autor/Rok</th><th>Sposób oceny</th><th>Rodzaj bazy wykorzystanej do testów</th><th>Wyniki</th></tr>
<tr><td>(A)
Verhasselt / 1996 [178]</td><td>Wsp. korelacji SPS, odchylenie standardowe błędu estymacji</td><td>TIMIT(Angielski) (podzbiór)</td><td>Wsp. korelacji: nie podano
odchylenie standardowe błędu predykcji: 1,36 SPS</td></tr>
<tr><td>(B)
Morgan / 1998 [106]</td><td>korelacja SPS
skuteczność</td><td>Switchboard (Angielski) (podzbiór)</td><td>korelacja SPS w obrębie zdania: 0,671
skuteczność:
<table>
<tr><td></td><td>wolne</td><td>średnie</td><td>szybkie</td></tr>
<tr><td>wolne</td><td>57,9%</td><td>29,0%</td><td>13,1%</td></tr>
<tr><td>średnie</td><td>24,6%</td><td>43,8%</td><td>31,6%</td></tr>
<tr><td>szybkie</td><td>9,6%</td><td>32,4%</td><td>58,0%</td></tr>
</table></td></tr>
<tr><td>(C)
Pfau / 1998 [141]</td><td>korelacja VPS, skuteczność detekcji samogłosek</td><td>German Verbmobil (Niemiecki) (podzbiór)</td><td>korelacja VPS w obrębie zdania: 0,796
błąd drugiego rzędu: 22,72%</td></tr>
<tr><td>(C)
Pellegrino / 2004 [139]</td><td>korelacja VPS</td><td>OGI MLTS [114]</td><td>Korelacja VPS w obrębie zdania dla różnych języków:
Angielski: 0,82
Niemiecki:0,73
Hinduski: 0,91
Japoński: 0,88
Chiński: 0,88
Hiszpański: 0,84</td></tr>
<tr><td>(D)
Xie / 2006 [190]</td><td>skuteczność detekcji sylab</td><td>TIMIT(Angielski)</td><td>skuteczność: 81,6%
błąd pierwszego rodzaju: 10,9%
błąd drugiego rodzaju: 18,4%</td></tr>
<tr><td>(B)
Wang / 2007 [184]</td><td>korelacja SPS, skuteczność detekcji sylab</td><td>ICSI Switchboard (podzbiór) (Angielski)</td><td>korelacja SPS w obrębie zdania: 0,745
skuteczność: 80,6%
błąd pierwszego rodzaju: 3,8%
błąd drugiego rodzaju: 15,6%</td></tr>
<tr><td>(B)
Wang / 2009 [198]</td><td>skuteczność detekcji sylab</td><td>TIMIT(Angielski)</td><td>błąd drugiego rodzaju: 0,8606</td></tr>
<tr><td>(B)
Zhang / 2009 [198]</td><td>skuteczność detekcji sylab</td><td>TIMIT(Angielski)</td><td>błąd drugiego rodzaju: 0,8659</td></tr>
<tr><td>(D)
De Jong / 2009 [29]</td><td>korelacja SPS</td><td>WISP (język Duński)
IFA (język Duński)</td><td>WISP:
korelacja SPS w obrębie frazy:0,71
korelacja SPS dla mówców: 0,88
IFA:
korelacja SPS w obrębie frazy: 0,77
korelacja SPS dla mówców: 0,8</td></tr>
</table>

2.4 Detekcja samogłosek

Detekcja samogłosek nie jest tematem często poruszanym w literaturze. Zagadnienie to jest mocno związany z dobrze znanymi tematami detekcji dźwięczności [12] [151] [7] [92] [163] oraz segmentacji mowy [26] [110] [148] [166] [202]. W niektórych zastosowaniach

sama informacja dotycząca tego, czy dana głoska jest dźwięczna czy też nie, nie jest wystarczająca. Za przykład mogą posłużyć algorytmy estymacji ROS opisane w rozdziale 2.4. W przypadku oceny tempa mowy zastąpienie detektora samogłosek detektorem dźwięczności prowadziłoby do uzyskiwania zawyżonych wartości ROS, ponieważ dźwięczne są nie tylko samogłoski, ale też niektóre spółgłoski. Uwzględnienie dźwięcznych spółgłosek podczas estymacji ROS, zwiększałoby liczbę wykrytych przez algorytm sylab. Innym zastosowaniem algorytmów detekcji samogłosek jest jej wykorzystaniem w procesie nierównomiernej modyfikacji czasu trwania sygnału mowy opisanej w rozdziale 2.1 oraz w systemach automatycznego rozpoznawaniu emocji [155] [154].

Algorytmy detekcji samogłosek można podzielić na cztery kategorie. Są to algorytmy:

a. detekcji początku samogłoski (ang. *Vowel Onset Point* – VOP*)* – zadaniem algorytmów z tej grupy jest wykrycie wystąpienia samogłoski w sygnale mowy i oznaczenie jej początku [95] [94],

b. detekcji wystąpienia samogłoski (ang. *Vowel Detection* – VD) – wykrywają wystąpienie samogłoski a nie jej obszar [141],

c. detekcji przedziału samogłoski (ang. *Vowel Landmark Detection* – VLD) – algorytmy wykrywają wystąpienie samogłoski w sygnale mowy oraz oznaczają obszar odpowiadający obszarowi maksymalnej energii pierwszego formantu samogłoski [56],

d. segmentacji przedziału samogłoski (ang. *Vowel Region Detection* – VRD) – wiąże się ze znalezieniem samogłoski w sygnale mowy oraz z wykryciem całego obszaru sygnału reprezentującego daną samogłoskę [136].

W tym podrozdziale przedstawiono najistotniejsze metody detekcji samogłosek opisane w literaturze. Z punktu widzenia możliwości bezpośredniego zastosowania wyniku klasyfikacji w procesie nierównomiernej TSM sygnału mowy, najistotniejsze są metody z grupy VRD. Algorytmy z grup a–c mogą być wykorzystane m.in. podczas estymację tempa mowy, ale nie niosą wystarczającej informacji, by umożliwić bezpośrednią współpracę z algorytmami modyfikującymi czas trwania sygnału.

W większości metod detekcji samogłosek zakłada się, że przed analizą sygnału dokonywana jest detekcja mowy, a znajdowanie samogłosek odbywa się wyłącznie w miejscach wystąpień sygnału mowy. Do wykrycia sygnału mowy wykorzystuje się

algorytmy z grupy VAD opisane w rozdziale 2.2 rozprawy. Autorowi nie są znane metody dokonujące detekcji samogłosek w czasie rzeczywistym. Opisane poniżej algorytmy, po zastosowaniu pewnych ograniczeń i modyfikacji mogą zostać zaadaptowane do pracy w czasie rzeczywistym.

2.4.1. Analiza energii w pasmach melowych

Pellegrino *et al.* [136] [137] [138] [139] [135] [156] opisali w swoich pracach możliwość wykorzystania analizy energii sygnału w celu detekcji wystąpień samogłosek w sygnale mowy. Metoda przez nich zaproponowana była wielokrotnie wykorzystywana, także przez innych badaczy [52] [155] [154] [1]. Założeniem zaproponowanego sposoby segmentacji sygnału na samogłoski i spółgłoski był fakt, iż energia samogłosek w paśmie poniżej 3500 Hz jest wyższa niż dla spółgłosek. Pellegrino *et al.* zaproponowali dwa parametry oparte na analizie energii sygnału w 24 pasmach melowych. Pierwszy z parametrów, nazwany SBEC (ang. *Spectral Band Energy Cumulating*), zdefiniowany jest za pomocą wzoru [137].

$$\mathrm{SBEC}_m = \sum_{j=1}^{24} wh_j \left| E_m^j - \overline{E_m} \right| \tag{2.47}$$

gdzie $SBEC_m$ jest wartością parametru w m-tej ramce sygnału, E_m^j jest wartością energii za czas ramki sygnału w j-tym filtrze melowym, $\overline{E}_m$ to wartość średnia m-tej ramki wyznaczona dla wszystkich energii obliczonych dla 24 pasm melowych, a wh_j jest wartością wagi każdego pasma. Autorzy nie zdefiniowali wartości wag filtrów. Rozsądnym wydaje się wybór wag taki, który przypisywałby wyższe wagi filtrom odpowiadającym częstotliwościom formantowym samogłosek. Segmentacja sygnału z wykorzystaniem parametru SBEC polega na znalezieniu szczytów w przebiegu $SBEC_m$. Wystąpienia szczytu w przebiegu SBEC_m jest tożsame z chwilą w sygnale odpowiadającym samogłosce. Wyznaczenie granic samogłosek wykonywane jest równolegle poprzez zastosowanie algorytmu segmentacji sygnału mowy nazwanego *Forward-Backward Divergence* (FBD) [4]. Segmenty w obszarze, których wykryto maksimum przebiegu $SBEC_m$ oznaczane są jako zawierające samogłoskę.

Głównym problemem sygnalizowanym przez autorów tego algorytmu było błędne klasyfikowani głosek bezdźwięcznych jako samogłoski, ponieważ w ich obszarze występowały szczyty w przebiegu $SBEC_m$. Zaproponowane przez autorów modyfikacje wiązały się z wprowadzeniem dodatkowego podziału energii na pasmo niskie E_{LF}^m (100–

1000 Hz), uściśleniem wartości wag filtrów melowych oraz określeniem minimalnego czasu trwania maksimum. Zalecane wartości wag są binarne i przyjmują wartość 1 w przypadku pasm odpowiadającym filtrom o częstotliwościach z przedziału 300–3200 Hz oraz wartość 0 dla wszystkich pozostałych pasm. Zasugerowany minimalny czas trwania szczytu wynosi 15 ms. Parametr uwzględniający powyższe zmiany nazwano REC (ang. *Reduced Energy Cumulating*) i opisano za pomocą wzoru [136]:

$$\mathrm{REC}_m = \frac{E_{LF}^m}{E_m} \sum_{j=1}^{24} wh_j \left| E_m^j - \overline{E_m} \right| \tag{2.48}$$

gdzie REC_m jest wartością parametru obliczoną w m-tej ramce sygnału, a E_m to energia w m-tej ramce sygnału. Jedną z głównych zalet segmentacji z wykorzystaniem parametru REC jest wykazana przez autorów niezależność od języka wypowiedzi [139]. Problem natomiast stanowi konieczność segmentacji za pomocą algorytmów FBD, która nie może zostać wykonana w czasie rzeczywistym.

2.4.2. Analiza zmodyfikowanej obwiedni amplitudowej sygnału

Pfau i Ruske [141], na podstawie założenia dotyczącego mówiącego o wysokiej energii sygnału samogłosek w niskich pasmach częstotliwości, przedstawił metodę detekcji samogłosek wykorzystującą analizę zmodyfikowanej obwiedni amplitudowej sygnału $N_m(t)$ (ang. *modified loudness*). Zdefiniował ją za pomocą wzorów (2.49) i (2.50), jako różnicę obwiedni amplitudowej wyznaczoną w niskich i wysokich pasmach krytycznych:

$$d(t) = \sum_{v=3}^{15} N_v(t) - \sum_{v=20}^{22} N_v(t) \tag{2.49}$$

$$N_m(t) = \begin{cases} d(t), & d(t) > 0 \\ 0, & d(t) \le 0 \end{cases} \tag{2.50}$$

gdzie v jest numerem pasma krytycznego, a $N_v(t)$ oznacza obwiednię amplitudową v-tego pasma krytycznego. Podczas detekcji samogłosek wykorzystuje się wygładzoną funkcję $N_m(t)$. Wygładzanie wykonywane jest poprzez zastosowanie filtru dolnoprzepustowego. Podobnie jak podczas analizy energii w pasmach melowych, miejsca wystąpień samogłosek znajdowane są jako szczyty w przebiegu $N_m(t)$. Za miejsca wystąpień samogłosek uznawane są tylko te szczyty, w których okolicy przynajmniej jedno ze zboczy przebiegu $N_m(t)$ (poprzedzające szczyt lub występujące po nim) w ograniczonym czasie d_t opadają poniżej adaptowanego progu T_r. Próg T_r jest to procentowa wartość analizowanego aktualnie szczytu. W celu zmniejszenia liczby błędów drugiego rodzaju,

zaproponowano wykorzystanie dodatkowego parametru pozwalającego na weryfikację tego, czy szczyt rzeczywiście reprezentuje samogłoskę. Do tego celu użyto parametru ZCR, którego wartość jest wyższa w obszarach spółgłosek niż samogłosek.

2.4.3. Metody detekcji początku samogłoski

Detekcja początku samogłoski może służyć, jako podstawa lub uzupełnienie opisanych w tym podrozdziale metod detekcji samogłosek. Szeroko zakrojone badania związane z detekcją VOP prowadził Mahadeva Prasanna [95] [94]. W swoich pracach przedstawił on możliwość detekcji VOP z wykorzystaniem analizy różnych parametrów.

W pierwszej metodzie zaproponował on wykorzystanie analizy energii sygnału będącego źródłem pobudzeniem traktu głosowego (ang. *source excitation*). Sygnał pobudzenia wyznaczany jest za pomocą analizy LPC. Przetwarzanie odbywa się w ramkach czasowych o długości 20 ms, a ramki przesuwane są z korkiem 10 ms. Dla każdej ramki wyznaczane jest 10 współczynników LPC. Predykcja wartości kolejnych próbek wykonywana jest zgodnie ze wzorem:

$$\hat{x}[n] = -\sum_{u=1}^{U} a_u x[n-u] \tag{2.51}$$

gdzie a_u jest wartością k-tego współczynnika LPC, a U jest rzędem analizy. Jako że współczynniki LPC niosą informację związaną z traktem głosowym, filtracja sygnału oryginalnego za pomocą filtru odwrotnego zdefiniowanego za pomocą wzoru (2.52) pozwala na uzyskanie sygnału pobudzającego trakt głosowy $e[n]$:

$$A(z) = 1 + \sum_{u=1}^{U} a_u z^{-u} \tag{2.52}$$

Mahadeva Prasanna postulował przeprowadzenie analizy obwiedni Hilberta [95] wyznaczonej dla sygnału pobudzenia. Sygnał analityczny związany z sygnałem rzeczywistym $e[n]$ zdefiniowany jest za pomocą wzoru:

$$y[n] = e[n] + je_h[n] \tag{2.53}$$

gdzie $e_h[n]$ jest transformatą Hilberta sygnału $e[n]$. Obwiednia Hilberta zdefiniowana jest w następujący sposób:

$$h[n] = \sqrt{e^2[n] + e_h^2[n]} \tag{2.54}$$

Gwałtowne zmiany obwiedni Hilberta odpowiadają miejscom występowania VOP.

W drugiej metodzie, nazwanej metodą analizy traktu głosowego, zasugerował on zastosowanie uproszczonej analizy formantów. Jako że energia formantów jest ściśle związana ze szczytami w widmie amplitudowym sygnału mowy, postulował on obliczanie sumy amplitud pierwszych 10 szczytów bieżącego widma amplitudowego, jako pewnego rodzaju odwzorowanie opisujące kształt traktu głosowego.

Analiza zmian obydwu parametrów wymaga jednak wykonania tak zwanego wykresu dokumentującego (ang. *evidence plot*). Wykres ten przedstawiania wygładzone wartości parametru. W celu uwypuklenia szczytów i dolin w obwiedni sygnału wszystkie wartości szczytów normalizowane są do wartości 1 a doliny do wartości –1. Powoduje to uwypuklenie nawet niewielkich szczytów, co znacząco uprasza algorytm detekcji. W ostatnim kroku ten zmodyfikowany przebieg splatany jest z odpowiedzią impulsową gaussowskiego cyfrowego filtru różniczkującego, którego długość wynosi 100 ms. Konieczność wykonania wykresu dokumentującego uniemożliwia implementację tych rozwiązań w czasie rzeczywistym.

2.4.4. Porównanie metod detekcji samogłosek

Główną miarą pozwalającą na ocenę skuteczności detekcji samogłosek jest zdefiniowana za pomocą wzoru (2.55) liczba błędów detekcji (ang. *Vowel Error Rate* – VER):

$$\mathrm{VER} = \left(\frac{n^{b}_{\mathrm{spół}} + n^{b}_{\mathrm{sam}}}{n^{o}_{\mathrm{sam}}} \right) \cdot 100 \tag{2.55}$$

gdzie $n^{b}_{\mathrm{spół}}$ jest liczbą niewykrytych samogłosek, n^{b}_{sam} to liczba samogłosek wykrytych w miejscu spółgłosek, a n^{o}_{sam} oznacza całkowitą liczbę samogłosek w analizowanym zbiorze.

W tab. 2.6 przedstawiono wartości VER osiągane przez algorytmy opisane w literaturze. Wyniki te nie mogą zostać porównane bezpośrednio, ponieważ uzyskano je dla różnych języków, różnych baz wypowiedzi, a algorytmy miały na celu detekcję zdarzeń akustycznych zdefiniowanych w różny sposób. W pierwszej kolumnie tabeli zamieszczono skrót reprezentujący kategorię algorytmu. Wyniki opisane przez Ringeval [154] zostały uzyskane z wykorzystaniem metody zaproponowanej przez Pellegrino [137]. Detekcja VOP zakładała możliwość błędnej klasyfikacji początku samogłoski w przedziale ± 40 ms. Dodatkowo skuteczność detekcji VOP została przedstawiona dla dwóch różnych baz nagrań. Pierwsza z baz zawierała krótkie wypowiedzi, druga zawierała wypowiedzi składające się z dwóch zdań.

Najmniej błędów uzyskano za pomocą metody wykrywającej miejsca VOP dla krótkich zdań. Jednak dla długich wypowiedzi skuteczność detekcji jest znacznie niższa. Algorytmy wykrywające obszar samogłosek (VRD) w zależności od języka oraz bazy wykorzystanej podczas testów, dokonują detekcji samogłosek z błędami w przedziale od 16,3% (Japoński) do 29,08%(Niemiecki). Algorytm zaproponowany przez Pfau [141] pozwala wykrywać samogłoski z wyższą dokładnością niż algorytm opracowany przez Pellegrino [137] (porównując skuteczność dla języka Niemieckiego).

Tab. 2.6 Porównanie wartości błędu VER uzyskanych dla różnym metod detekcji samogłosek

Rodzaj algorytmu	**Autor/Rok**	**Język**	**Rodzaj bazy wykorzystanej do testów**	**VER [%]**
VRD	Pellegrino [137]	Francuski	OGI MLTS	19,5
		Japoński		16,3
		Koreański		28,5
		Hiszpański		19,2
		Wietnamski		31,1
		Średnia		22,9
VRD	Ringeval [154]	Niemiecki	Berlin	29,08
		Angielski	TIMIT	19,5
		Angielski	NTIMIT	24,07
		Baskijski	Aholab	24,28
VD	Pfau [141]	Niemiecki	Verbimobil	22,72
VOP	Mahadeva Prasanna [94]	Angielski	TIMIT – krótkie zadania	6,92[7]
				8,8[8]
VOP	Prasanna [94]	Angielski	TIMIT – długie zdania	21,39[7]
				25,8[8]

[7] Wynik uzyskany dla algorytmy analizującego sygnał pobudzający trakt głosowy
[8] Wynik uzyskany dla algorytmy uproszczonej analizy formantów

3 Opracowane metody analizy i modyfikacji sygnału mowy

W rozdziale tym przedstawiono metody modyfikacji czasu trwania sygnału, opracowane w ramach niniejszej rozprawy. Zostały one zaprojektowane tak, by umożliwiały modyfikację w czasie rzeczywistym mowy rejestrowanej przez mikrofon znajdujący się blisko ust mówcy lub mowy emitowanej przez różnego rodzaju urządzenia (np. telefon komórkowy czy telewizor). Zaproponowano dwa rozwiązania różniące się sposobem dobierania chwilowych wartości współczynnik skali α – metoda B i C. W metodzie B chwilowa wartość α zależna jest od określonej przez użytkownika wartości oczekiwanego współczynnika skali α_{de}. W metodzie C użytkownik określa maksymalne tempo mowy spowolnionej, a chwilowe wartości współczynnika skali dobierane są w sposób automatyczny w zależności od wykrytego przez algorytm tempa mowy wejściowej. Należy tu wspomnieć, iż nietypowa numeracja metod rozpoczynająca się od litery B związana jest z tym, iż w dalszej części rozprawy porównano wpływ opracowanych metod na rozumienie mowy oraz metody równomiernego spowalniania mowy, którą oznaczono jako metoda A.

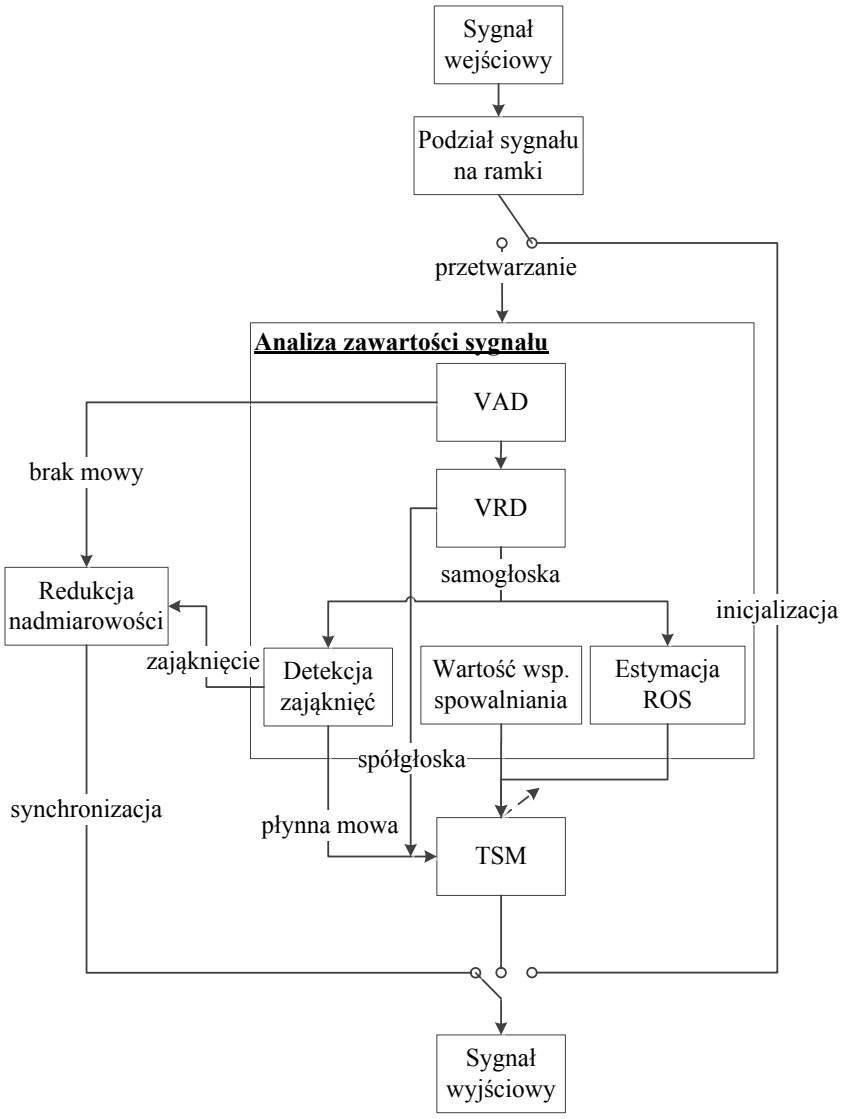
Opracowane metody nierównomiernej modyfikacji czasu trwania sygnału oparte zostały na założeniu, iż w sygnale na wejściu algorytmu znajdują się redundantne informacje tj. fragmenty ciszy (np. pauzy pomiędzy słowami, zdaniami lub wypowiedziami) czy wydłużone nienaturalnie samogłoski (nazywane na potrzeby tej rozprawy) zająknięciami. Nadmiarowe fragmenty sygnału są usuwane, jeżeli sygnał wyjściowy jest opóźniony w stosunku do sygnału wejściowego i niespowalniane, jeżeli opóźnienie nie występuje. Takie podejście pozwala „zaoszczędzić” dodatkowy czas, podczas którego odtwarzana jest spowolniona wypowiedź.

Dodatkowo, zgodnie z postulatami Coyle [22], Chu [16] i Demol [31], w celu uzyskania wysokiej naturalności mowy spowolnionej, struktura czasowa sygnału modyfikowana jest w sposób nierównomierny. Jest to wykonywane poprzez wykorzystanie różnych wartości współczynników skali dla samogłosek i spółgłosek. Wartości α dobierane są tak, by zachować prozodię mowy wejściowej, tj. samogłoski spowalniane są z wykorzystaniem większych wartości współczynnika skali niż spółgłoski. W zależności od tempa mowy wejściowej, wartości α zmieniane są tak, by zapewnić możliwe najmniejsze opóźnienie pomiędzy sygnałem wejściowy i wyjściowym. Należy tu zauważyć, iż dla

mowy wypowiadanej w wolnym tempie możliwe jest wykorzystanie mniejszych α niż dla mowy wypowiadanej w tempie szybkim. Sposób doboru chwilowych wartości współczynnika skali zależny jest od metody (B, C).

3.1 Opis metod

Na rys. 3.1 przedstawiono schemat blokowy ilustrujący sposób przetwarzania sygnału przez opracowane metody modyfikacji czasu trwania sygnału. Blok analizy zawartości sygnału składa się z następujących algorytmów: VAD, VRD, detektora zająknięć i estymatora ROS. Rdzeniem metody nierównomiernej TSM sygnału jest algorytm SOLA. Jak pokazują badania, zapewnia on wysoką jakość mowy spowolnionej, a równocześnie nie jest wymagający obliczeniowo [36]. Dodatkowo algorytm SOLA wykorzystuje ramki analizy o stałej długości przesuwane ze stałym krokiem (nie zależnym od zawartości sygnału wejściowego). To pozwala na bezpośrednią integrację algorytmów analizy zawartości sygnału z procedurą TSM, tj. każda ramka wejściowa sygnału analizowana jest w celu określenia jej zawartości. Następnie, na podstawie wyników analizy wykonywana jest modyfikacja struktury czasowej sygnału. Zastosowanie innego algorytmu TSM sygnału, niż SOLA (lub OLA), wiązałoby się z koniecznością stworzenia dwóch niezależnych torów przetwarzania, gdzie analiza sygnału wykonywana byłaby niezależnie od procesu TSM sygnału (podobnie jak zrobili to Demol *et al.* [31]). Innym możliwym rozwiązaniem byłoby zastosowanie algorytmów analizy zawartości sygnału, w których możliwe jest zastosowana ramek analizy o zmiennej długości (takie podejście zastosowali Nejime *et al.* [121]). Pierwsze rozwiązanie skutkuje zwiększeniem złożoności obliczeniowej algorytmu powodowanej przez konieczność podwójnego przetwarzania sygnału (analiza i TSM sygnału). Dodatkowo, ponieważ TSM sygnału wykonywana jest z użyciem innych ramek czasowych niż te wykorzystywane podczas analizy, dokładność oceny zawartości ramki aktualnie przetwarzanej przez algorytm jest niższa niż w jednoczesnej analizie i modyfikacji sygnału. Drugie rozwiązanie skutkuje wprowadzeniem ograniczeń w skuteczności segmentacji sygnału oraz uniemożliwia zastosowania większości metod analizy. W celu zbadania tego, jaki wpływ na jakość zmodyfikowanej mowy ma rodzaj zastosowane metody TSM sygnału, przeprowadzono serię testów subiektywnych, które wykazały, iż dla wartości α większych od 1,5 różnice w jakości mowy zmodyfikowanej za pomocą różnych metod TSM sygnału zaczynają być zauważalne [74]. W związku z tym celowe wydaje się zastosowanie algorytmu SOLA, podobnie jak zrobili to Covell *et al.* [21] (punkt 2.1.6).



Rys. 3.1 Schemat blokowy algorytm modyfikacji czasu trwania sygnału mowy.

Sygnał wejściowy przetwarzany jest w ramkach czasowych o trzech różnych długościach: 11,6 ms, 23,2 ms, 46,4 ms. Długości ramek są wielokrotnością najkrótszej z nich. Na wejście każdego z algorytmów podawana jest ramka sygnału o długości 11,6 ms. Każdy z nich buforuje odpowiednią liczbę ramek o podstawowej długości i tworzy z nich ramki długością odpowiadające długości ramki wykorzystywanej dalej przez ten algorytm.

Przetwarzanie sygnału wejściowego wykonywane jest zgodnie z poniższą procedurą:

1) Sygnał wejściowy zapisywany jest w ramce o długości 11,6 ms,
2) algorytm detekcji mowy analizuje zawartości otrzymanej ramki,
3) w ramkach niezawierających sygnału mowy przeprowadzana jest operacja redukcji nadmiarowości i ramka sygnału nie jest dalej analizowana; jeżeli sygnał wyjściowy nie jest zsynchronizowany z sygnałem wejściowym, wtedy ramka nie jest wysyłana

na wyjście algorytmu, w przeciwnym razie ramka jest wysyłana na wyjście algorytmu,

4) ramka oznaczona jako zawierająca mowę analizowana jest przez algorytm detekcji samogłosek,

5) informacja na temat pozycji samogłosek zapisywana jest w buforze algorytmu estymacji tempa wypowiedzi, oraz wykorzystywana jest przez algorytm detekcji zająknięć,

6) w kolejnym kroku sprawdzane jest wystąpienie zająknięcia. Po wykryciu zająknięcia przeprowadzana jest operacja redukcji nadmiarowości zgodnie z procedurą opisaną w punkcie 3,

7) ramki zawierające mowę bez zająknięć podawane są na wejście algorytmu TSM i przeprowadzana jest procedura modyfikacji czasu trwania sygnału z zastosowaniem różnych współczynników skali.

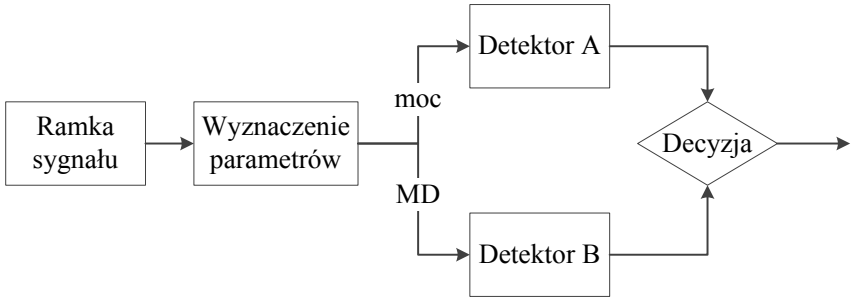
Wszystkie algorytmy będące częścią opracowanych metod nierównomiernej modyfikacji czasu trwania sygnału zostały zaprojektowane lub zaadaptowane tak, by umożliwiały przetwarzanie sygnału w czasie rzeczywistym.

3.2 Detekcja mowy

Niezawodna detekcja mowy jest kluczowym elementem opracowanej metody, ponieważ pozwala ona na uzyskanie wysokiego stopnia synchronizacji pomiędzy sygnałem wyjściowym a sygnałem spowolnionym. Zapewnia ona także odpowiednią jakość oraz zrozumiałość mowy wyjściowej. Niska skuteczności detekcji fragmentów sygnału niezawierających wypowiedzi (HR0) będzie powodować, iż fragmenty ciszy nie zostaną usunięte, a co więcej będą one spowalniane. W efekcie różnica w długości sygnału wejściowego i sygnału zmodyfikowanego będzie rosnąć. Natomiast błędy powodowane przez niską skuteczność detekcji mowy (HR1) spowodują, iż niektóre fragmenty mowy zostaną usunięte z sygnału. W konsekwencji obniżona zostanie zarówno jakość jaki i zrozumiałość przetworzonego sygnału fonicznego. Oczywiście drugi rodzaj błędów jest o wiele bardziej szkodliwy. Dlatego, podobnie, jak w znanych w literaturze algorytmów VAD, opracowywanych do celów kodowania sygnału, tak i tu, podczas opracowywania detektora główny nacisk został położony na niezawodną detekcję mowy (wysoką wartość HR1).

W oparciu o te założenia opracowano algorytm bazujący na analizie dwóch krótkookresowych parametrów sygnału: mocy widmowa amplitudowego oraz MD (ang. *Mean Delta*) (parametry te opisano w punkcie 2.2.1). Decyzja, dotycząca zawartości danej ramki sygnału, podejmowana jest z wykorzystaniem dwustopniowej adaptacji wartości progu. Nie wykorzystano tu skutecznych metod detekcji opartych np. na metodzie SVM (ang. *Support Vector Machine*), ponieważ ich użycie wiązałoby się z koniecznością przechowywania modeli klasyfikatorów. W implementacji na urządzeniu mobilnym, które posiada niewielką pamięć, powodowałoby to wprowadzenie ograniczeń co do wielkości modelu. Takie ograniczenie prowadziłoby do obniżenia skuteczność klasyfikacji lub uniemożliwiałoby całkowicie wytrenowanie klasyfikatora.

Na rys. 3.2 przedstawiono schemat blokowy opracowanego algorytmu detekcji mowy. Dokonuje on analizy sygnału w ramkach o długości 46,4 ms przesuwanych z krokiem 23,2 ms. Dla każdej ramki obliczona jest wartość mocy widma amplitudowego oraz MD. Dla obu parametrów tworzony jest osobny detektor progowy (detektor A i B). Końcowa decyzja, dotycząca przynależności analizowanej ramki do jednej z dwóch kategorii, podejmowana jest na podstawie wyniku dostarczonych przez dwa niezależne detektory. Jeżeli przynajmniej jeden z detektorów wykryje sygnał mowy, to jego decyzja jest ostateczną decyzją detektora. W przeciwnym razie uznaje się, że analizowana ramka sygnału nie zawiera mowy. Takie podejście pozwala na zwiększenie prawdopodobieństwa wykrycia sygnału mowy, ale jednocześnie skutkuje zwiększeniem liczby błędów pierwszego rodzaju.



Rys. 3.2 Schemat blokowy algorytmu detekcji mowy.

Każdy z detektorów mowy (A i B) dokonuje analizy parametrów w ten sam sposób. Zakłada się, iż na początku pracy algorytmu, przez pierwsze 1000 ms, w analizowanym sygnale nie występuje mowa. Założenie to pozwala wyznaczyć początkową wartość progu VAD_p jako średnią wartości parametrów wyznaczonych w pierwszej sekundzie nagrania. W celu uniknięcia zbytniego dopasowania wartości VAD_p, do tła akustycznego występującego w początkowej części sygnału, jego wartość jest wymnażana przez

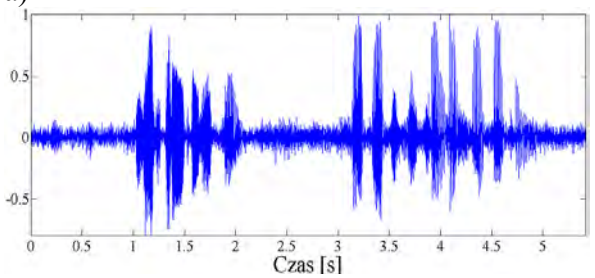
współczynnik $C = 1{,}1$. Wartość C została dobrana eksperymentalnie w taki sposób, by zapewniała możliwie najwyższą skuteczność detekcji sygnału mowy na początku nagrania [75]. Podczas pracy algorytmu, gdy żaden z detektorów nie wykryje sygnału mowy, wartość progu jest poddawana adaptacji z użyciem filtru dolnoprzepustowego opisanego wzorem (2.31). Wartość współczynnika a dobierana jest zgodnie z procedurą opisana przez Sangwana *et al.* [159]. Tak więc na początku a przyjmuje wartość 0,2. Za każdym razem, gdy algorytm VAD nie wykryje sygnału mowy, wyznaczana jest wariancja wartości analizowanego parametru w ostatnich 10 szumowych ramkach sygnału licząc od ramki $m_{\text{szum}-1}$ (gdzie m_{szum} oznacza numer aktualnej ramki szumowej, a $m_{\text{szum}-1}$ to poprzednia ramka zawierająca szum) oraz wariancja wyznaczona dla ostatnich 10 ramek szumowych licząc od ramki m_{szum}-*tej*. Pierwsza z wartości nazywana jest „starą" wariancją szumu i oznaczana jest jako σ^2_{stara}, a druga wartość nazywana jest „nową" wariancją szumu (σ^2_{nowa}). Jeżeli wariancja zaczyna wzrastać, to a powinno przyjmować większe wartości, tak by aktualna wartość parametru miała większy wpływ na wartość progu. Aktualna wartość współczynnika a dobierana jest zgodnie z zależnościami przedstawionymi w tab. 3.1.

Tab.3.1 Relacja pomiędzy wariancją parametru a wartością współczynnika a [159].

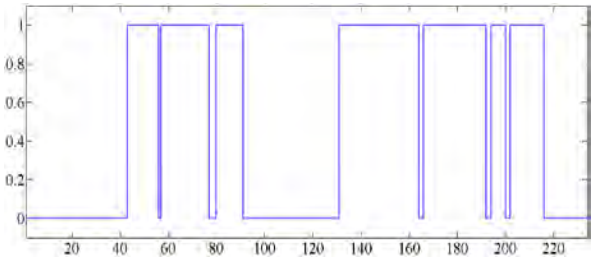
$\sigma^2_{\text{nowa}}/\sigma^2_{\text{stara}}$	a
$\sigma^2_{\text{nowa}}/\sigma^2_{\text{stara}} > 1{,}25$	0,25
$1{,}25 \geq \sigma^2_{\text{nowa}}/\sigma^2_{\text{stara}} > 1{,}1$	0,2
$1{,}1 \geq \sigma^2_{\text{nowa}}/\sigma^2_{\text{stara}} > 1{,}0$	0,15
$\sigma^2_{\text{nowa}}/\sigma^2_{\text{stara}} \leq 1{,}0$	0,1

W celu wygładzenia decyzji i zmniejszenie prawdopodobieństwa błędnej klasyfikacji ramek zawierających mowę jako szum, ostateczna decyzja jest wygładzana poprzez oznaczanie wszystkich fragmentów ciszy krótszych niż 200 ms jako mowa. Na rys. 3.3 przedstawiono przykład działania opracowanego algorytmu VAD podczas detekcji mowy w warunkach zakłóceń szumem typu *babble noise* przy SNR = 10 dB. Poprawna detekcja mowy w warunkach zaszumienia tego typem szumu przy niskiej wartości SNR jest bardzo trudna, ponieważ charakterystyka widmowa szumu jest mocno zbliżona do charakterystyki widmowa sygnału, który ma zostać wykryty (mowa).

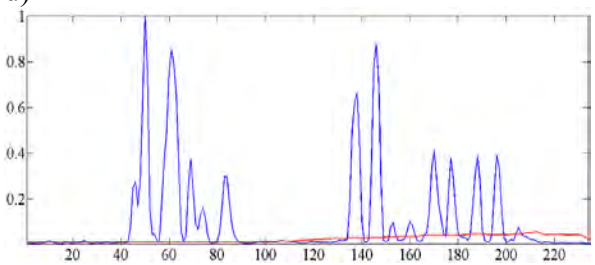
a)



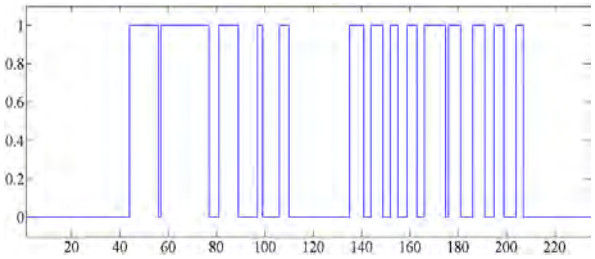
b)



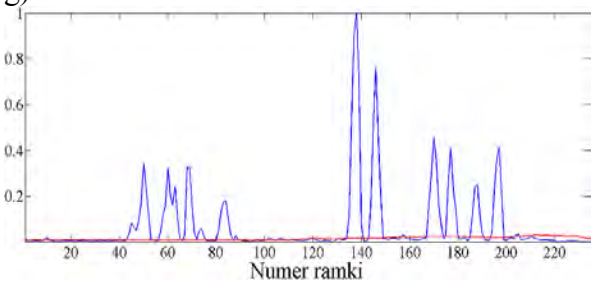
d)



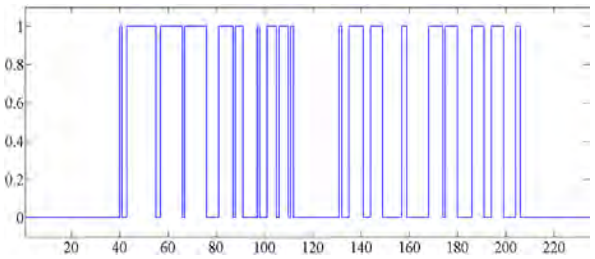
e)



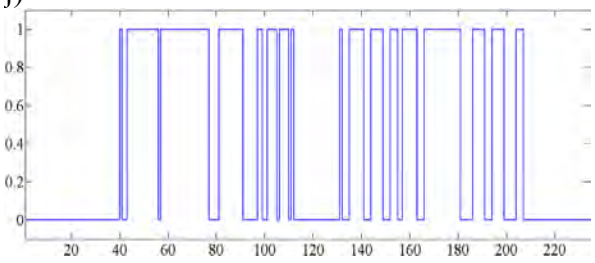
g)



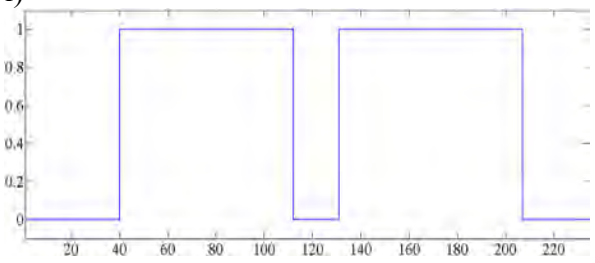
h)

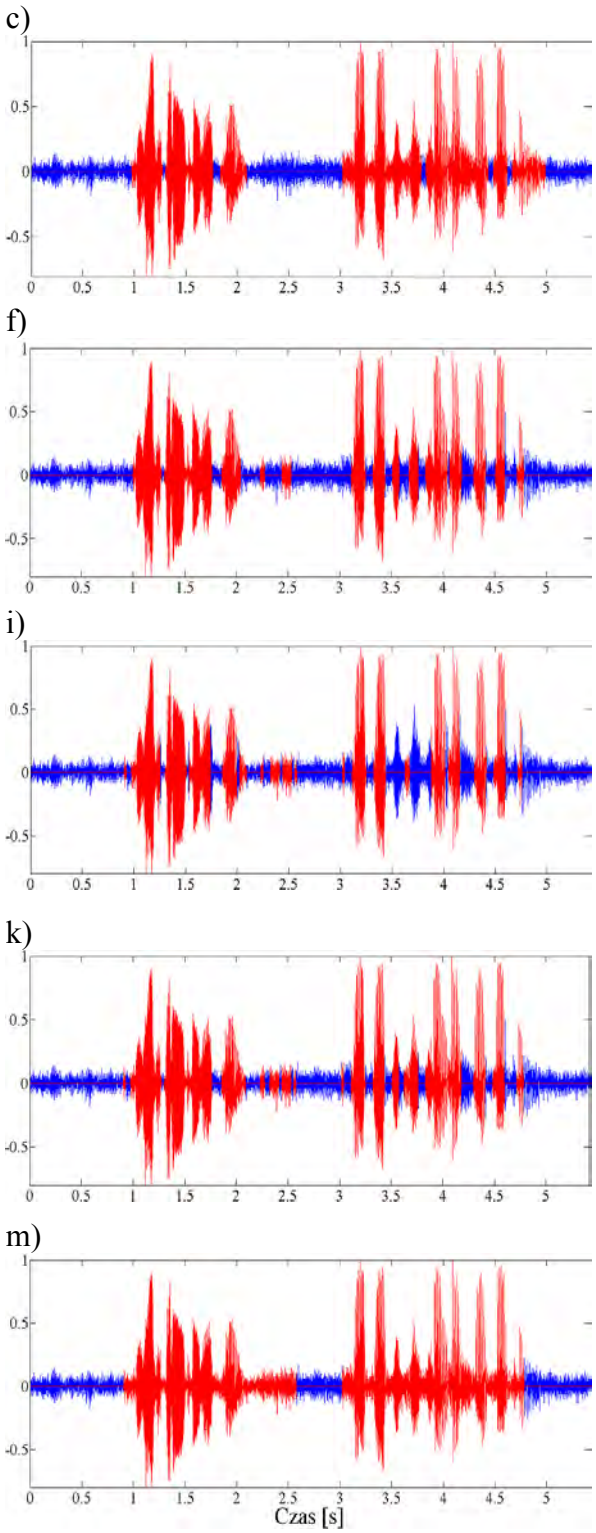


j)



l)





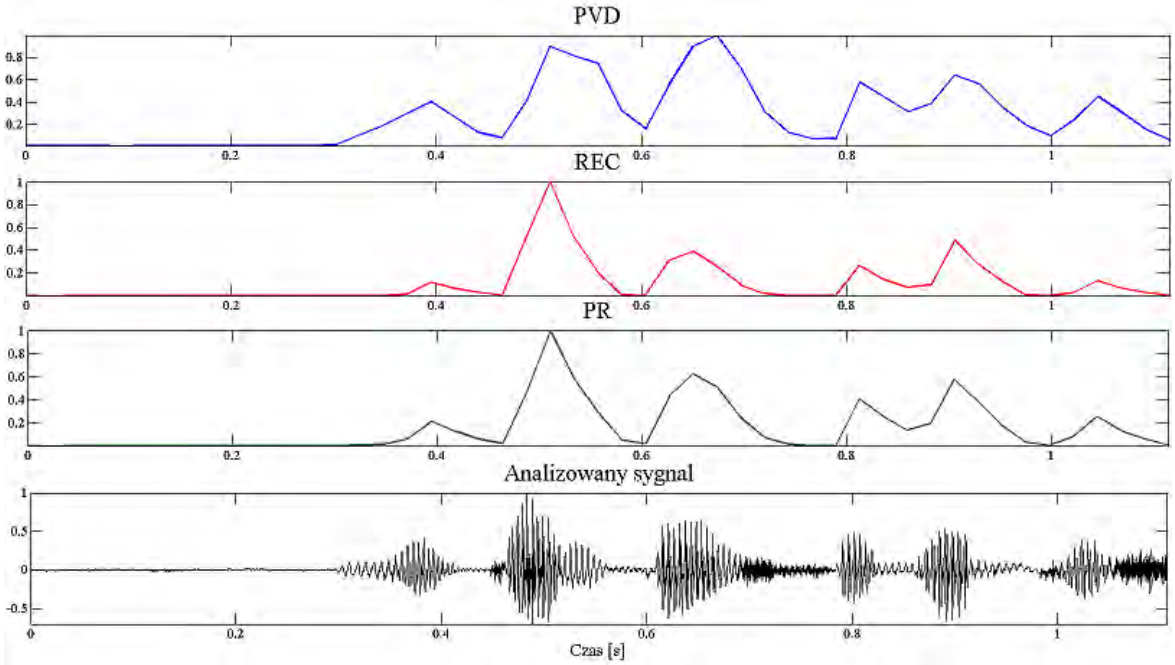
Rys. 3.3 Przykład detekcji mowy zakłóconej szumem typu *babble noise* (SNR = 10dB). Wypowiedź składa się z dwóch zdań: „*Dziewczynka śpi.*” i „*Chłopiec sprząta swój pokój.*”. Pierwsze zdanie wypowiadane jest przez kobietę a drugie przez mężczyznę.

Na wykresie a) umieszczono oscylogram analizowanego sygnału. Jest to pięciosekundowy fragment rozmowy kobiety z mężczyzną. Poniżej zamieszczono wykresy ilustrujące zmiany mocy widmowa amplitudowego (rys. d)) oraz parametru MD (rys. g)) w kolejnych ramkach sygnału. Czerwonym kolorem zaznaczono wartości progu wykorzystywanego podczas detekcji. „Schodowy" wykres b) pokazuje wynik ręcznego oznaczenia nagrania, a na wykresach e), h), j) oraz l) wykreślono wyniki detekcji mowy uzyskane odpowiednio dla algorytmu VAD korzystającego z analizy: mocy widmowa amplitudowego, parametru MD, obu parametrów, a także wygładzoną decyzję uzyskaną dla opracowanego algorytmu VAD. Na rys. c), f), i), k), m) czerwonym kolorem zaznaczono fragmenty sygnału odpowiadające mowie wykrytej przez kolejne algorytmy detekcji a kolorem niebieskim fragmenty zaklasyfikowane jako szum. Na podstawie ręcznego oznaczenia nagrania można stwierdzić, iż w nagraniu znajdują się dwie dłuższe wypowiedzi. W każdej wypowiedzi można zaobserwować krótkie przerwy pomiędzy słowami trwające nie dłużej niż dwie kolejne ramki analizy (69,6 ms). Wynik detekcji oparty jedynie na analizie jednego parametru prowadzi do nie wykrycia fragmentów o niskiej mocy występujących w pierwszej części wypowiedzi. W konsekwencji prowadzi to do podniesienia wartości progu detekcji (rys. d) i g)), przez co w drugiej części wypowiedzi liczba niewykrytych ramek zawierających mowę jest znacznie większa. Wynik detekcji opracowanego algorytmu jest sumą logiczną wyników dwóch osobnych detektorów, przez co sumaryczna liczba niewykrytych ramek zawierających mowę jest niższa. Jednocześnie algorytm powoduje powstanie wielu błędów pierwszego rodzaju. Dzięki zastosowaniu wygładzania liczba niewykrytych ramek zawierających mowę została prawie całkowicie wyeliminowana kosztem klasyfikacji niektórych ramek szumowych jako mowa. Na rys. m) widać, iż wygładzanie powoduje podtrzymanie decyzji o wykryciu mowy przez kolejne 500 ms po zakończeniu pierwszej części wypowiedzi. Natomiast początek mowy wykrywany jest poprawnie. Za to drugiej części wypowiedzi końcowa część zdania nie została wykryta poprawnie.

3.3 Detekcja samogłosek i zająknięć

Opracowany algorytm VRD (ang. *Vowel Region Detector*) został oparty na analizie dwóch parametrów: REC (ang. *Reduced Energy Cumulating*) oraz PVD (ang. *Peak Valley Difference*). Jego zadaniem jest detekcja wystąpienia oraz wykrycie obszaru samogłoski. Algorytm analizuje tylko te ramki sygnału, które zostały oznaczone przez VAD jako mowa. Wykorzystanie parametru REC wydaje się tu naturalnym rozwiązaniem, ponieważ

parametr ten został wynaleziony przez Pellegrino i Andre-Obrecht [137] właśnie do celów detekcji samogłosek. Nie jest jednak możliwe bezpośrednie wykorzystanie algorytmu proponowanego przez autorów parametru do detekcji obszaru samogłoski, ponieważ algorytm segmentacji FBD (ang. *Forward-Backward Divergence*), będący nieodłączną częścią tej metody, nie działa w czasie rzeczywistym (p. 2.4.1). W literaturze znany jest także inny parametr nazwany PVD. Został oparty na założeniu, iż widma amplitudowe wszystkich samogłosek są spójne, przez co możliwe jest znalezienia podobieństwa aktualnie analizowanej ramki sygnału do średniego widma amplitudowego sygnału samogłoski. Jak wspominano w punkcie 2.2.1 parametr ten został opracowany w celu detekcji mowy. Możliwe jest jednak jego wykorzystanie w algorytmie detekcji samogłosek. Autor rozprawy wykorzystywał ten parametr do tego celu w swoich wcześniejszych pracach [73] [72] [76]. Jako, iż oba parametry opisują podobieństwo aktualnie analizowanej ramki sygnału do charakterystyki widmowej typowej samogłoski, możliwe jest ich równoczesne wykorzystanie. W pracy użyto informacje niesione przez obie liczby poprzez utworzenie nowego parametru nazwanego PR (od pierwszych liter skrótów PVD i REC). PR wyznaczane jest jako suma wartości PVD i REC w danej ramce analizy. Na rys. 3.4 przedstawiono parametry PVD, REC i PR obliczone dla fragmentu sygnału zawierającego wypowiedź mężczyzny.

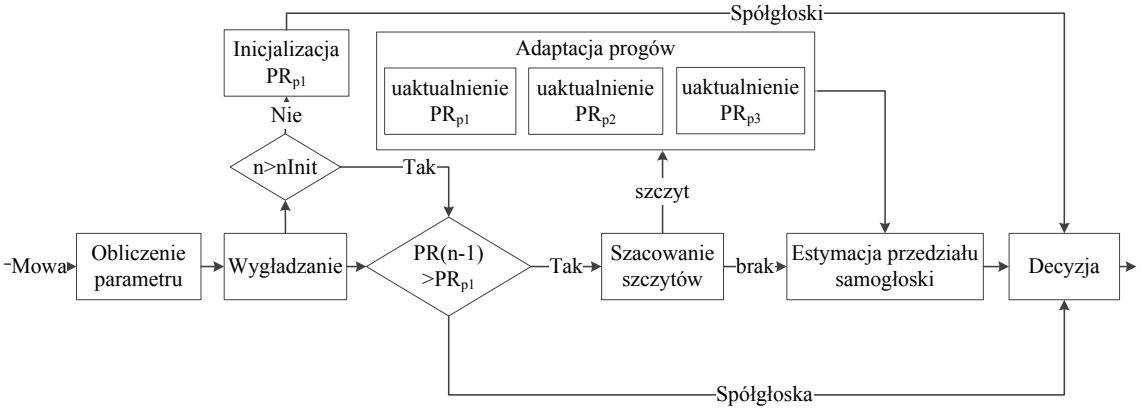


Rys. 3.4 Porównanie parametrów PVD, REC i PR wyznaczonych dla wypowiedzi o treści: „*Witam państwa bardzo s…*” (głos męski).

Detekcja samogłosek wykonywana jest na podstawie analizy zmienności parametru PR. Na rys. 3.5 przedstawiono schemat blokowy opracowanego algorytmu. Jest on oparty

na analizie szczytów w przebiegu parametru PR, których wartość przekracza próg PR_{p1}. Wartość PR_{p1} jest adaptacyjnie zmieniana podczas procesu detekcji. Przed rozpoczęciem pracy wymagany jest krok inicjalizacji, pozwalający na wyznaczenie początkowej wartości PR_{p1}. Jest ona obliczana jako suma wartości średniej i odchylenia standardowego wyznaczonych podczas pierwszej sekundy pracy algorytmu (inicjalizacja progu odbywa się równolegle z inicjalizacją wartości progowych detektora mowy). Podczas analizy, wartość progu PR_{p1} dostosowywana jest do amplitudy aktualnie wykrytego szczytu w przebiegu zmienności parametru PR. Próg ustawiany jest jako 5% amplitudy ostatnio wykrytego szczytu. Wartość ta została dobrana eksperymentalnie w taki sposób, by zapewnić jak najmniejszą liczbę błędów detekcji.

W celu oszacowanie obszaru samogłoski, wykonywana jest analiza wartości PR w sąsiedztwie wykrytego szczytu. Wszystkie wartości parametru znajdujące się przed szczytem, których amplituda jest większa niż 25% aktualnego szczytu (PR_{p2}) oraz wszystkie wartość występujące po szczycie, których amplituda jest większą niż 50% szczytu (PR_{p3}) oznaczane są jako obszar samogłoski. Wartości progów PR_{p2} i PR_{p3} zostały dobrane w sposób eksperymentalny i odzwierciedlają powtarzającą się prawidłowość: wartości parametru PR odpowiadające obszarowi samogłoski występującemu przed szczytem są niższe, niż wartości parametru PR odpowiadające obszarowi samogłoski występującej po szczycie.



Rys. 3.5 Schemat blokowy algorytm detekcji samogłosek.

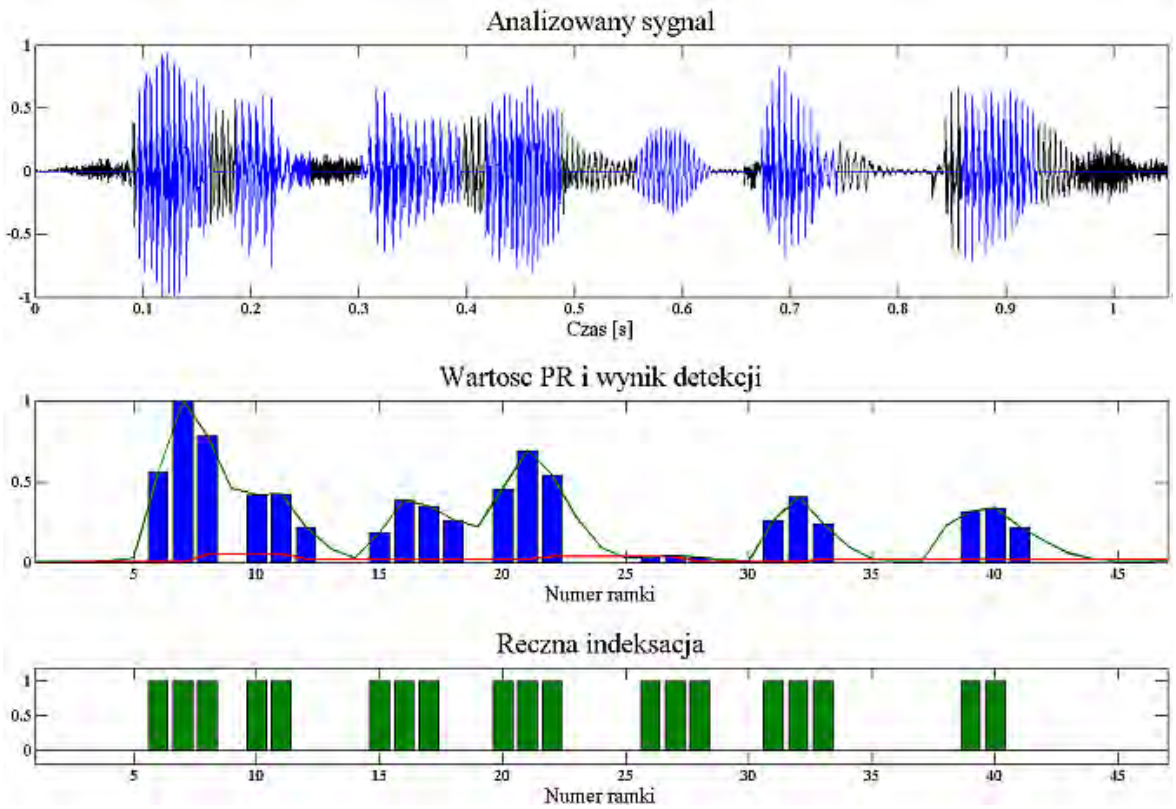
W implementacji działającej w czasie rzeczywistym wykorzystano bufor zawierający trzy sąsiednie wartości parametru PR. Dodatkowo zastosowano wygładzanie polegające na wyznaczeniu średniej wartości ważonej parametrów znajdujących się w buforze. Taki sposób analizy powoduje opóźnienie. Jego wielkość równa jest czasowi 2 ramek analizy. W celu zminimalizowania wielkości opóźnienia wartości parametru wyznaczane są z

użyciem krótkich ramek. Długość ramki i krok analizy wykorzystywane przez detektor wynoszą 23,2 ms. Wprowadza to opóźnienie decyzji równe 46,4 ms.

Reguł decyzyjne opisane powyżej zostały zrealizowane w czasie rzeczywistym poprzez zastosowanie następującego algorytmu przetwarzania:

1) jeżeli wartość PR dla ramki *m–1* jest większa niż dla ramek *m* i *m–2*, gdzie *m* jest numerem aktualnie analizowanej ramki, i większa od progów PR_{p1} i PR_{p3}, wtedy ramka *m–2* uznawana jest za ramkę zawierającą samogłoskę i ustawiana jest informacja na temat wystąpienia szczytu w ramce *m–1*,
2) jeżeli warunek 1 nie zostanie spełniony, sprawdzany jest warunek 2, tj.: jeżeli informacja o wystąpieniu szczytu jest aktualna i wartość PR dla ramki *m–2* jest większa od progu PR_{p2}, wtedy ramka *m–2* oznaczana jest jako zawierająca samogłoskę,
3) we wszystkich innych przypadkach ramka *m–2* oznaczana jest jako ramka zawierająca spółgłoskę i informacja na temat wystąpienia szczytu jest usuwana.

Na rys. 3.6 przedstawiono przykładowy wyniki detekcji. Na górnym wykresie umieszczono przebieg czasowy analizowanego sygnału. Kolorem niebieski zaznaczono wykryty obszar samogłosek. Środkowy wykres przedstawia przebieg wartości PR wraz z zaznaczonym wynikiem detekcji. Czerwona linia ilustruje wartość progu PR_{p1}. Natomiast na najniższym wykresie pokazano wynik ręcznego oznaczenia nagrania. Jak można zauważyć wszystkie samogłoski zostały wykryte poprawnie, jednak nie wszędzie ich przedziały został znaleziony prawidłowo (np. w 12 ramce analizy).



Rys. 3.6 Przykład detekcji samogłosek w wypowiedzi o treści: „*Halo, halo witam państwa ...*” (głos męski).

Omówiona powyżej implementacja ma jedno główne ograniczenie: tylko jedna ramka przed szczytem, występującym w przebiegu wartości PR, może zostać oznaczona jako samogłoska. Możliwa jest minimalizacja tego obostrzenia polegająca na wydłużeniu bufora analizy, jednak spowoduje to zwiększenie opóźnienia w podjęciu decyzji przez algorytm. Jako, że opóźnienie opracowywanego algorytmu modyfikacji czasu trwania sygnału musi być możliwie najmniejsze nie ma możliwości eliminacji wspominanego ograniczenia.

Na podstawie analizy informacji dostarczanych przez algorytm VRD wykonywana jest detekcja zająknięć. Zająknięcie zdefiniowano jako samogłoskę trwająca dłużej niż 371 ms (16 ramek o długości 23,2 ms). Ten czas dobrano na podstawie analizy długości samogłosek występujących w 30 oznaczonych ręcznie wypowiedziach (mówca męski i żeński). Ramka sygnału oznaczana jest jako zająknięcie, jeżeli zawiera samogłoskę i przed tą ramką przynajmniej 16 ramek sygnału zawierało samogłoski. Informacja na temat wystąpienia zająknięć pozwala na uzyskanie wyższej naturalności sygnału spowolnionego poprzez wykluczenie ramek zawierających zająknięcie z procesu spowalniania oraz na uzyskanie lepszej synchronizacji sygnału wejściowego z sygnałem spowolnionym przez usuwanie fragmentów samogłosek wypowiadanych dłużej niż 371 ms.

3.4 Estymacja tempa wypowiedzi

Opracowany algorytm estymacji ROS, podobnie jak metoda zaproponowana przez Pfau i Ruske [141], został oparty na analizie liczby wystąpień samogłosek. Nie analizuje on bezpośrednio sygnału fonicznego, tylko wyjście z algorytmu VRD. Algorytm VRD na wejście estymatora ROS dostarcza informację dotyczące wystąpienia szczytów w sygnale zmienności parametru PR. Miejsca te zostały nazwane VR (ang. *Vowel Region*), i reprezentują one przedział samogłoski. Zastosowanie takiego rozwiązania pozwala na eliminację konieczności wykonywania dodatkowych obliczeń polegających na wyznaczaniu różnego rodzaju parametrów oraz ich analizy, dzięki czemu estymacja ROS staje się bezkosztowa obliczeniowo. Algorytm estymuje lokalną wartość *netto* tempa mowy dla każdej 23,2 ms ramki sygnału (czas trwania ramki jest taki jak w algorytmie VRD). Chwilowa wartość ROS wykorzystywana jest w metodzie C w celu wyznaczenie wartości współczynnika skali i jest ona obliczana zgodnie ze wzorem:

$$\mathrm{ROS}[m] = \frac{n_{vr}}{\Delta t} \tag{3.1}$$

gdzie ROS[m] oznacza wartość estymowanego tempa mowy w m-tej ramce, n_{vr} liczbę wystąpień obszarów VR w przedziale czasu Δt, a Δt reprezentuje przedział czasu, dla którego zliczona została liczba wystąpień VR. Chwilowa wartość ROS[m] obliczana jest jako średnia liczba samogłosek, które wystąpiły w przedziale wynoszącym 1,5 sekundy (w przedziale wystąpienia mowy). Czas uśredniania został dobrany eksperymentalnie, tak by możliwe było uchwycenie lokalnych zmian tempa wypowiedzi. Należy zaznaczyć, iż wartość ROS jest obliczana tylko wtedy, gdy obecnie analizowana ramka sygnału zawiera mowę i nie zawiera zająknięcia. W celu eliminacji sytuacji, w której wartość tempa będzie wzrastać od 0; na początku przetwarzania algorytm domyślnie przyjmuje estymowaną wartość tempa równą 4 samogłoski/s.

Dodatkowo na podstawie analizy chwilowej wartość tempa mowy, każda z ramek przypisywana jest do jednej z dwóch klas tempa: szybkie albo wolne. Informacja dotyczą klasy tempa wykorzystywana jest w metodzie B w celu doboru odpowiedniej wartości współczynnika skali. Podział na dwie klasy wykonywany jest na podstawie progu $\mathrm{ROS}_{\mathrm{th}}$. Jego wartość wyznaczono poprzez analizę średnich wartości ROS obliczonych dla nagrań zawierających trzy klasy tempa: wolne, normalne i szybkie. Nagrania zawierały zdania wypowiadane przez dwóch mówców (kobietę i mężczyznę). Każda osoba czytała pięć różnych fraz (ich treść zamieszczono w załączniku nr 1). W nagraniach oznaczono miejsca

wystąpień samogłosek i na ich podstawie dla każdej wypowiedzi obliczono średnią wartość wartości *netto* ROS. Uzyskane wyniki zamieszczono w tab. 3.2.

Tab. 3.2 Wartość średnia oraz odchylenie standardowe wartości *netto* ROS wyznaczone dla mowy wypowiadanej w trzech różnych tempach.

tempo mowy	**wolne**	**normalne**	**szybkie**
μ(ROS)[samogłosek/s]	3,60	4,41	5,03
σ(ROS)[samogłosek/s]	0,20	0,28	0,40

Na podstawie uzyskanych statystyk, wyznaczono wartość ROS_{th} poprzez obliczenie wartość średnią dwóch skrajnych klas:

$$ROS_{th} = \frac{\mu(ROS_{slow}) + \mu(ROS_{fast})}{2} \tag{3.2}$$

gdzie $\mu(ROS_{slow})$ oznacza wartość średnią tempa mowy wolnej, a $\mu(ROS_{fast})$ wartość średnią tempa mowy szybkiej.

3.5 Modyfikacja czasu trwania sygnału mowy

Proces modyfikacji czasu trwania sygnału mowy został oparty na algorytmie SOLA, ponieważ (jak wspomniano na początku tego rozdziału), budowa tego algorytmu pozwala na bezpośrednie wykorzystanie go w realizacji działającej w czasie rzeczywistym. Stała długość ramek analizy oraz kroku analizy pozwala na prostą współpracę algorytmu TSM z opracowanymi algorytmami detekcji.

W celu modyfikacji czasu trwania sygnału zaproponowano dwa różne rozwiązania pozwalające na ustalenie tempa mowy spowolnionej. W pierwszym algorytmie tempo sterowane jest, tak jak w tradycyjnych algorytmach, przy użyciu wartości współczynnika skali. Drugie podejście zakłada, iż osoba korzystająca z algorytmu chciałaby słyszeć mowę w tempie określanym w liczbie samogłosek na sekundę.

W celu uzyskanie wysokiej jakości mowy spowolnionej, długości ramek analizy/syntezy oraz krok przesunięcie analizy, zostały dobrane w taki sposób by długość ramki L zawierała przynajmniej dwa okresy podstawowe mowy (około 40 ms), a w kroku syntezy dla wszystkich wykorzystywanych wartości α, wartość zakładki pomiędzy kolejnymi ramkami wynosiła $L/3$ [36]. Dlatego algorytm operuje na ramkach o długości 46,4 ms a przesunięcie Sa wynosi 11,6 ms.

3.5.1. Tempo sterowane współczynnikiem spowalniania (metoda B)

W metodzie tej chwilowa wartość współczynnika skali α zależna jest od wartości α_o (określanej przez użytkownika) oraz od zawartości i tempa mowy wejściowej. Aktualna wartość α obliczana jest dla każdej ramki zgodnie zasadami przedstawionymi w tab. 3.3. Informacje dotyczące tempa mowy oraz zawartości sygnału dostarczane są przez opisane powyżej algorytmy: estymacji ROS, detekcji mowy i detekcji samogłosek. Zastosowanie takiego sposobu doboru chwilowej wartości współczynnika skali skutkuje tym, iż mowa szybka spowalniana jest wyższym stopniu niż mowa wolna. Dzięki tej strategii doboru chwilowej wartość współczynnika skali zarówno mowa wypowiadana w tempie szybkim jaki i wolnym, na wyjściu algorytmu mają zbliżone tempo. Wprowadzono także dodatkowe obostrzenie związane z wartościami, jakie może przyjmować α: mianowicie wartość ta nie może być mniejsza niż 1,0. W sytuacji, gdy obliczona chwilowa wartość współczynnika skali jest mniejsza niż 1,0, jest ona zmieniana na 1,0. Dodatkowo należy zaznaczyć, iż w przypadku fragmentów ciszy α przyjmuje wartość równą 1,0 jedynie wtedy, gdy sygnał wyjściowy jest zsynchronizowany z sygnałem wejściowym, w przeciwnym razie sygnał ten nie jest przekazywany na wyjście algorytmu.

Tab. 3.3 Zależność chwilowej wartości współczynnika skali od wykrytej zawartości sygnału oraz od estymowanego tempa wypowiedzi.

Klasa ROS	***α***		
	Samogłoski	**Spółgłoski**	**Cisza**
Szybka	α_o	0,8 α_o	1
Wolna	0,95 α_o	0,75 α_o	1

3.5.2. Tempo sterowane wartością (metoda C)

W tej metodzie, w celu uzyskania komfortowego tempa mowy, chwilowa wartość współczynnika skali ustalana jest na podstawie wartości ROS_o (określonej przez użytkownika). ROS_o określa oczekiwaną wartością tempa mowy wyjściowej. Wtedy chwilową wartość α należy zdefiniować w następujący sposób (3.3):

$$\alpha = \frac{\mathrm{ROS}[m]}{\mathrm{ROS_O}} \tag{3.3}$$

gdzie ROS[m] oznacza wartość tempa wypowiedzi wyznaczoną przez algorytm estymacji ROS dla m-tej ramki sygnału.

Współczynniki skali dla samogłosek i spółgłosek wyznaczane są na podstawie tych samych założeń co w metodzie B mówiącej, iż samogłoski w mowie spowolnionej powinny być wydłużone bardziej niż spółgłoski. Chwilową wartość współczynnika skali dla ramek zawierających spółgłoski wyznacza się zgodnie ze wzorami:

$$\alpha_{\text{cons}} = \frac{\alpha \cdot \Delta t}{(\eta - 1) \cdot \Delta t_{\text{vowel}} + \Delta t} \tag{3.4}$$

$$\alpha_{\text{vowel}} = \eta \cdot \alpha_{\text{cons}} \tag{3.5}$$

gdzie α_{cons} oznacza współczynnik skali stosowany dla m-tej ramki przy wykryciu spółgłoski, α_{vowel} współczynnik skali stosowany dla m-tej ramki przy wykryciu samogłoski, Δt przedziału czasu wykorzystywany przez algorytm estymacji ROS (1,5 s), Δt_{vowel} to czas trwania samogłosek w przedziale czasu Δt, a η jest współczynnikiem określającym relację pomiędzy wartościami współczynników skali wykorzystywanych do spowalniania samogłosek i spółgłosek (w tej implementacji wartość ta została ustalona na 1,7).

4 Badanie wpływu opracowanych metod na proces rozumienia mowy

W ramach tego rozdziału przedstawiono wyniki badań poświęconych ocenie wpływu opracowanych metod modyfikacji sygnału mowy na proces jej rozumienia przez osoby z pogorszoną rozdzielczością czasową słuchu. Prace te zostały także opublikowane w artykule [77]. Jako grupy docelowe w badaniu uczestniczyły dzieci ze zdiagnozowanymi obwodowymi zaburzeniami słuchu (4–14 lat) oraz osoby starsze (64–91 lat). Taki dobór grup słuchaczy podyktowany był brakiem dostępu osób mających zaburzenia słuchu związane jedynie z (C)APD. W związku z tym badania przeprowadzono z udziałem grup słuchaczy, dla których problem z rozdzielczością czasową słuchu jest wysoce prawdopodobny. W badaniach wzięło udział 34 ochotników (siedemnaścioro dzieci głuchych i 17 osób starszych).

4.1 Metodologia badań

Badania rozumienia mowy przeprowadzono z wykorzystaniem dwóch testów: testu mowy przyspieszonej oraz testu mowy spowolnionej. Pierwszy z nich pozwolił ocenić rozdzielczość czasową słuchaczy. Wyniki drugiego testu pokazały czy i kiedy opracowana metoda modyfikacji mowy jest użyteczna.

4.1.1. Materiał zdaniowy

W obydwu badaniach wykorzystano ten sam materiał zdaniowy. Były to opracowany przez Ozimka *et al.* test macierzowe dla osób dorosłych (PTM – Polski Test Macierzowy) [130] oraz test dla dzieci (PPTM – Polski Pediatryczny Test Macierzowy) [131]. Użyteczność tego materiału zdaniowego w badaniach rozumienia mowy została wykazana przez jego autorów [129]. Struktura syntaktyczna zdań, każdego z testów, jest stała, natomiast zdania poprawne gramatycznie lecz o małej zawartości kontekstowej. Składają się one z 5 (test PTM) i 3 słów (test PPTM). Wypowiedzi tworzone są poprzez syntezę pojedynczych słów zgodnie z procedurą typowego testu macierzowego opracowanego przez Hagermana [50], tj.: zbiór słów jest ograniczony (do 50 w teście PTM i 48 w teście PPTM), zdania tworzone są poprzez losowy wybór słów zgodnie z określoną strukturą gramatyczną. Pozwala to na syntezę 100000 unikatowych zdań w teście PTM i 256 w teście PPTM. Różnica w liczbie kombinacji związana jest z faktem, iż w teście PPTM możliwe jest łączenie ze sobą wyłącznie tych słów, które należą do tego samego

podzbioru, a w teście PTM możliwe są dowolne kombinacje słów w ramach określonej struktury syntaktycznej. W tab. 4.1 i 4.2 zamieszczono materiał słowny wykorzystywany w tych testach.

Tab. 4.1 Zbiór słów wykorzystywany w teście PTM [130]

Rzeczownik	**Czasownik**	**Liczebnik**	**Przymiotnik**	**Rzeczownik**
Tomasz	nosi	pięć	dobrych	piłek
Paweł	woli	sześć	tanich	gazet
Adam	widzi	siedem	drogich	soków
Maciej	bierze	osiem	pięknych	dzwonów
Michał	daje	dziewięć	nowych	opon
Anna	ma	dużo	starych	stołów
Ewa	robi	sto	białych	klocków
Maria	kupi	tysiąc	żółtych	toreb
Zofia	wygra	wiele	czarnych	okien
Julia	sprzeda	kilka	dziwnych	koszy

Tab. 4.2 Zbiór słów wykorzystywany w teście PPTM [131]

Numer zbioru	**Rzeczownik**	**Czasownik**	**Rzeczownik**
1	Babcia	maluje	dom
	Chłopiec	ogląda	pałac
	Dziecko	otwiera	samochód
	Strażak	rysuje	szafę
2	Kot	je	marchewki
	Miś	niesie	ogórki
	Zając	podlewa	truskawki
	Żaba	zrywa	wisienki
3	Dziadek	nosi	koszulę
	Dziewczyna	pierze	płaszcz
	Mama	prasuje	spodnie
	Tata	wiesza	sweter
4	Król	ma	herbatę
	Królewna	miesza	kawę
	Pani	nalewa	mleko
	Żołnierz	pije	wodę

Przykładowym zdaniem wygenerowanym dla testu PTM jest:

„Maria nosi dużo drogich toreb.”

Zdania tworzone w ramach tego testu nie niosą istotnej informacji, co uniemożliwia słuchaczowi zgadnięcia niezrozumianych słów poprzez analizę kontekstu.

W teście PPTM zdania muszą być tworzone ze słów należących do tego samego zbioru. Ograniczenie to jest wymagane, ponieważ w teście PPTM zakłada się, iż zdania nie mogą być abstrakcyjne, a połączenie słów z różnych zbiorów mogłoby prowadzić do powstania

tego rodzaju zdań (np. *Kot nalewa pałac*). Poniżej przedstawiono dwa poprawnie utworzone zdania w ramach testu PPTM:

„Babcia ogląda szafę."

„Zając je ogórki."

Słowa wykorzystane do tworzenia materiału zdaniowego zostały zarejesrtowane w studiu KSM. Jest to pomieszczenie przeznaczonym do nagrań lektorskich. Zarejestrowano mówcę męskiego. Miał on za zadanie odczytać słowa w trzech rożnych tempach: normalnym, średnim i szybkim. Słowa nagrano w formie dziesięciu (test PTM) i szesnastu zdań (testu PPTM). Zawierały one wszystkie wymagane słowa i uwzględniały strukturę syntaktyczną testów. Taki sposób rejestracji materiału słownego pozwolił na zachowanie naturalnej prozodii wypowiadanych zdań. W procesie edycji wycięto każde ze słów i dołączono na jego początku i końcu 200 ms naturalnego szumu, będącego szumem pomieszczenia. Miało to na celu umożliwić naturalne łączenie ze sobą kolejnych słów w zdania. Synteza wypowiedzi polegała na połączeniu odpowiednich słów poprzez zsumowanie obszarów ciszy z zastosowaniem liniowego *cross-fade'u*. Aplikację syntezującą zdania zaimplementowano w środowisku MATLAB. W załączniku nr 2 umieszczono przykładowe zestawy wypowiedzi wygenerowane dla obu testów, a na płycie dołączonej do rozprawy – odpowiadające im pliki dźwiękowy.

Korzystając z opracowanej aplikacji wyznaczono średnie wartości ROS dla różnych grup tempa wypowiedzi lektora. Wartości ROS obliczono dla wszystkich możliwych kombinacji zdań w testach PTM i PPTM. Uzyskane wyniki zamieszczono w tab. 4.3. Wykorzystano je w późniejszej analizie wyników badań. Pozwoliły one również na wybór odpowiedniego tempa mowy do poszczególnych testów.

Tab. 4.3 Wartość średnia i odchylenie standardowe ROS wyznaczone dla wszystkich kombinacji zdań zarejestrowanych w ramach testów PTM i PPTM

Rodzaj testu	**Statystyka**	**ROS^{P}_{wolne}**	**$ROS^{P}_{średnie}$**	**$ROS^{P}_{szybkie}$**
PTM	μ(ROS) [samogłosek/s]	2,72	4,88	6,48
	σ(ROS)[samogłosek/s]	0,23	0,42	0,44
		ROS^{PP}_{wolne}	**$ROS^{PP}_{średnie}$**	**$ROS^{PP}_{szybkie}$**
PPTM	μ(ROS) [samogłosek/s]	3,87	6,43	7,58
	σ(ROS)[samogłosek/s]	0,55	0,62	0,92

4.1.2. Test rozumienia mowy przyspieszonej

Test rozumienia mowy przyspieszonej (ang. *Time Compressed Speech Test* – TCST) należy do grupy testów rozumienia mowy niskoredundantnej (ang. *Low-redundancy*

speech test) [113]. Testy te przeznaczone są do diagnozowania ośrodkowych zaburzeń słuchu. TCST przeprowadzany jest w celu oceny rozdzielczości czasowej słuchu i jest on alternatywę dla popularnego testu rozpoznawania przerw (ang. *Random Gap Detection Test* – RGDT) [191] [111] [66]. Test RDGT polega na odtwarzaniu osobie badanej sekwencji dwóch tonów prostych lub dwóch trzasków. Przerwa pomiędzy bodźcami jest zmienna, a badanie pozwala wyznaczyć minimalny czas przerwy słyszanej przez pacjenta. Uzyskana minimalna długość przerwy wyznacza próg rozdzielczości czasowej słuchu osoby badanej. W ramach badań nie wykorzystano testu RGDT, ponieważ jego wyniki nie odnoszą się bezpośrednio do procesu rozumienia mowy.

Test TCST polega na odtwarzaniu słuchaczowi przyspieszonych zdań lub słów. Modyfikacja tempa mowy wykonywana jest z wykorzystaniem jednego z algorytmów TSM. W teście badana jest procentowa liczba poprawnie rozpoznawanych słów dla różnych wartości współczynnika skali [66][9]. Należy zauważyć, iż wynik testu TCST nie niesie pełnej informacji na temat rozdzielczości słuchowej osoby badanej, ponieważ jego wynik zależy od tempa mowy oryginalne wykorzystanej w teście[10]. Dlatego w badaniach użyto zmodyfikowanej przez Versfelda [181] wersji testu TCST. Modyfikacja ta została wprowadzona tak, by wynikiem badania była wartość tempa mowy, dla której badana osoba jest w stanie poprawnie zrozumieć 50% zdań. Jest ona wyrażana w liczbie sylab na sekundę odpowiadających progowi rozumienia 50% zdań (ang. *50% time-compressed speech threshold* – TCT_{50}). Próg TCT_{50} jest pewnego rodzajem analogią do progu SRT_{50} (ang. *Speech Reception Threshold*) uzyskiwanego w testach rozumienia mowy w szumie [145]. Dodatkowo Versfeld wprowadził mechanizm adaptacyjnego dobierania wartości współczynników skali. Procedura testu jest następująca:

- w teście wykorzystuje się 13 różnych zdań,
- wartość współczynnika skali modyfikowana jest dla każdego zdania zgodnie z następującymi regułami:
    - jeżeli wszystkie słowa w zdaniu zostały rozpoznane poprawnie, to wartość współczynnika skali jest zwiększana,
    - w przeciwnym razie wartość współczynnika skali jest zmniejszana.

[9] W teście wykorzystuje się kilka stałych wartości współczynnika skali
[10] Mowa o niskim tempie przyspieszona o 50 % jest prostsza do zrozumienia niż mowa szybka poddana tej samej modyfikacji.

- próg TCT_{50} jest wyznaczany jako średnia geometryczna wartości ROS obliczonych dla 10 ostatnich zdań testu.

Wartość TCT_{50} pozwala określić, jakie tempo jest maksymalnym tempem mowy poprawnie rozumianej przez słuchacza. Informacja ta jest istotna, ponieważ umożliwia określić, to czy dana osoba ma problemy z rozdzielczością czasową słuchu. W ramach prac nad rozprawą wykonano w środowisku MATLAB implementację opisanego powyżej testu TCST. Została ona wykorzystana w badaniach opisanych w tym rozdziale. Podczas testów modyfikacji poddawane były zdania tworzone z wykorzystaniem testów macierzowych PTM i PPTM. Do syntezy zdań wykorzystano mowę zarejestrowaną w tempie „wolnym", co umożliwiło analizę rozumienia mowy z szerokiego zakresu temp. Jako że opracowane metody modyfikacji sygnału mowy bazują na tempie określonym w liczbie samogłosek/s, próg TCT_{50} wyznaczany w ramach tego testu, był określany także w tej jednostce, a nie jako liczba sylab/s, jak proponował Versfeld [181].

4.1.3. Test rozumienia mowy spowolnionej

W ramach TRMS zbadano stopień rozumienia mowy spowolnionej z wykorzystaniem metod modyfikacji sygnału opracowanych w ramach rozprawy. Badanie składało się z dwóch części, w których użyto różnego tempa mowy oryginalnej. W badaniach wykorzystano ten sam materiał zdaniowy, co w teście TCST. Jedyną różnicą było to, iż tempa wypowiadanej mowy było „średnie" i „szybkie" (tab 4.3). W każdej części testu słuchacz miał za zadanie powtórzyć 40 losowych i niepowtarzających się zdań. Zbiór składał się z czterech dziesięciozdaniowych podzbiorów, w skład których wchodziły: zdania niezmodyfikowane oraz zdania zmodyfikowane za pomocą metod A, B i C. Na podstawie odpowiedzi wyznaczano liczbę błędnie powtarzanych słów (ang. *Word Error Rate* – WER) oraz średnią poprawę rozumienia mowy (PRM). Obie wartości obliczono niezależnie dla każdego podzbioru zdań. WER został obliczony, jako procent niepoprawnie powtórzonych słów, a PRM, jako różnica pomiędzy WER uzyskanym dla mowy niemodyfikowanej a WER osiągniętym dla mowy zmodyfikowanej za pomocą jednego z algorytmów.

W badaniu wykorzystano następujące metody modyfikacji sygnału:

A. Równomierne spowalnianie mowy za pomocą algorytmu SOLA

B. Nierównomierne spowalnianie sygnału mowy sterowane wartością współczynnika skali (algorytm opracowany w ramach rozprawy)

C. Nierównomierne spowalnianie sygnału mowy sterowane wartością maksymalnego tempa mowy (algorytm opracowany w ramach rozprawy)

Wartości współczynnika skali α_o dla metod A i B ustawiono na 1,5, a dla metody C wartość ROS_o wynosiła 5,5 samogłoski/s. Wartości te zostały dobrane na podstawie wyników badań przeprowadzonych przez Nejime *et al.* [122], gdzie wykazano, iż największą poprawę w rozumieniu mowy można zaobserwować dla współczynnika skali wynoszącego 1,5. Jak pokazały badania Nejime *et al.* wyższe wartości współczynnika skali powodowały spadek rozumienia wypowiedzi. Wartość ROS_o została dobrana w taki sposób, by wydłużenie sygnału było takie samo, jak w zastosowaniu metody B.

Metoda A została wykorzystana w badaniach w celu zbadania, czy spowolnienie sygnału z wykorzystaniem tradycyjnych metod równomiernego spowalniania sygnału pozwala na uzyskanie tego samego efektu, w postaci poprawy rozumienia mowy, co w zastosowaniu jednej z metod opracowanych w ramach rozprawy. Należy tu podkreślić, iż metoda A nie może zostać bezpośrednio wykorzystana do spowolniania sygnału w czasie rzeczywistym, ponieważ nie zawiera ona mechanizmów zapewniających synchronizację sygnału wejściowego i wyjściowego. Testy rozumienia mowy zostały zaimplementowany w środowisku MATLAB.

4.1.4. Przebieg badania

Obie grupy słuchaczy składały się z 17 ochotników. Mediana wieku w grupie dzieci głuchych wynosiła 9 lat a minimalny i maksymalny wiek odpowiednio 4 i 14 lat. W teście udział wzięło 9 dziewczynek i 8 chłopców. W grupie osób starszych mediana wieku wynosiła 83 lata, a zakres wieku mieścił się w granicach 64–91 lat. W teście udział wzięło 11 kobiet i 6 mężczyzn. U osób starszych wykonano dodatkowo badanie audiometryczne pozwalające wyznaczyć próg słyszenia. Wyniki badań audiometrycznych dzieci udostępnione zostały dzięki uprzejmości Specjalistycznego Ośrodka Diagnozy i Rehabilitacji Dzieci i Młodzieży z Wadą Słuchu Polskiego Związku Głuchych w Gdańsku.

Badania przeprowadzono korzystając z oprogramowania zaimplementowanego w środowisku MATLAB. Dźwięk odtwarzany był za pośrednictwem obuusznych słuchawek, podłączonych do komputerowej karty dźwiękowej za pośrednictwem wzmacniacza słuchawkowego. Zastosowanie wzmacniacza pozwoliło na dopasowywanie poziomu sygnału do potrzeb każdego słuchacza osobno w taki sposób, by zapewnić mu komfortowy

odsłuch. Wszystkie osoby korzystające na co dzień z aparatów słuchowych lub implantów ślimakowych miały urządzenia włączone podczas badań.

Osoby starsze brały udział w badaniu trwającym 40 minut. Testy wykonywane przez dzieci głuche zostały podzielone na dwie sesje, ponieważ zbyt długie badanie powodowało u nich problemy z koncentracją i wyniki testów mogłyby być obciążone dodatkowym błędem. Podczas pierwszej sesji dzieci wykonywały test TCST i TRMS przygotowany dla mowy oryginalnie wypowiadanej w tempie „szybkim" (sesja trwała około 20 minut). Podczas drugiej sesji wykonywano test TRMS z wykorzystaniem mowy wypowiadanej w tempie „średnim" (sesja trwała około 12 minut).

4.2 Wyniki badań

Uzyskane wyniki zostały przeanalizowane w celu wykazania pierwszej tezy rozprawy mówiącej, iż: **nierównomierna modyfikacja czasu trwania sygnału powoduje wzrost współczynnika rozumienia mowy u osób o pogorszonej rozdzielczości czasowej słuchu**. Współczynnikiem określającym rozumienie mowy była tu procentowa liczba błędnie powtórzonych słów (WER). Podczas badań, jako miarę rozdzielczości czasowej słuchu wykorzystano parametr TCT_{50}. Dla obu grup słuchaczy wyznaczono średnią wartość progu TCT_{50}, odchylenie standardowe oraz 95% przedział ufności. Wyniki te zamieszczono w tab. 4.4.

Tab. 4.4 Wyniki testu TCST.

Grupa słuchaczy	TCT_{50}[samogłosek/s]		
	Wartość średnia	Odchylenie standardowe	95% przedział ufności
Dzieci	6,33	1,21	5,7/6,96
Osoby starsze	4,81	1,59	3,99/5,63

Test TCST na tym samym materiale zdaniowym wykonał także Ozimek *et al.* [129]. Grupa słuchaczy składała się z 30 osób dorosłych ze słuchem prawidłowym (średni wiek 37,3 lat) oraz 31 osób starszych z ubytkiem słuchu (średni wiek 72,4 lata). Wartości średnie TCT_{50} uzyskane podczas tamtych badań wynosiły odpowiednio 11 sylab/s i 6,4 sylab/s. Można zauważyć, iż wartości progów dla osób starszych przedstawione w badaniach Ozimka *et al.* są wyższe od tych uzyskanych podczas testów przeprowadzonych w ramach tych badań. Może to być związane z następującymi czynnikami:

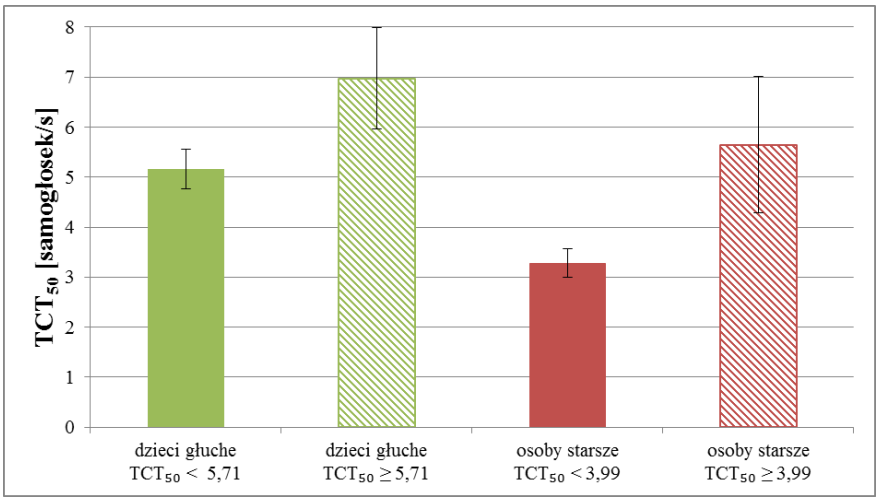
- różną liczebnością grup słuchaczy,

- różnym tempem, mowy oryginalnej użytej w badaniu: Ozimek *et al.* nie podali tempa mowy oryginalnej użytej podczas testu, a to, od jakiego tempa mowy rozpoczyna się badanie, ma wpływ na to, jaką maksymalną wartość tempa można osiągnąć, oraz na to, jakie najniższe wartości tempa mowy będą brane pod uwagę podczas obliczania średniej geometrycznej wyznaczającej próg,
- procedurą wyznaczania tempa mowy – Ozimek *et al.* nie określili sposoby obliczania tempa mowy,
- różnym stopieniem ubytku rozdzielczości czasowej słuchu.

W celu podziału obu badanych grup słuchaczy na osoby z normalną i pogorszoną rozdzielczością czasową słuchu wykorzystano założenia poczynione przez Versfelda [181]. Mówią one, że osoby, dla których wartość progu TCT_{50} jest niższa niż dolna wartość 95% przedziału ufności danej grupy, mają pogorszoną rozdzielczości czasowej słuchu. Na tej podstawie przyjęto dwa progi TCT_{50} dzielące obie grupy słuchaczy na podzbiory zawierające:

- dzieci głuche z normalną rozdzielczością czasową słuchu,
- dzieci głuche z pogorszoną rozdzielczością czasową słuchu,
- osoby starsze z normalną rozdzielczością czasową słuchu,
- osoby starsze z pogorszoną rozdzielczością czasową słuchu.

Przy czym przez normalną rozdzielczość czasową słuchu rozumie się rozdzielczość będącą typową dla danej grupy słuchaczy. Próg dzielący grupę dzieci głuchych wynosił 5,71 samogłosek/s, a próg dzielący osoby starsze 3,99 samogłosek/s. Na rys. 4.1 przedstawiono średnie wartości progu TCT_{50} wyznaczone dla każdego podzbioru słuchaczy. Wartości średnie progów TCT_{50} dla osób powyżej granicy normalnej rozdzielczości czasowej słuchu są wyższe o 1,8 samogłoski/s w grupie dzieci i o 2,3 samogłosek/s w grupie osób starszych od grup osób, których próg TCT_{50} znajduje się poniżej tej granicy. W podzbiorach osób o pogorszonej rozdzielczości czasowej słuchu znalazło się sześcioro dzieci i sześć osób starszych, a w podzbiorach osób o normalnej rozdzielczości czasowej słuchu jedenaścioro dzieci i jedenaście osób starszych.



Rys. 4.1 Wartości progu TCT_{50} dla dzieci głuchych i osób starszych.

W dalszej części tego podrozdziału przedstawiono analizę wyników testu rozumienia mowy spowolnionej. Została ona przeprowadzono niezależnie dla każdej z grup słuchaczy. W celu zbadania, czy opracowane metody modyfikacji sygnału mowy mają istotny statystycznie wpływ na poprawę jej rozumienia wykonano analizę różnić średnich wartości błędów popełnionych przez każdą grupę słuchaczy podczas słuchania mowy niezmodyfikowanej oraz zmodyfikowanej za pomocą metod A, B, i C. Ocena istotności została wykonana za pomocą dwóch testów: jednoczynnikowego testu ANOVA z powtórzeniami (ang. *Repeated Measures ANalysis Of Variance* – RM ANOVA), oraz jego nieparametrycznego odpowiednika – testu Friedmana. Test RM ANOVA bada wpływ różnych warunków pomiarowych (np. pomiar tego samego parametru wykonanego w różnych odstępach czasu lub bada wpływ różnych metod terapii na ten sam czynnik) na tą samą grupę badanych. Czynnikiem w obu testach była procentowa liczba błędów popełnionych przez słuchaczy w obrębie jednej z grup (np. dzieci głuche z pogorszoną rozdzielczością czasową słuchu) osiągnięta dla mowy odtwarzanej w różnych warunkach (brak modyfikacji lub modyfikacja za pomocą jednej z metod). Test nieparametryczny był wykorzystywany tylko wtedy, gdy dane poddawane analizie nie spełniały założeń parametrycznego testu RM ANOVA. Założeniami testu RM ANOVA są:

- rozkład normalny danych uzyskanych podczas każdego z pomiarów,
- sferyczność danych w obrębie grupy (ang. *sphericity*).

Rozkład normalny zbadano z wykorzystaniem nieparametrycznego testu Shapiro-Wilka [165], a sferyczność za pomocą testu Mauchly'ego [99]. Hipotezą zerową testu Shapiro-Wilka jest rozkład normalny danych wewnątrz grupy. W przypadku, gdy uzyskana w

teście wartość istotności (ang. *p-value*) jest mniejsza od założonego poziomu istotności[11], przyjmuje się hipotezę alternatywną mówiącą o tym, iż dane nie mają rozkładu normalnego. W rozprawie dla wszystkich testów statystycznych przyjęto wartość $p_i = 0{,}05$. W teście Mauchly'ego hipoteza zerowa zakłada sferyczny rozkład danych. Oznacza to równość wariancji różnic pomiędzy poziomami powtórzonych czynników testu RM ANOVA. Hipoteza alternatywna mówi o braku sferyczności danych. Szczegółowe wyniki analizy rozkładu danych oraz ich sferyczności przedstawiono w załączniku nr 3.

4.2.1. Rozumienie mowy zmodyfikowanej przez dzieci głuche

W grupie dzieci głuchych znajdowało się dziewięcioro dzieci z implantem ślimakowym i ośmioro dzieci noszących aparat słuchowy. W tab. 4.5 i 4.6 umieszczono wyniki badań uzyskane dla grupy z obniżoną i normalną rozdzielczością czasową słuchu. Dla celów informacyjnych każdy ze słuchaczy został oznaczony jednym z dwóch symboli: IS (implant ślimakowy) lub AS (aparat słuchowy). Tabele zawierają średnie wartości progu słyszenia, wartości progów TCT_{50} oraz wartości WER wyznaczone dla mowy niezmodyfikowanej wypowiadanej w tempie $ROS^{PP}_{\text{średnie}}$ i $ROS^{PP}_{\text{szybkie}}$. W załączniku nr 4 zamieszczono szczegółowe wyniki badań audiometrycznych oraz wartości PRM uzyskane dla poszczególnych metod modyfikacji jak również dla różnych temp mowy oryginalnej.

Jak można zauważyć (tab. 4.5 i 4.6), procent błędnie powtórzonych słów uzyskany dla mowy wypowiadanej w tempie średnim jest znacznie mniejszy od tej liczby dla mowy wypowiadanej w tempie szybkim. Prawidłowość ta widoczna jest w przypadku obydwu grup (dzieci głuchych z pogorszoną i normalną rozdzielczością czasową słuchu). Może to być związane z tym, że średnia wartość tempa $ROS^{PP}_{\text{średnie}}$ (6,43 samogłosek/s) jest bliska średniej wartość progu TCT_{50} wyznaczonej dla obu grup (odpowiednio $\mu(TCT_{50}) = 5{,}16$ samogłosek/s i $\mu(TCT_{50}) = 6{,}98$ samogłosek/s). Natomiast druga wartość tempa ($ROS^{PP}_{\text{szybkie}} = 7{,}48$ samogłosek/s) przekracza średnią wartość progu TCT_{50} w obu grupach. Widoczna jest również nieznaczna różnica w liczbie popełnianych błędów pomiędzy grupami dzieci głuchych z pogorszoną i normalną rozdzielczością czasową słuchu. W przypadku tempa $ROS^{PP}_{\text{szybkie}}$ średnie wartości WER wyniosły odpowiednio 56,11% ($TCT_{50} < 5{,}71$) i 40,15% ($TCT_{50} \geq 5{,}71$), a dla tempa $ROS^{PP}_{\text{średnie}}$ 17,78% oraz 13,63%. Dlatego można założyć, iż obie grupy dzieci miały podobne problemy ze zrozumieniem mowy niezmodyfikowanej.

[11] Typowo oznaczane symbolem α. Jednak w ramach rozprawy symbol ten odnosi się do współczynnika skali dlatego poziom istotności oznaczono tu jako p_i.

Tab. 4.5 Wyniki badań słuchu wykonane dla dzieci głuchych z obniżoną rozdzielczością czasową słuchu ($TCT_{50} < 5{,}71$ samogłosek/s)

L.p.	Wiek	Próg słyszenia [dB HL]	TCT_{50} [samogłosek/s]	WER [%] $ROS^{PP}_{średnie}$	WER [%] $ROS^{PP}_{szybkie}$
1 (IS)	9	85	5,04	23,33	58,33
2 (IS)	9	95	5,34	26,67	53,33
3 (AS)	14	100	5,01	36,67	66,67
4 (AS)	12	68,75	4,53	0	62,50
5 (AS)	12	65	5,39	13,33	66,67
6 (AS)	6	33,75	5,66	6,67	29,17
wartość średnia	10,33	74,58	5,16	17,78	56,11
odchylenie standardowe	2,87	24,35	0,39	13,61	14,16

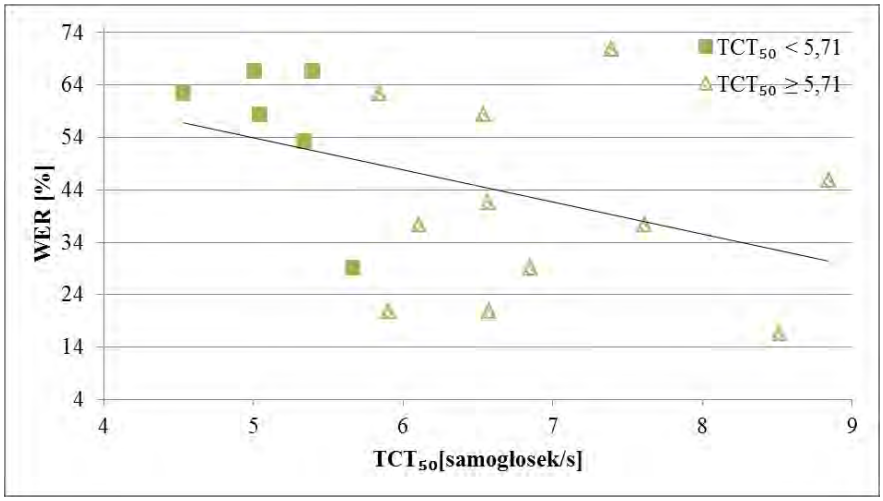
Tab. 4.6 Wyniki badań słuchu wykonane dla dzieci głuchych z normalną rozdzielczością czasową słuchu ($TCT_{50} \geq 5{,}71$ samogłosek/s)

L.p.	Wiek	Próg słyszenia [dB HL]	TCT_{50} [samogłosek/s]	WER [%] $ROS^{PP}_{średnie}$	WER [%] $ROS^{PP}_{szybkie}$
1 (IS)	14	100	5,92	10,00	20,83
2 (IS)	14	90	5,84	3,33	62,50
3 (IS)	8	100	6,54	23,33	58,33
4 (IS)	6	110	6,56	6,67	41,67
5 (IS)	14	86,25	6,10	23,33	37,50
6 (IS)	4	85	7,61	10,00	37,50
7 (IS)	12	105	6,57	13,33	20,83
8(AS)	4	35	7,39	23,33	70,83
9 (AS)	6	26,25	8,51	0	16,67
10 (AS)	7	44,375	8,85	33,33	45,83
11 (AS)	8	55	6,85	3,33	29,17
wartość średnia	8,82	76,08	6,98	13,63	40,15
odchylenie standardowe	3,97	30,21	1,01	10,69	17,99

Relacja pomiędzy wartością TCT_{50} a współczynnikiem WER dla mowy niezmodyfikowanej wypowiadanej w tempie szybkim

W celu zbadania czy istnieje relacja pomiędzy wyznaczoną w badaniu wartością rozdzielczości czasowej słuchu wyrażoną w formie progu TCT_{50}, a rozumieniem szybko wypowiadanej mowy opisanym liczbą błędnie powtórzonych słów (WER), wyznaczono współczynnik korelacji linowej Pearsona. Na rys. 4.2 przedstawiono zależność pomiędzy wartościami TCT_{50} a liczbą błędów, uzyskaną dla wszystkich dzieci głuchych biorących udział w badaniach. Kwadratami oznaczono dzieci z pogorszoną rozdzielczością czasową

słuchu a trójkątami dzieci z prawidłową rozdzielczością czasową słuchu. Linia ciągła reprezentuje linię trendu obliczoną dla wszystkich dzieci biorących udział w badaniach. Analizę wykonano osobno dla całej grupy dzieci głuchych oraz dla grupy dzieci głuchych z normalną i obniżoną rozdzielczością czasową słuchu. Dla wymienionych powyżej grup słuchaczy wartości współczynnika korelacji wyniosły odpowiednio: –0,41, –0,09 i –0,62. Wskazują one, iż jedynie u dzieci głuchych z pogorszoną rozdzielczością czasową słuchu można dopatrywać się liniowej korelacji pomiędzy progiem TCT_{50} a stopniem rozumienia mowy wyrażonym poprzez WER. Ujemny znak współczynnika korelacji świadczy o zależności odwrotnej. Przekłada się to na następujący wniosek dotyczący dzieci głuchych z pogorszoną rozdzielczością czasową słuchu: im mniejszą wartość TCT_{50} uzyskano w teście TCST, tym większą liczbę błędów popełni dane dziecko w teście rozumienia mowy.

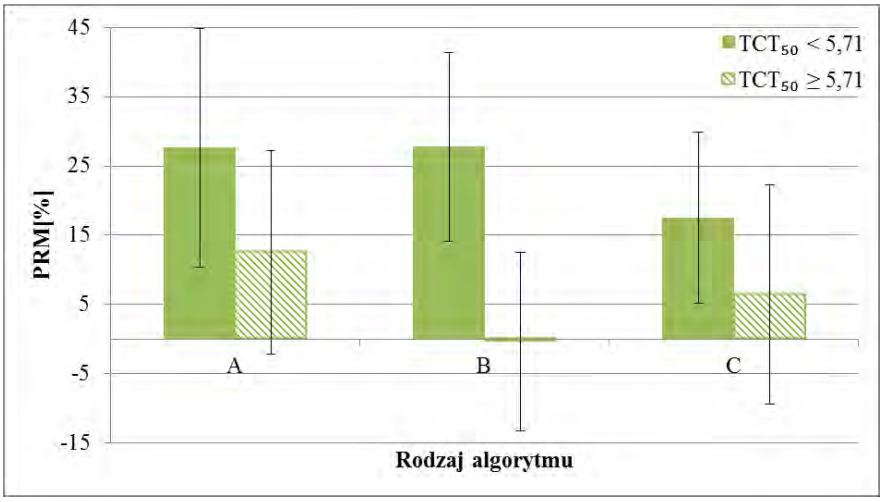


Rys. 4.2 Relacja pomiędzy progiem TCT_{50} a wartością WER wyznaczona dla dzieci głuchych dla mowy wypowiadana w tempie $ROS^{PP}_{szybkie}$.

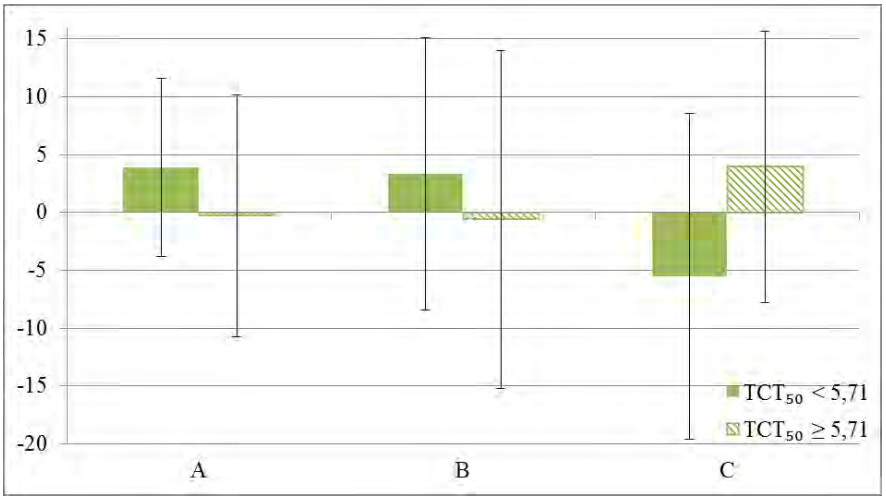
Analiza wpływu modyfikacji tempa na rozumienie mowy

Wpływ opracowanych metod modyfikacji sygnału mowy został oceniony niezależnie dla dwóch grup słuchaczy ($TCT_{50} < 5{,}71$ i $TCT_{50} \geq 5{,}71$) oraz osobno dla mowy wypowiadanej w tempie szybkim ($ROS^{PP}_{szybkie}$) i średnim ($ROS^{PP}_{średnie}$). Podczas analizy wpływu metod wykorzystano wyniki testu TRMS. Wartości PRM uzyskane podczas badań dla obu grup zamieszczono na rys. 4.3 (mowa wypowiadana w tempie szybkim) i rys. 4.4 (mowa wypowiadana w tempie średnim). Znaczącą poprawę w rozumieniu mowy można zaobserwować głównie dla mowy wypowiadanej w tempie szybkim. Jest ona wyraźna dla obu grup słuchaczy. U dzieci z pogorszoną rozdzielczością czasową słuchu wartości PRM mieszczą się w przedziale od 17,5% (metoda C) do 27,78 % (metoda B). Wpływ opracowanych metod na rozumienie mowy przez dzieci głuche z normalną rozdzielczością

słuchu jest także zauważalny jednak jedynie w metodach A i C (12,49% i 6,44%). Metoda B miała prawie niezauważalny wpływ na wartość współczynnika PRM (–0,38%).



Rys. 4.3 Poprawa rozumienia mowy w grupie dzieci głuchych uzyskana dla mowy wypowiadanej w tempie szybkim (ROS_{mean} = 7,58 [samogłosek/s])



Rys. 4.4 3 Poprawa rozumienia mowy w grupie dzieci głuchych uzyskana dla mowy wypowiadanej w tempie średnim (ROS_{mean} = 6,43 [samogłosek/s])

Analiza wpływu modyfikacji tempa na rozumienie mowy wypowiadanej w tempie szybkim

W celu weryfikacji istotności statystycznej uzyskanych wyników przeprowadzono serię testów analizujących wartości WER. Jako, że nie wszystkie rozkłady wartości WER dla kolejnych prób (mowa niezmodyfikowane i mowa zmodyfikowana z wykorzystaniem opracowanych metod) w grupie dzieci z progiem $TCT_{50} < 5,71$ miały rozkład normalny, dla danych wykonano analizę z wykorzystaniem nieparametrycznego testu Friedmana. W grupie dzieci z progiem $TCT_{50} \geq 5,71$ wartość WER dla kolejnych metod modyfikacji (lub jej braku) miały rozkład normalny. Dodatkowo dane miały rozkład sferyczny

($\chi^2(5) = 11{,}38$; $p = 0{,}051$). W związku z tym możliwe było wykonanie parametrycznego testu RM ANOVA.

W grupie dzieci z pogorszoną rozdzielczością czasową słuchu zamiast analizy wartości istotności wykonano porównanie wartość statystyki testu z wartościami tablicowymi uwzględniającymi niewielką liczebność grupy (poniżej 8 osób)[12]. Hipotezą zerową testu, była zgodność rozkładu prawdopodobieństwa WER w ramach różnych algorytmów. W grupie z pogorszoną rozdzielczością słuchu wartość statystyki testu wyniosła $\chi^2(3) = 113{,}5$, a wartość krytyczna odczytana z tablic wynosi $\chi^2(3)_{cv} = 76$. Ponieważ wartość statystyki jest większa od wartości krytycznej, należy odrzucić hipotezę zerową testu i tym samym przyjąć hipotezę alternatywną mówiącą, iż przynajmniej jedna dystrybuanta wartości WER różni się od pozostałych dystrybuant WER. W celu zbadania, które rozkłady prawdopodobieństw WER różnią się między sobą wykonano nieparametryczny test *post hoc* będący odpowiednikiem parametrycznego testu Fishera LSD (ang. *Least Significant Difference*) [18]. W tab. 4.7 zamieszczono wyniki testu. Wartości otrzymanych statystyk porównano z wartością krytyczną równą 6,2. W wyniku porównania można zauważyć, iż hipoteza zerowa została odrzucona ze względu na różnice dystrybuant WER występujące pomiędzy mową niezmodyfikowaną a mową spowolniona z wykorzystaniem metod A, B i C. Oznacza to, iż wszystkie metody modyfikacji sygnału mowy miały istotny statystycznie wpływ na poprawę jej rozumienie przez dzieci z pogorszoną rozdzielczością czasową słuchu.

Tab. 4.7 Wyniki nieparametrycznego testu *post hoc* uzyskane dla grupy dzieci z pogorszoną rozdzielczością czasową słuchu.

rodzaj modyfikacji	**brak**	**metoda A**	**metoda B**	**metoda C**
brak	0	12.5	13.5	8.0
metoda A	12.5	0	1.0	4.5
metoda B	13.5	1.0	0	5.5
metoda C	8.0	4.5	5.5	0

W drugim teście za pomocą testu RM ANOVA zbadano hipotezę zerową dla grupy dzieci z normalną rozdzielczością czasową słuchu. Hipotezą zerową testu było założenie, iż wartości średnie WER uzyskane dla mowy niezmodyfikowanej oraz mowy zmodyfikowanej z wykorzystaniem metod A–C są równe. Wartości tych statystyk wyniosły $F(3,30) = 3{,}44$ i $p = 0{,}03$. Uzyskany poziom istotności testu jest niższy od

[12] Metoda ta przy wykonywaniu testu Friedmana dla nielicznej grupy pozwala uzyskać wynik o wyższej wiarygodności.

założonego p_i, więc hipotezę zerową należy odrzucić i przyjąć hipotezę alternatywną mówiącą, iż przynajmniej jedna ze średnich różni się znacząco od pozostałych. W celu określenie tego, które średnie różnią się od siebie wykonano test *post hoc*. Został on przeprowadzony zgodnie z metodą LSD (ang. *Least Significant Difference*) opracowaną przez Fishera [54]. W metodzie tej dokonuję się porównania wartości średnich pomiędzy parami z wykorzystaniem t-testu. Hipotezą zerową t-testu jest równość średnich w parze. W tab. 4.8 przedstawiono wyniki porównań. Jedynie dla pierwszej i czwartej pary wartości poziomu istotności p są mniejsza od p_i. Oznacza to, że różnice pomiędzy średnimi wartości WER w tych parach są istotne statystycznie. Istotność różnic w parze pierwszej świadczy o tym, iż zastosowanie metody A pozwala na uzyskanie poprawy w rozumieniu mowy u dzieci z normalną rozdzielczością czasową słuchu. Nie wykryto natomiast statystycznie istotnych różnic pomiędzy rozumieniem mowy niezmodyfikowanej, a rozumieniem mowy zmodyfikowanej za pomocą metod B i C.

Tab. 4.8 Wyniki parametrycznego testu *post hoc* uzyskane dla dzieci z normalną rozdzielczością czasową słuchu.

L.p.	Para	Wartość statystyki testu (t)	Ilość stopni swobody	Istotność (p)
1	brak modyfikacji – metoda A	–2,83	10	**0,02**
2	brak modyfikacji – metoda B	0,10	10	0,92
3	brak modyfikacji – metoda C	–1,35	10	0,20
4	metoda A – metoda B	2,54	10	**0,03**
5	metoda A – metoda C	1,72	10	0,12
6	metoda B – metoda C	–1,16	10	0,27

Analiza wpływu modyfikacji tempa na rozumienie mowy wypowiadanej w tempie średnim

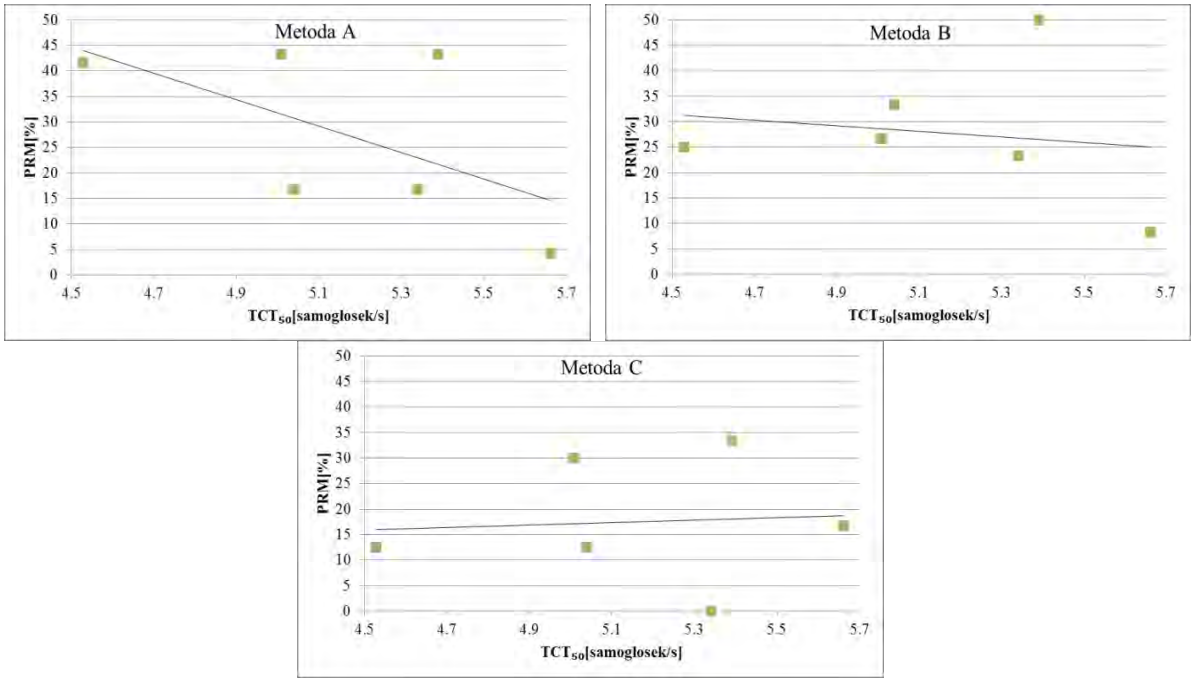
Dla mowy wypowiadanej w tempie średnim można zaobserwować niewielką PRM jedynie u dzieci głuchych z pogorszona rozdzielczością czasową słuchu przy zastosowania metod A i B (odpowiednio 3,89 % i 3,34 %) oraz u dzieci głuchych z normalną rozdzielczością czasową słuchu dla metody C (3,93 %). W pozostałych badaniach zaobserwowano nieznaczne pogorszenie rozumienia mowy mieszczące się przedziale od –5,55% do –0,30%. W celu sprawdzenia istotności statystycznej różnic w stopniu rozumienia mowy niezmodyfikowanej i zmodyfikowanej z wykorzystaniem opracowanych metod dla grupy dzieci głuchych z pogorszoną rozdzielczością słuchu wykonano test RM ANOVA. Test ten mógł być wykonany, ponieważ wartości WER w obrębie metod

miały rozkład normalny i dodatkowo założenie o sferyczności rozkładów nie zostało odrzucono. W teście Muachly'ego uzyskano wartość statystyki $\chi^2(5) = 4{,}55$ i $p = 0{,}49$. W grupie dzieci głuchych z normalną rozdzielczością słuchu, nie możliwe było wykonanie testu parametrycznego, ze względu na brak rozkładu normalnego wartości WER dla wszystkich metod modyfikacji w ramach tej podgrupy (załącznik nr 5). Dlatego w analizie użyto nieparametrycznego testu Friedmana. Dla grupy dzieci głuchych z pogorszoną rozdzielczością czasową słuchu wartość statystyki wyniosła $F(3,15) = 1{,}51$ przy $p = 0{,}25$. Jako że uzyskana wartość p jest większa od p_i, nie ma podstaw do odrzucenia hipotezy zerowej testu. Stąd wniosek, iż nie istnieje statystycznie istotna różnica pomiędzy rozumieniem mowy niezmodyfikowanej oraz zmodyfikowanej za pomocą opracowanych metod przez dzieci głuche z pogorszoną rozdzielczością czasową słuchu. Dla grupy dzieci głuchych z normalną rozdzielczością czasową słuchu wartość statystyki testu Friedmana wyniosła $\chi^2(3) = 0{,}30$, a poziom istotności $p = 0{,}82$. Ponieważ poziom istotności znacznie przewyższa założoną wartość poziomu istotności p_i, hipoteza zerowa testu nie została odrzucona. Oznacza to, iż nie istnieją istotne statystycznie różnice pomiędzy rozumieniem mowy niezmodyfikowanej oraz zmodyfikowane z wykorzystaniem metod A–C w grupie dzieci z normalną rozdzielczością czasową słuchu.

Relacja pomiędzy wartością TCT_{50} a współczynnikiem PRM dla mowy zmodyfikowane wypowiadanej w tempie szybkim

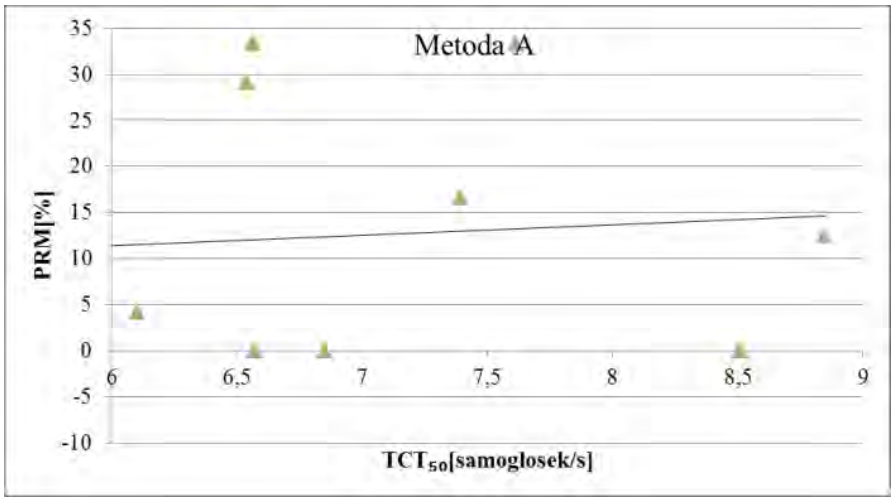
W kolejnym kroku zbadano czy istnieje liniowa zależność pomiędzy poprawą rozumienia mowy osiągniętą w badaniu z użyciem mowy szybkiej dla grupy dzieci z normalną i pogorszoną rozdzielczością czasową słuchu, a wartością progu TCT_{50} osiągniętą w badaniu TCST. Przeanalizowano tą relację, jedynie dla metod modyfikacji, dla których wykazano istotne statystycznie różnice pomiędzy rozumieniem mody niezmodyfikowanej i mowy zmodyfikowanej. Na rys. 4.5 przedstawiono relację pomiędzy progiem TCT_{50} a PRM wyznaczoną dla dzieci z pogorszoną rozdzielczością czasową słuchu dla trzech metod modyfikacji sygnału mowy. Współczynnik korelacji pomiędzy tymi wartościami wyniósł odpowiednio dla metody A –0,59, dla metody B –0,16 oraz dla metody C 0,08. Wynik ten pozwala zakładać, iż jedynie dla metody A można zaobserwować bezpośrednią liniową relację pomiędzy progiem TCT_{50}, a wpływem modyfikacji sygnału na jej rozumienie. Co więcej, w tej metodzie można oczekiwać tego, że wraz ze wzrostem wartości progu TCT_{50} poprawa rozumienia mowy będzie maleć. Należy podkreślić, iż zależność ta spełniona jest jedynie w przypadku dzieci głuchych z

obniżoną rozdzielczością czasową słuchu przy modyfikacji mowy z wykorzystaniem metody A.



Rys. 4.5 Relacja pomiędzy progiem TCT_{50} a wartością PRM wyznaczona dla grupy dzieci z progiem $TCT_{50} < 5{,}71$ dla metody A, B i C.

Relację pomiędzy wartością TCT_{50} a współczynnikiem PRM dla dzieci głuchych z normalną rozdzielczością czasową słuchu przedstawiono na rys. 4.6. Dla tej grupy dzieci głuchych jedynie metoda A pozwoliła osiągnąć istotne statystycznie wyniki. Współczynnik korelacji Pearsona dla tych danych wyniósł 0,07. Oznacza to, iż nie istnieje korelacja linowa pomiędzy wartością progu TCT_{50} a współczynnikiem PRM dla dzieci głuchych z normalną rozdzielczością czasową słuchu.



Rys. 4.6 Relacja pomiędzy progiem TCT_{50} a wartością PRM wyznaczona dla grupy dzieci z progiem $TCT_{50} \geq 5{,}71$ dla metody A.

4.2.2. Rozumienie mowy zmodyfikowanej przez osoby starsze

W grupie osób starszych znajdowały się trzy osoby noszące aparat słuchowy. W tab. 4.9 i 4.10 zamieszczono wyniki badań uzyskane dla osób starszych z pogorszoną i normalną rozdzielczością czasową słuchu. Osoby noszące aparat słuchowy zostały wyróżnione poprzez umieszczenie przy ich numerze symbolu AS. Tab. 4.9 i 4.10 przedstawiają średnią wartość progu słyszenia, wartości progów TCT_{50} oraz wartości WER wyznaczone dla mowy niezmodyfikowanej wypowiadanej w tempie $ROS^{P}_{średnie}$ i $ROS^{P}_{szybkie}$. Szczegółowe wyniki badań audiometrycznych oraz wartość współczynników PRM dla różnych metod modyfikacji sygnału mowy umieszczono w załączniku nr 5. Tak jak dla grupy dzieci głuchych, tak i wśród osób starszych zaobserwowano, iż procentowa liczba błędnie powtórzonych słów uzyskana dla mowy wypowiadanej w tempie szybkim jest wyższa, niż mowy wypowiadanej w tempie średnim. Przyczyny tej prawidłowości można dopatrywać się w fakcie, iż $ROS^{P}_{szybkie} = 6{,}48$ samogłosek/s, więc przekracza ona $\mu(TCT_{50})$ wyznaczone dla obu grup słuchaczy. Inaczej, niż w grupie dzieci głuchych, dla grupy osób starszych widać większą różnicę między liczbą błędnie powtórzonych słów pomiędzy grupami osób z pogorszoną rozdzielczością czasową słuchu (31,67% i 51,33%) i normalną rozdzielczością czasową słuchu (6,18%, 12,36%).

Tab. 4.9 Wyniki badań słuchu wykonane dla osób starszych z pogorszoną rozdzielczością czasową słuchu ($TCT_{50} < 3{,}99$ samogłosek/s)

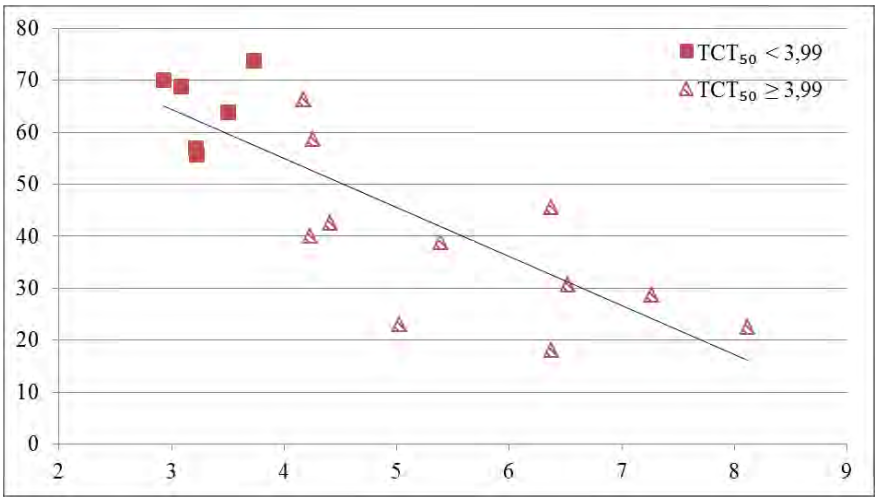
L.p.	**Wiek**	**Średni próg słyszenia [dB HL]**	**TCT_{50} [samogłosek/s]**	**WER [%] $ROS^{PP}_{średnie}$**	**WER [%] $ROS^{PP}_{szybkie}$**
1	83	56,87	3,22	22	28
2	87	68,75	3,08	34	28
3	85	63,75	3,50	48	54
4	91	70	2,93	56	78
5(AS)	89	73,75	3,73	22	66
6	86	55,62	3,23	8	54
wartość średnia	86,83	64,79	3,28	31,67	51,33
odchylenie standardowe	2,86	7,36	0,29	17,95	20,15

Tab. 4.10 Wyniki badań słuchu wykonane dla osób starszych z normalną rozdzielczością czasową słuchu ($TCT_{50} \geq 3{,}99$ samogłosek/s)

L.p.	Wiek	Średni próg słyszenia [dB HL]	TCT_{50} [samogłosek/s]	WER [%] $ROS^{PP}_{średnie}$	WER [%] $ROS^{PP}_{szybkie}$
1 (AS)	86	58,75	4,25	10	14
2 (AS)	77	38,75	5,39	10	20
3	82	66,25	4,17	8	28
4	74	30,62	6,52	8	18
5	83	23,12	5,02	10	22
6	76	40	4,23	8	10
7	89	45,62	6,38	6	12
8	76	28,75	7,26	0	0
9	81	42,5	4,41	4	6
10	64	22,5	8,12	0	0
11	64	18,12	6,38	4	6
wartość średnia	77,45	37,73	5,65	6,18	12,36
odchylenie standardowe	8,05	15,19	1,37	3,73	9,07

Relacja pomiędzy wartością TCT_{50} a współczynnikiem WER dla mowy niezmodyfikowanej wypowiadanej w tempie szybkim

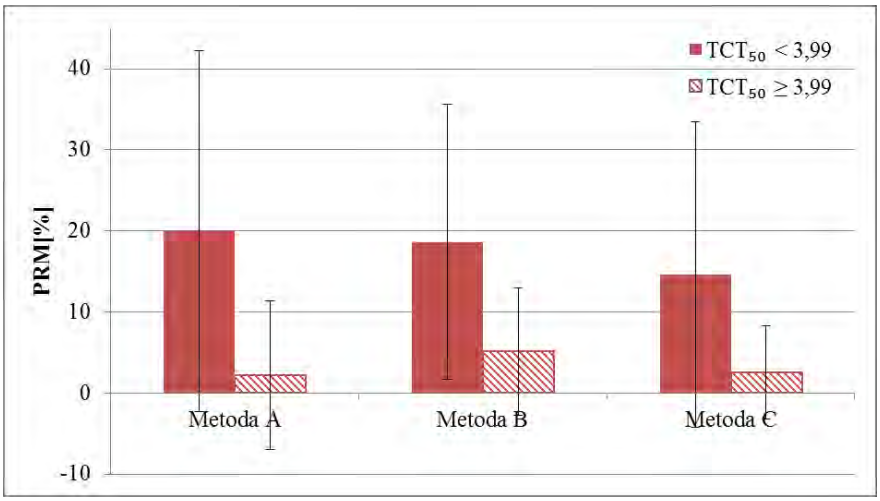
Na podstawie uzyskanych wyników zbadano, czy istnieje korelacja między wyznaczoną wartością progu TCT_{50}, a rozumieniem mowy wypowiadanej w tempie szybkim. Na rys. 4.7 umieszczono relację wartości TCT_{50} i średniej liczby błędów wyrażoną przez WER. Pełnymi kwadratami oznaczono osoby starsze z pogorszoną rozdzielczością czasową słuchu, a trójkątami osoby starsze z normalną rozdzielczością czasową słuchu. Linia ciągła reprezentuje trend dla wszystkich osób. Obliczone wartości współczynników korelacji Pearsona wyniosły –0,71 (dla wszystkich osób biorących udział w teście), 0,12 (dla osób z pogorszoną rozdzielczością czasową słuchu), oraz –0,58 (dla osób z normalną rozdzielczością czasową słuchu). Na podstawie uzyskanych wartości współczynnika korelacji można wnioskować, iż nie istnieje liniowa relacja między wartością progu TCT_{50} i liczbą błędów wyrażoną jako WER dla grupy osób starszych z pogorszoną rozdzielczością czasową słuchu. Dlatego też na podstawie wartości progu TCT_{50} nie można ocenić jakie problemy w rozumieniu mowy będą miały osoby z tej grupy słuchaczy. U osób starszych z normalną rozdzielczością czasową słuchu widoczna jest ujemna korelację pomiędzy parametrami. Można więc założyć, iż w tej grupie osób słuchacze z niższym progiem TCT_{50} będą mieli większe trudności w rozumieniu szybko wypowiadanej mowy niż słuchacze mający wyższy próg TCT_{50}.



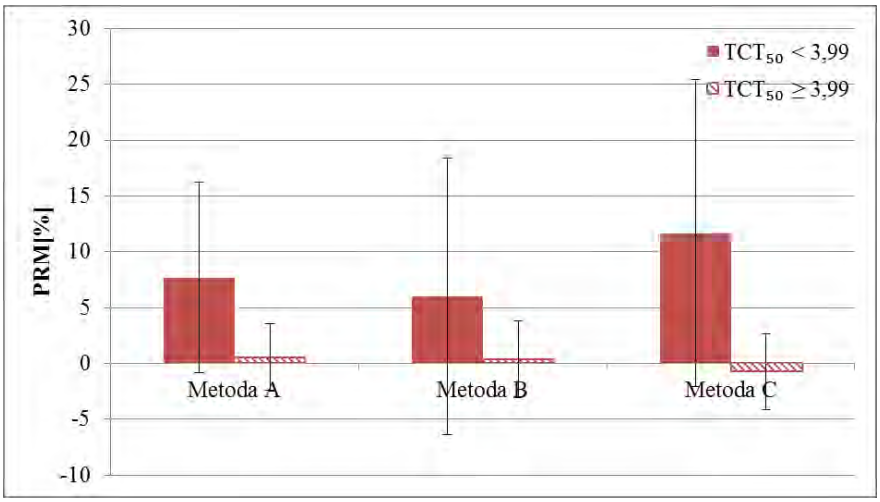
Rys. 4.7 Relacja pomiędzy progiem TCT_{50} a wartością WER wyznaczona dla osób starszych.

Analiza wpływu modyfikacji tempa na rozumienie mowy

Wpływ opracowanych metod modyfikacji sygnału na rozumienie mowy przez osoby starsze, podobnie jak w przypadku grupy dzieci głuchych, został oceniony niezależnie dla grupy słuchaczy z pogorszoną rozdzielczością czasową oraz z normalną. Analizę przeprowadzono z uwzględnieniem dwóch różnych temp mowy oryginalnej $ROS^{P}_{\text{średnie}} = 4{,}88$ samogłosek/s i $ROS^{P}_{\text{szybkie}} = 6{,}48$ samogłosek/s (tab. 4.3). Na rys. 4.8 i 4.9 przedstawiono średnie wartości poprawy rozumienia mowy uzyskane przez osoby starsze dla mowy wypowiadanej w tempie szybkim i średnim. Największą różnice w stopniu rozumienia mowy można zaobserwować dla grupy osób z progiem $TCT_{50} < 3{,}99$ samogłosek/s (obniżona rozdzielczość czasowa słuchu). Dla tej grupy słuchaczy tendencja ta jest widoczna zarówno dla mowy wypowiadanej w szybkim (PRM w przedziale od 14,66% do 20%) jak i w średnim tempie (PRM od 6% do 11,66%). Warty podkreślenia jest fakt, iż poprawa jest większy dla mowy wypowiadanej w szybkim tempie. W grupie osób starszych z normalną rozdzielczością czasową słuchu ($TCT_{50} \geq 3{,}99$ samogłosek/s), niewielką poprawę w rozumieniu mowy można zauważyć jedynie dla mowy wypowiadanej w szybkim tempie (PRM od 2,18% do 5,09%). Jednocześnie dla mowy wypowiadanej w tempie średnim znikomą poprawę da się zaobserwować dla metod A i B (PRM odpowiednio 0,54% i 0,36%). W metodzie C i mowie wypowiadanej w średnim tempie widać niewielkie pogorszenie stopnia rozumienia mowy (PRM = –0,72%).



Rys. 4.8 Poprawa rozumienia mowy u osób starszych uzyskana dla mowy wypowiadanej w tempie szybkim ($ROS^{P}_{szybkie}$ = 6,48 [samogłosek/s])



Rys. 4.9 Poprawa rozumienia mowy u osób starszych uzyskana dla mowy wypowiadanej w tempie średnim ($ROS^{P}_{średnie}$ = 4,88 [samogłosek/s])

Analiza wpływu modyfikacji tempa na rozumienie mowy wypowiadanej w tempie szybkim

W celu statystycznej weryfikacji istotności wyników TRMS przeprowadzono odpowiednią analizę. Jako że w przypadku mowy wypowiadanej w tempie` szybkim w obrębie obu grup słuchaczy ($TCT_{50} < 3,99$ i $TCT_{50} \geq 3,99$) wartości błędu WER, dla mowy niezmodyfikowanej jak i dla mowy spowolnionej z wykorzenianiem metod A–C, miały rozkład normalny, a założenie o sferyczności danych[13] zostało spełnione, możliwe było zastosowanie testu RM ANOVA.

Test RM ANOVA wykonano osobno dla grupy osób starszych z pogorszoną rozdzielczością czasową słuchu i dla grupy z normalną rozdzielczością czasową słuchu.

[13] Szczegółowe wynik testów Shapiro-Wilka i Mauchly'ego umieszczono w załączniku nr 3.

Hipotezą zerową testu było założenie, iż wartości średnie WER uzyskane dla mowy niezmodyfikowanej oraz mowy zmodyfikowanej z wykorzystaniem metod A–C są równe. W pierwszej grupie słuchaczy wartość statystyki wyniosła $F(3,15) = 4,36$ przy poziomie istotności $p = 0,021$. Jako że wartość istotności jest mniejsza od założonej wartości p_i, należy odrzucić hipotezę zerową i przyjąć hipotezę alternatywną mówiącą, iż przynajmniej w przypadku jednej z metod modyfikacji średnia wartość WER różni się od wartości WER uzyskanych dla pozostałych metod modyfikacji. W celu zbadania, która z metod powoduje zmianę średniej wartości WER przeprowadzono test *post hoc* z wykorzystaniem metody LSD. W tab. 4.11 zamieszczono wyniki porównania par metodą LSD. Jedynie w przypadku pary numer 2 wartość istotności jest mniejsza od założonego poziomu p_i. Oznacza to, iż dla tej pary należy odrzucić hipotezę zerową i przyjąć hipotezę alternatywną mówiącą o tym, że wartości średnie WER w obrębie pary różnią się między sobą w sposób istotny statystycznie. W związku z tym należy odrzucić hipotezę zerowa testu RM ANOVA. Powyższa analiza dowodzi tego, iż u osób starszych mających pogorszoną rozdzielczość czasową słuchu, spowalnianie mowy z wykorzystaniem metody B powoduje uzyskanie istotnego statystycznie wzrostu rozumienia mowy (18,67%) wypowiadanej w szybkim tempie.

Tab. 4.11 Wyniki parametrycznego testu *post hoc* uzyskane dla grupy osób starszych z pogorszoną rozdzielczością czasową słuchu.

L.p.	**Para**	**Wartość statystyki testu (t)**	**Ilość stopni swobody**	**Istotność (p)**
1	brak modyfikacji – metoda A	–2,20	5	0,079
2	brak modyfikacji – metoda B	–2,68	5	**0,043**
3	brak modyfikacji – metoda C	–1,91	5	0,115
4	metoda A – metoda B	0,29	5	0,780
5	metoda A – metoda C	1,66	5	0,158
6	metoda B – metoda C	1,19	5	0,286

Dla grupy osób starszych z normalną rozdzielczością czasową słuchu w teście RM ANOVA, wartość statystyki wyniosła $F(3,30) = 1,25$ przy poziomie istotności $p = 0,30$. Jako że wartość p jest większa od założonego poziomu p_i, należy przyjąć hipotezę zerową mówiąca o równości średnich w obrębie wszystkich metod modyfikacji sygnału. Oznacza to, że dla grupy osób starszych z normalną rozdzielczością czasową słuchu żadna z opracowanych metod modyfikacji sygnału nie powoduje istotnych statystycznie zmian w stopniu rozumienia mowy wypowiadanej w szybkim tempie.

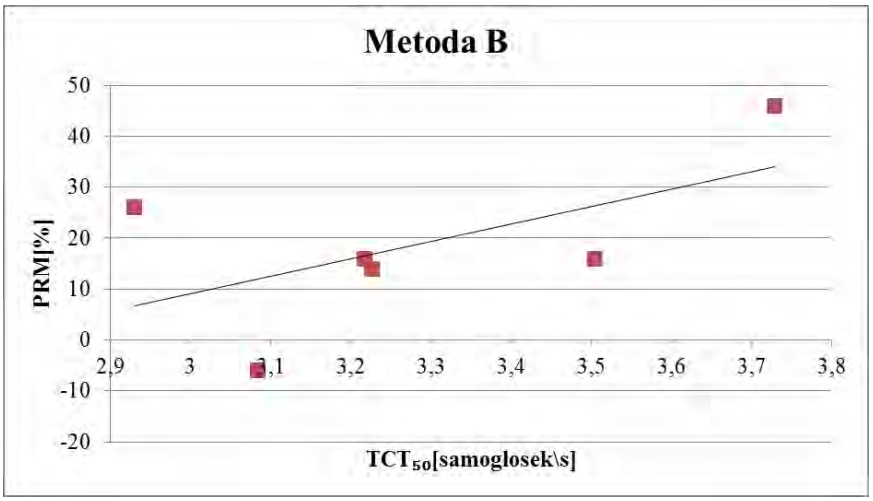
Analiza wpływu modyfikacji tempa na rozumienie mowy wypowiadanej w tempie średnim

W dalszej części dokonano analizy wyników TRMS uzyskanych dla mowy wypowiadanej w tempie średnim. Ponieważ zarówno w obrębie grupy osób starszych z progiem $TCT_{50} < 3{,}99$, jak i w obrębie grupy osób starszych z progiem $TCT_{50} \geq 3{,}99$ dla wszystkich metod modyfikacji sygnału oraz dla mowy niezmodyfikowanej wartości WER miały rozkład normalny oraz w teście Mauchly'ego nie potwierdzono hipotezy zerowej[14], możliwe było wykonania testu RM ANOVA. Hipoteza zerowa testu była taka sama jak w przypadku mowy wypowiadanej w tempie szybkim. Dla grupy osób starszych z pogorszoną rozdzielczością czasową słuchu wartość statystyki wyniosła $F(3,15) = 2{,}62$ przy poziomie istotności $p = 0{,}089$. Wartość p jest mniejsza od założonej wartości p_i, w związku z tym hipoteza zerowa została potwierdzona. Prowadzi to do wniosku, iż różnice w stopniu rozumienia mowy niezmodyfikowanej oraz spowolnionej z wykorzystaniem opracowanych metod, dla mowy wypowiadanej w tempie średnim, nie są istotne statystycznie. Dla grupy osób z normalną rozdzielczością czasową słuchu uzyskano następujące wartości testu: $F(3, 30) = 0{,}48$ przy $p = 0{,}69$. Wartość p jest większa od p_i dlatego nie ma podstaw do odrzucenia hipotezy zerowej testu RM ANOVA. Stąd wniosek, iż modyfikacja mowy z wykorzystaniem opracowanych metod nie ma statystycznie istotnego wpływu na stopień rozumienia mowy wypowiadanej w tempie średnim.

Relacja pomiędzy wartością TCT_{50} a współczynnikiem PRM dla mowy zmodyfikowane wypowiadanej w tempie szybkim

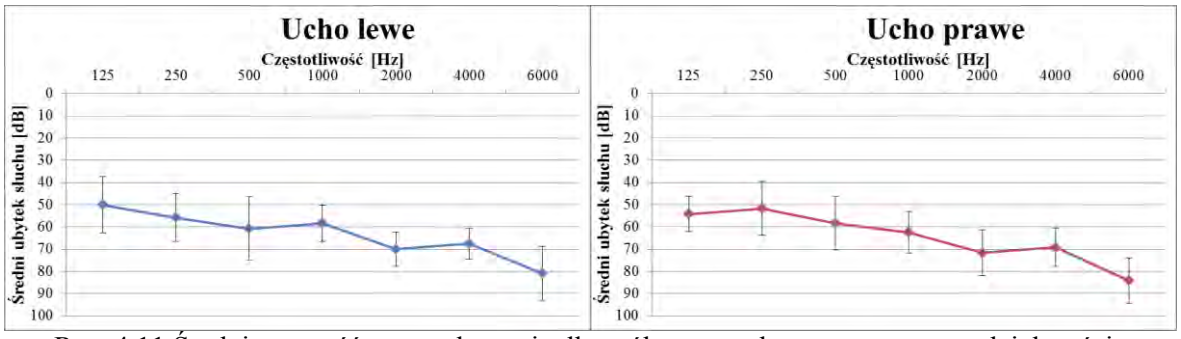
Powyższa analiza wykazała, iż jedynie metoda B, dla mowy wypowiadanej w tempie szybkim dla grupy osób starszych z pogorszoną rozdzielczością czasową słuchu, powoduje istotny statystycznie wzrost rozumienia wypowiedzi. W zawiązku z tym, dla tego przypadku zbadano, czy istnieje liniowa korelacja pomiędzy PRM i wartością TCT_{50}. Na rys. 4.10 przestawiono zależność tych dwóch parametrów dla grupy osób starszych z pogorszoną rozdzielczością czasową słuchu. Obliczony współczynnik korelacji Pearsona wyniósł 0,58. Sugeruje to istnienie zależności pomiędzy wartościami. Dodatnia wartość współczynnika świadczy o tym, iż wzrost wartości progowej TCT_{50} powoduje wzrost współczynnika PRM.

[14] Szczegółowe wynik testów Shapiro-Wilka i Mauchly'ego umieszczono w załączniku nr 3.

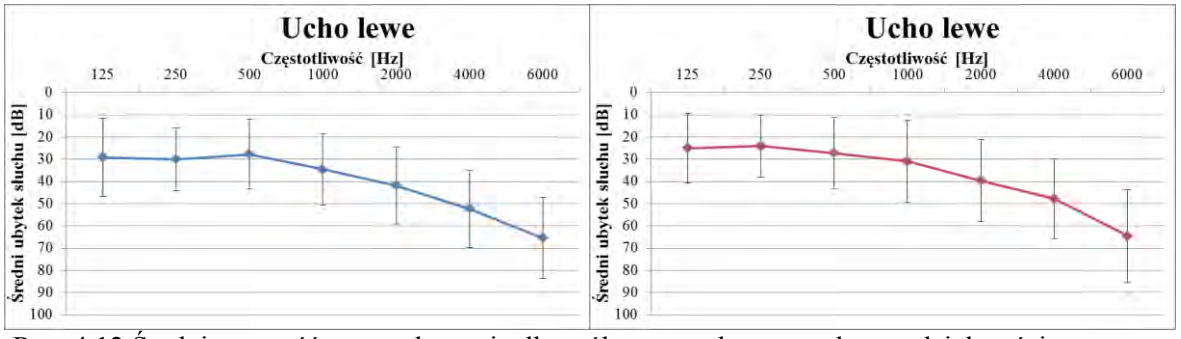


Rys. 4.10 Relacja pomiędzy progiem TCT_{50} a wartością PRM wyznaczona dla grupy osób starszych z progiem $TCT_{50} < 3{,}99$.

Wniosek ten wydaje się dziwny, ponieważ można by oczekiwać zależności odwrotnej (im niższa rozdzielczość czasowa słuchu tym wyższa procentowa poprawa rozumienia wypowiedzi). W celu wyjaśnienie tej wątpliwości wykonano dodatkową analizę wyników badań audiometrii tonalnej. Na rys. 4.11 i 4.12 przedstawiono średnie wartości progu słyszenia oraz wartości odchylenia standardowego dla osób starszych z pogorszoną i normalną rozdzielczością czasową słuchu.



Rys. 4.11 Średnie wartość progu słyszenia dla osób starszych z pogorszoną rozdzielczością czasową słuchu.



Rys. 4.12 Średnie wartość progu słyszenia dla osób starszych z normalną rozdzielczością czasową słuchu.

Wyraźnie widoczna jest około 25 dB różnica w progu słyszenia obu grup słuchaczy (potwierdzają to także wartości średnie progu słyszenia zamieszone w tab. 4.9 i 4.10).

Warte podkreślenia jest to, iż o ile dla grupy dzieci głuchych wszyscy słuchacze podczas badań korzystali z aparatu słuchowego bądź implantu ślimakowego, o tyle u osób starszych jedynie trzy osoby nosiły aparat słuchowy. Dlatego problem z niską rozdzielczością czasową słuchu nakładał się tu na problem z niskim progiem słyszenia, co mogło wprowadzać dodatkowy błąd w wynikach TRMS. W pewnym stopniu miała temu zapobiec możliwość dobierania, przez każdego z uczestników, poziomu amplitudy odsłuchiwanej wypowiedzi. Jednak nie pozwoliło to w pełni wyeliminować błędu związanego z brakiem pełnej korekcji obwodowego zaburzenia słuchu.

5 Badanie opracowanych metod

W niniejszym rozdziale przedstawiono wyniki przeprowadzonych badań skuteczności algorytmów detekcji mowy, detekcji samogłosek oraz estymacji tempa wypowiedzi. Dodatkowo w kolejnej części rozdziału zamieszczono wyniki eksperymentów mających na celu wyznaczenie zakresu zastosowania opracowanych metod. Zakres ten jest ograniczony przez wartości współczynnika skali, które zapewniają niewielkie przesunięcie pomiędzy sygnałem wejściowym, a sygnałem spowolnionym. Na koniec określono jakości i naturalności mowy spowolnionej za pomocą opracowanych metod modyfikacji czasu trwania sygnału. Mowa zmodyfikowana została poddana ocenie w serii testów subiektywnych.

Skuteczność wszystkich analizowanych algorytmów została zbadana z wykorzystaniem wypowiedzi zarejestrowanych w wytłumionym pomieszczeniu. W celu symulacji warunków pracy algorytmu, sygnał rejestrowano za pomocą mikrofonu komputerowego i karty dźwiękowej zintegrowanej z płytą główną komputera. W nagraniach wzięło udział 8 nieszkolonych mówców (1 kobieta/7 mężczyzn). Każdy miał za zadanie przeczytać 5 różnych wypowiedzi składających się z kilku zdań. Średnia liczba słów przypadających na wypowiedź wynosiła 35,2 (ich treść zamieszczono w załączniku nr 1). Duża liczba słów oraz ich różnorodność sprawiała, iż zadanie klasyfikacji postawione przed detektorami było trudniejsze w stosunku do typowo stosowanych wypowiedzi testowych (np. zarejestrowanych w bazie TIMIT), składających się ze zdań zawierających kilka słów. Każda wypowiedź była czytana w trzech różnych tempach: wolnym, średnim i szybkim. W efekcie uzyskano 120 nagrań o średnim czasie trwania 16,3 s. We wszystkich 120 nagraniach ręcznie oznaczono miejsca występowania mowy i dodatkowo w 30 nagraniach (15 odpowiadających mówcy żeńskiemu i 15 mówcy męskiemu) oznaczono obszar występowania samogłosek.

5.1 Skuteczność detekcji mowy

Ocena skuteczności użytego algorytmu detekcji mowy została wykonana z wykorzystaniem miar opisanych w rozdziale 2.2.3. Były to skuteczność detekcji mowy (HR1), skuteczność detekcji ciszy (HR0) oraz procentowa liczba błędów klasyfikacji (FAR). Wartości te zostały wyznaczone dla mowy nie zawierającej szumu oraz dla mowy z dodanym szumem białym, *babble noise* oraz *volvo noise*. Szum dodano w takich

proporcjach by uzyskać siedem poziomów SNR w zakresie od 30 dB do 0 dB z krokiem co 5 dB. Całkowity czas trwania ciszy we wszystkich nagraniach wynosił 401 s, a czas trwania mowy 2639 s. W tab. 5.1 i 5.2 umieszczono wyniki klasyfikacji detektora mowy bez wygładzania decyzji. Detektor ten dla celów uproszczenia analizy wyników został nazwany detektorem VAD_{czysty}. Przy braku zakłóceń szumowych uzyskał on wysoką skuteczność detekcji sygnału mowy 97,21% przy 84,05% skuteczności detekcji szumu. Wraz z obniżeniem wartości SNR, niezależnie od rodzaju szumu, wzrasta wartość HR0 a maleje HR1. Najsłabiej detektor radzi sobie przy szumie typu *babble noise* (średnia wartość HR1 wynosi 79,68%) a najlepiej przy szumie typu *volvo noise* (średnia wartość HR1 95%). Jest to związane z charakterystyką obu zakłóceń. Szum typu *babble noise* zawiera wiele składowych widma amplitudowego w paśmie mowy, które powodują wzrost wartości zarówno mocy widmowa amplitudowego, jaki i MD co prowadzi do błędów detekcji. Przy szumie typu *volvo noise* problem ten nie występuję, ponieważ szum jest niskoczęstotliwościowy i znajduje się on w widmie poniżej pasma mowy. Niezależnie od rodzaju szumu, przy SNR $\geq$ 10 dB skuteczność detekcji mowy jest większa od 87%. Ta wartość SNR jest istotna, ponieważ dla rozumienia mowy przez osoby z zaburzeniami słuchu, SNR = 15 dB jest poziomem, przy którym zakłada się, iż nie występują znaczące utrudnienia w rozumieniu wypowiedzi [24]. Dlatego została ona przyjęta, jako wartość graniczna dla opracowanej metody modyfikacji sygnału mowy. Dla niższych wartości SNR zakłócenia związane z szumem powodują problemy w rozumieniu wypowiedzi i stosowanie opracowanej metody u osób z zaburzeniami słuchu nie przyniesie korzyści [123] [175]. Typową wartością SNR, przy której szum nie wpływa na percepcję mowy przez osoby z (C)APD jest wartość 20 dB. Dlatego ta wartość jest progiem komfortowego słuchania przez osoby z zaburzeniami słuchu.

Tab. 5.1 Skuteczność algorytmu VAD uzyskana dla różnych rodzajów szumu i różnych wartości SNR.

SNR	**Biały**			***Babble***			***Volvo***			**Liczba ramek/ czas trwania [s]**	
	HR0	**HR1**	**FAR**	**HR0**	**HR1**	**FAR**	**HR0**	**HR1**	**FAR**	**HR0**	**HR1**
0	95,76	49,78	43,73	92,17	37,61	54,69	90,02	88,10	11,62	17276 (401s)	113666 (2639s)
5	96,06	71,13	25,35	90,31	63,33	32,86	88,11	93,23	7,48		
10	95,73	84,00	14,34	87,88	81,49	17,60	86,31	95,61	5,69		
15	94,87	91,26	8,22	87,42	89,51	10,77	85,51	96,61	4,94		
20	91,56	94,90	5,56	87,47	93,56	7,29	84,38	97,03	4,74		
25	88,25	96,41	4,73	86,88	95,66	5,57	84,34	97,16	4,64		
30	85,93	96,89	4,65	86,31	96,58	4,85	83,99	97,23	4,62		
Średnia	92,59	83,48	15,23	88,35	79,68	19,09	86,09	95,00	6,25		
Bez szumu	84,05	97,21	4,63								

Tab. 5.2 Średnie wartości skuteczności algorytmu VAD uzyskane dla różnych wartości SNR.

	Średnia		
SNR	**HR0**	**HR1**	**FAR**
0	92,65	58,50	36,68
5	91,49	75,89	21,90
10	89,97	87,03	12,54
15	89,27	92,46	7,98
20	87,80	95,16	5,86
25	86,49	96,41	4,98
30	85,41	96,90	4,71
Średnia	89,01	86,05	13,52

W tab. 5.3 i 5.4 przedstawiono skuteczność detekcji mowy osiągniętą przez opracowaną metodę przy zastosowaniu wygładzania decyzji. Algorytm ten został nazwany detektorem $VAD_{wygładzony}$. W detekcji mowy w warunkach bezszumowych skuteczność detekcji mowy wyniosła 98,92% a skuteczność detekcji szumu 72,78%. W porównaniu z detektorem VAD_{czysty} widać wzrost wartości HR1 i spadek HR0. Wynik ten nie jest zaskoczeniem, gdyż wygładzanie decyzji miało na celu zwiększenie skuteczności detekcji mowy. Podobnie, jak dla detektora VAD_{czysty}, najniższą skuteczność detekcji mowy uzyskano dla szumu typu *babble noise*. Istotny jest jednak fakt, iż dla wartości SNR $\geq$ 10 dB skuteczność detekcji mowy jest większą bądź równa 92,50%, a dla SNR = 15 dB HR1 = 96,06%.

Tab. 5.3 Skuteczność algorytmu VAD z wygładzaniem decyzji uzyskana dla różnych rodzajów szumu i różnych wartości SNR.

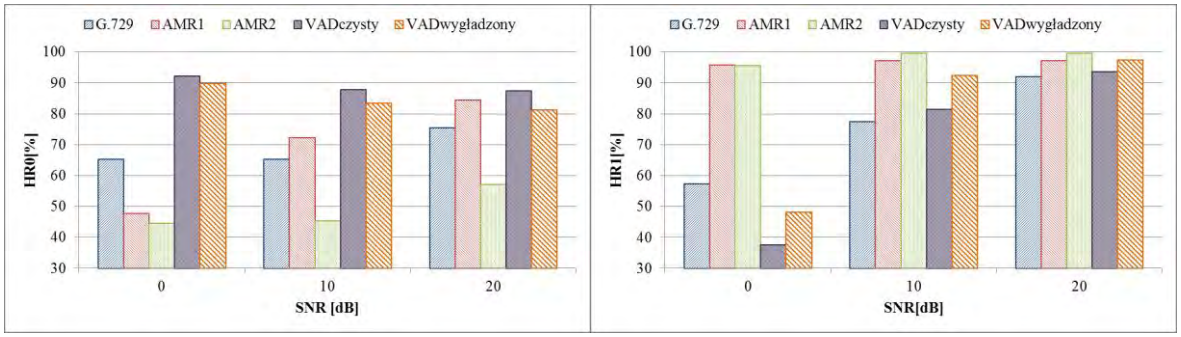
	Biały			***Babble***			***Volvo***			**Liczba ramek/ czas trwania [s]**	
SNR	**HR0**	**HR1**	**FAR**	**HR0**	**HR1**	**FAR**	**HR0**	**HR1**	**FAR**	**HR0**	**HR1**
0	90,80	71,53	25,75	89,85	48,11	46,01	73,84	96,49	6,69		
5	91,06	88,76	10,91	87,01	78,30	20,46	70,85	97,78	6,01		
10	90,80	94,38	6,12	83,52	92,50	8,75	68,95	98,45	5,70		
15	89,28	96,70	4,33	81,99	96,06	5,91	70,31	98,74	5,25	17276	113666
20	82,35	97,99	4,20	81,25	97,38	4,88	71,35	98,89	4,98	(401s)	(2639s)
25	77,99	98,65	4,25	78,90	98,18	4,53	73,08	98,89	4,73		
30	75,21	98,79	4,52	77,54	98,59	4,37	72,34	98,92	4,81		
Średnia	85,35	92,40	8,58	82,87	87,02	13,56	71,53	98,31	5,45		
Bez szumu	72,78	98,92	4,75								

Tab. 5.4 Średnie wartości skuteczność algorytmu VAD z wygładzaniem uzyskana dla różnych wartości SNR.

	Średnia		
SNR	**HR0**	**HR1**	**FAR**
0	84,83	72,04	26,15
5	82,97	88,28	12,46
10	81,09	95,11	6,86
15	80,53	97,17	5,17
20	78,32	98,09	4,69
25	76,66	98,57	4,51
30	75,03	98,77	4,57
Średnia	79,92	92,58	9,20

Skuteczność opracowanego algorytmu VAD porównano ze skutecznościami innych metod znanych z literatury. Na rys. 5.1 przedstawiono wartości HR0 i HR1 uzyskane dla opracowanych detektorów VAD_{czysty} i $VAD_{wygładzony}$ oraz dla opisanych w normach algorytmów VAD takich jak detektora mowy G.729, AMR1 i AMR2. Prezentowane wyniki uzyskane zostały dla sygnału mowy zaszumionej szumem typu *babble noise*. W przykładzie wykorzystano właśnie ten szum, ponieważ jest on typowym zakłóceniem występującym m.in. podczas zajęć w szkole, podczas rozmowy w biurze lub w centrum handlowym. Jak można zauważyć, niezależnie od poziomu SNR najwyższą skuteczność detekcji mowy osiąga algorytm AMR2 (powyżej 90%). Jednak skuteczność detekcji szumu w przypadku tego algorytmu jest niska (niezależnie od SNR znacznie niższa niż 60%). Podobną zależność można zaobserwować także dla algorytmów G.729 i AMR1 z tą różnicą, iż skuteczność HR0 jest tu znacznie wyższa niż dla algorytmu AMR2. Natomiast

opracowane detektory VAD_{czysty} i $VAD_{wygładzony}$ uzyskują wysoką skuteczność detekcji HR0 (zawsze powyżej 80%) i w warunkach $SNR \geq 10dB$ skuteczność detekcji HR1 powyżej 80%. Detektor $VAD_{wygładzony}$ znacząco przewyższa detektor VAD_{czysty} pod względem skuteczność detekcji mowy przy niewiele słabszej skuteczności detekcji szumu. Zastosowanie w opracowanej metodzie modyfikacji sygnału mowy detektorów o niskiej skuteczność detekcji HR0 powodowałoby wprowadzenie dodatkowego opóźnienia, o którym pisano w rozdziale 3.2. Jednak w uzasadnionych sytuacjach (np. praca w warunkach $SNR < 10dB$) użycie tych detektorów przyniesie zysk w postaci znacząco wyższej skuteczność detekcji mowy.



Rys. 5.1 Porównanie skuteczności detekcji szumu i mowy uzyskanych dla różnych algorytmów w warunkach zakłócenia szumem typu *babble noise*.

Jak pokazały badania opracowany detektor sygnału mowy (detektor $VAD_{wygładzony}$) przy $SNR \geq 10dB$ zapewnia wysoką skuteczność detekcji sygnału mowy zaszumionej różnego rodzaju szumem przy jednocześnie wysokiej skuteczności detekcji fragmentów szumowych.

5.2 Skuteczność detekcji samogłosek

Opracowany algorytm VRD został oceniony przy użyciu dwóch miar. Pierwszą z nich była skuteczność detekcji ramek zawierających obszar samogłosek. Została ona obliczona poprzez porównanie wyników detekcji z ręczną indeksacją nagrań. Wszystkie ramki, które nie zostały zaklasyfikowane jako samogłoski, oznaczano jako spółgłoski. Założenie, iż ramki niezawierające samogłosek zawierają spółgłoski było poprawne, ponieważ algorytm VRD analizuje jedynie ramki sygnału, które zostały zaklasyfikowane przez algorytm VAD jako mowa. Drugą miarą oceny algorytm VRD był współczynnik VER. Jest on często stosowany w celu oceny algorytmów detekcji samogłosek (podrozdział 2.4). Podczas obliczania wartości VER, jako liczbę samogłosek uznawano liczbę wykrytych VR. W celu zbadania, czy zaproponowany parametr PR pozwala uzyskać wysoką skuteczność detekcji obszaru samogłosek, wyniki uzyskane dla opracowanego algorytmu porównano ze

skutecznością detekcji uzyskaną dla algorytmów bazujących na tym samym algorytmie detekcji, ale wykorzystujących inne parametry: PVD i REC. W dalszej części rozprawy symbolem PR_{det} oznaczany będzie algorytm VRD bazujący na parametrze PR, a symbolami PVD_{det} i REC_{det} algorytmy bazujące na analizie parametrów PVD i REC.

W tab. 5.5 przedstawiono skuteczność detekcji ocenianych algorytmów VRD. Najwyższą skuteczność detekcji obszaru samogłosek uzyskał algorytm PVD_{det} (75,80%), a detekcji obszaru spółgłosek algorytm REC_{det} (87,18%). Opracowany algorytm PR_{det} osiągnął nieznacznie niższą skuteczność detekcji obszaru samogłosek (75,7%) i znacznie wyższą od algorytmu PVD_{det} skuteczność detekcji obszaru spółgłosek (83,03%). Skuteczność detekcji spółgłosek uzyskana przez algorytm PR_{det} jest niższa od tej osiągniętej przez algorytm REC_{det}, jednak algorytm PR_{det} pozwala na osiągnięcie o wiele wyższą skuteczność detekcji obszaru samogłosek, co jest niezmiernie istotne podczas nierównomiernej modyfikacji czasu trwania sygnału. Dlatego można stwierdzić, iż zastosowanie parametru PR w opracowanym algorytmie VRD przyniosło oczekiwany efekt w postaci wysokiej skuteczności detekcji obu rodzaju głosek.

Tab. 5.5 Skuteczność detekcji samogłosek uzyskana przez algorytmy PR_{det}, PVD_{det}, i REC_{det}.

	PR_{det}		**PVD_{det}**		**REC_{det}**	
	samogłoski	**spółgłoski**	**samogłoski**	**spółgłoski**	**samogłoski**	**spółgłoski**
Średnia[%]	75,7	83,03	**75,80**	73,89	70,5	**87,18**
Liczba ramek	6183	12523	6183	12523	6183	12523
Czas trwania [s]	143,5	290,8	143,5	290,8	143,5	290,8

Dodatkowo wyniki te zostały potwierdzone przez wyniki VER przedstawione w tab. 5.6. Najmniej skuteczny pod względem poprawnie wykrytych samogłosek okazał się algorytm PVD_{det}. Algorytmy PR_{det} i REC_{det} uzyskały zbliżoną częstość błędów, jednak algorytm PR_{det} osiągną znacznie niższą częstość niewykrytych samogłosek (6,93%) niż algorytm PVD_{det} (10,10%).

Tab. 5.6 Błąd detekcji samogłosek uzyskany dla algorytmów PR_{det}, PVD_{det}, i REC_{det}.

	PR_{det}	**PVD_{det}**	**REC_{det}**
n^b_{sam}	14,21	22,80	**11,71**
$n^b_{spół}$	**6,93**	8,34	10,10
VER	**21,14**	31,15	21,82
n^o_{sam}	2048		

Skuteczność opracowanego algorytmu VRD może zostać porównana ze skutecznością metod znanych z literatury jedynie na poziomie błędów wyrażonych za pomocą miary VER. Autorowi rozprawy nie są znane opracowania, w których przedstawiona byłaby skuteczność detekcji obszarów samogłosek oraz rezultaty osiągnięte przez algorytmy detekcji analizujące wypowiedzi w języku Polskim. W tab. 2.6 przedstawiono VER uzyskane dla czterech algorytmów proponowanych w literaturze. Wyniki te przedstawiają częstość błędów detekcji (VER) wyznaczoną dla różnych języków (podrozdział 2.4). Należy tu podkreślić, iż metody, których skuteczność przedstawiono w tab. 2.6, miały za zadanie wykrycie różnych zdarzeń akustycznych związanych z samogłoskami (VOP, VRD, VD) i w przeciwieństwie do opracowanego algorytmu PR_{det}, nie operowały w czasie rzeczywistym. Dlatego najbardziej odpowiednim wydaje się porównanie skuteczności algorytmu PR_{det} z metodami detekcji VRD. I tak algorytm PR_{det} uzyskał wartość błędu mniejszą niż średnia wartość VER osiągnięta przez algorytm proponowany przez Pellegrino i Andre-Obrecht [137] (22,9%). Dla języków japońskiego, hiszpańskiego i francuskiego, algorytm opracowanych przez Pellegrino *et al.* osiągną niższą wartość VER niż algorytm PR_{det} dla języka polskiego. Natomiast w przypadku koreańskiego i wietnamskiego algorytm Pellegrino i Andre-Obrecht. uzyskał wyższą liczbę błędów. W swoich badaniach Ringeval i Chetouani [154] wykorzystali metodę Pellegrino i Andre-Obrecht oraz przedstawili wyniki detekcji VRD dla języka angielskiego, baskijskiego i niemieckiego. Jedynie dla języka angielskiego i klasyfikacji wykonanej z wykorzystaniem bazy TIMIT, osiągnięta przez nich liczba błędów detekcji była niższa niż dla algorytmu PR_{det} i języka polskiego. Wyniki te pokazują, iż opracowana metoda VRD pozwala na skuteczną detekcję obszaru samogłosek w czasie rzeczywistym.

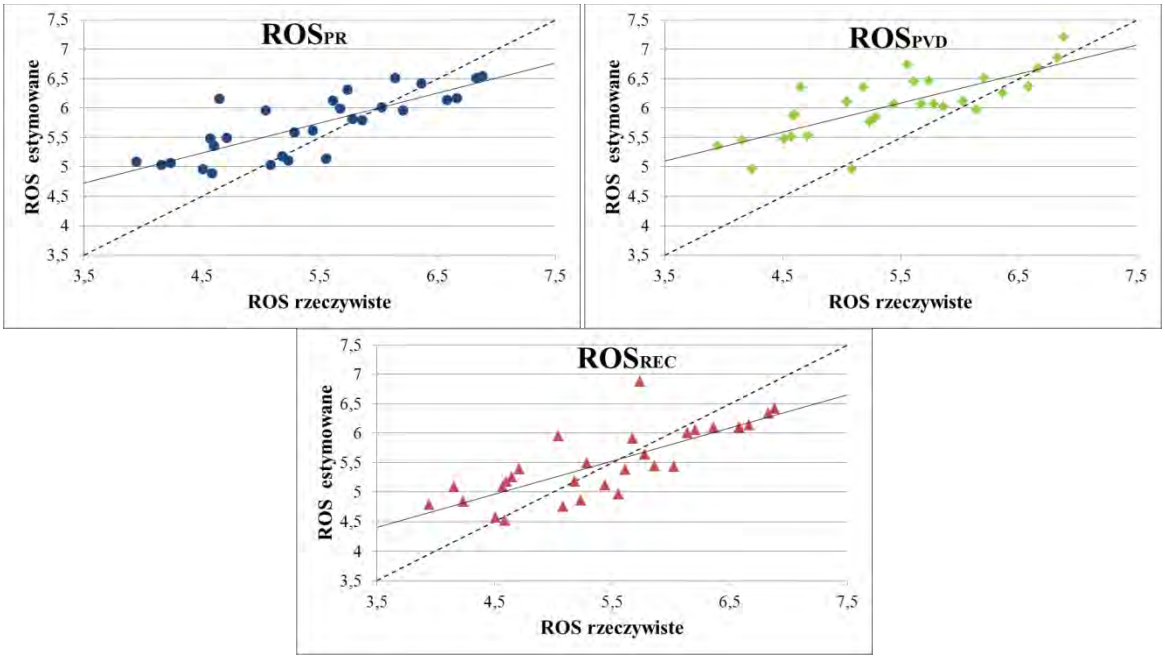
5.3 Skuteczność estymacji tempa wypowiedzi

Skuteczność opracowanego algorytm estymacji tempa wypowiedzi została oceniona osobno dla dwóch wariantów estymacji: lokalnej i globalnej. Dokładność lokalnej estymacji tempa wypowiedzi oceniono na podstawie analizy 30 nagrań, w których miejsca występowania samogłosek oznaczono ręcznie. Globalną estymację oceniono korzystając ze wszystkich 120 nagrań, a rzeczywiste wartości tempa wypowiedzi obliczono na podstawie transkrypcji nagrań. W badaniach użyto trzech jednakowych algorytmów estymacji tempa wypowiedzi opartych na algorytmach VRD przebadanych w rozdziale 5.2. W celu uproszczenia analizy wyników algorytmy estymacji bazujące na różnych parametrach oznaczono w następujący sposób: ROS_{PR}, ROS_{PVD}, ROS_{REC}.

5.3.1. Estymacja lokalnej wartości tempa wypowiedzi

Jak wspomniano w rozdziale 2.3, skuteczność algorytmów estymacji tempa mowy oceniona może zostać poprzez wyznaczenie współczynnika korelacji Pearsona. Współczynnik ten obliczany jest pomiędzy rzeczywistą wartością ROS, a wartością estymowaną przez algorytm. W ramach przedstawionych tu badań rzeczywista lokalna wartość ROS wyznaczona została na podstawie analizy położenia samogłosek oznaczonych podczas procesu ręcznej indeksacji. Dodatkowo, dla każdego nagrania wyznaczono wartość średnią estymowanego oraz rzeczywistego tempa wypowiedzi. Najwyższą wartość współczynnika korelacji uzyskano dla algorytmu ROS_{PR} (0,79). Algorytmy ROS_{PVD} i ROS_{REC} uzyskały nieznacznie niższą wartości współczynnika korelacji (0,77).

Rys. 5.2 ilustruje relacje występujące pomiędzy rzeczywistymi i estymowanymi lokalnymi wartościami ROS. Linia ciągła przedstawia prostą regresji liniowej a linia przerywana jest prostą o nachyleniu jednostkowym. Można zauważyć, iż wszystkie analizowane algorytmy mają tendencję do przeszacowywania wartości tempa wypowiedzi w przypadku mowy, której tempo jest mniejsze niż 5,5 samogłosek/s. Algorytm ROS_{PVD} przedstawia większą skłonność do zawyżania wartość ROS, niż dwa pozostałe algorytmy. Dla mowy wypowiadanej w tempie wyższym niż 5,5 samogłosek/s algorytmy ROS_{PR} i ROS_{REC} niedoszacowująca wartości tempa wypowiedzi.



Rys. 5.2 Relacja pomiędzy rzeczywistą i estymowaną średnią lokalną wartością ROS uzyskana dla algorytmów ROS_{PR}, ROS_{PVD} i ROS_{REC}.

Uzyskane wyniki porównano z wynikami przedstawionymi w literaturze (podrozdział 2.3). Należy tu podkreślić trzy istotne kwestie: 1) wyniki przedstawione w tab. 2.5 zostały uzyskane dla algorytmów estymujących globalną wartość ROS dla całego zdania lub wypowiedzi; 2) algorytmy opisane w literaturze nie dokonywały estymacji tempa wypowiedzi w czasie rzeczywistym; 3) podobnie jak w przypadku algorytmów VRD autorowi nie są znane badania przedstawiające wyniki skuteczności estymacji ROS dla mowy polskiej. Obliczona dla algorytmu ROS_{PR} wartość współczynnik korelacji jest większa lub równa , wartościom współczynnika uzyskanym dla języka niemieckiego (Pellegrino *et al.* [139]; Pfau i Ruske [141]), angielskiego (Morgan i Fosler-Lussier [106]; Wang i Narayanan [184] i duńskiego (De Jong [29]). Jedynie dla algorytmu opracowanego przez Pellegrino *et al.* [139] (metoda bazująca na parametrze REC) algorytm ROS_{PR} uzyskał mniejszą wartość współczynnika korelacji.

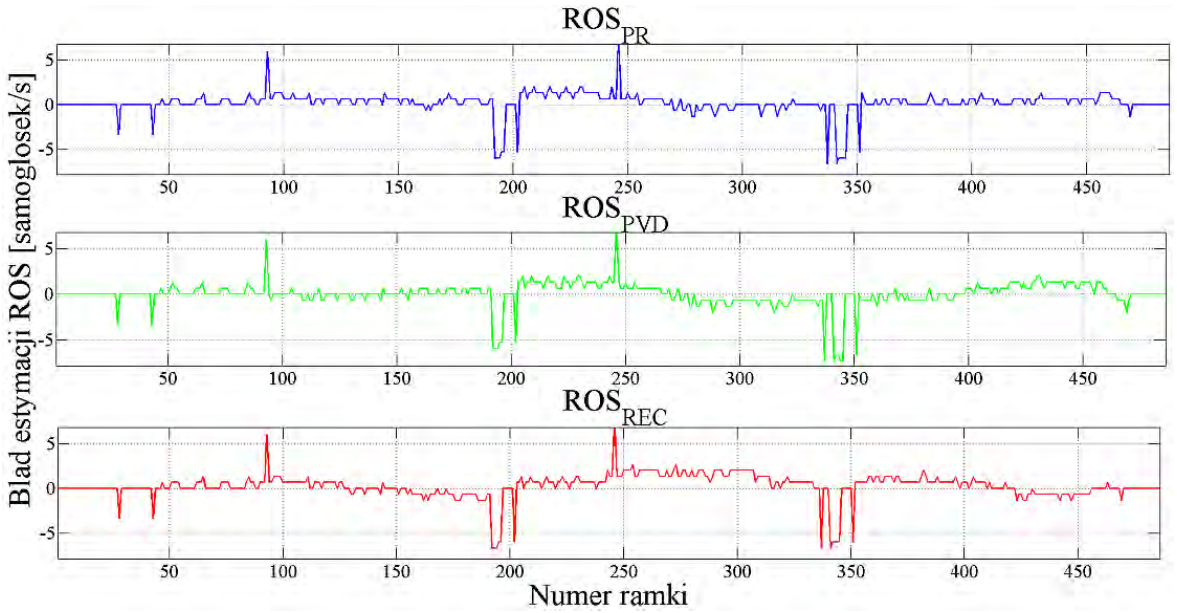
W celu określenia, który z ocenianych algorytmów estymacji ROS pozwala na uzyskanie najdokładniejszej estymacji lokalnej wartości ROS, dla każdego nagrania obliczono wartości błędu średnio kwadratowego (ang. *Mean Squere Error* – MSE). W tab. 5.7 zamieszczono średnie wartości MSE obliczone dla algorytmów ROS_{PR}, ROS_{PVD} i ROS_{REC}. Algorytm ROS_{REC} uzyskał najniższą wartość błędu. Jak można zauważyć, że wartość MSE dla wszystkich algorytmów jest wysoka. Jest to związane z dwoma faktami: 1) błędna detekcja ciszy jako mowy powoduje pojawienie się błędów estymacji o dużej wartości, ponieważ w miejscu ciszy oczekiwana wartość ROS wynosi 0 samogłosek/s, a algorytm przypisuje temu miejscu pewną wartość, 2) błędy algorytmu VRD powodują niedoszacowanie lub przeszacowanie lokalnej wartości ROS. Obydwa rodzaje pomyłek prowadzą do kumulacji błędu estymacji w kolejnych ramkach. Pomimo, iż oba algorytmy detekcji(VAD i VRD) osobno zapewniają wysoką skuteczność detekcji zdarzeń, w procesie estymacji ROS błędy przez niepowodowane wpływają na znaczące przeszacowanie/niedoszacowanie chwilowych wartości tempa wypowiedzi.

Tab. 5.7 Średnie wartości MSE wyznaczone dla algorytmów ROS_{PR}, ROS_{PVD} i ROS_{REC}.

	ROS_{PR}	**ROS_{PVD}**	**ROS_{REC}**
MSE	2,08	2,63	**1,96**

W celu wizualizacji tego problemu na rys. 5.3 zamieszczono wartości błędu estymacji ROS uzyskane przez algorytmy ROS_{PR}, ROS_{PVD} i ROS_{REC}. Każdy wykres ilustruje wielkość błędu estymacji (wyrażoną w liczbie samogłosek/s) obliczoną dla kolejnych ramek analizy jako różnica pomiędzy chwilowymi wartościami ROS wyznaczonymi na

podstawie ręcznego oznaczenia nagrania, a wartościami ROS estymowanymi przez kolejne algorytmy. Na górnym wykresie umieszczono chwilowe wartości błędu estymacji ROS wyznaczone dla algorytmu ROS_{PR}, środkowy wykres przedstawia błąd estymacji algorytmu ROS_{PVD}, a dolny – algorytmu ROS_{REC}. Na wykresach widoczne są skoki wartości błędu estymacji, których wartość przekracza 5 samogłosek/s. Miejsca te korelują z fałszywymi alarmami pojawiającymi się podczas detekcji mowy. W pozostałych miejscach błędy estymacji oscylują w przedziale od –2 do 2 samogłosek/s.



Rys. 5.3 Błąd lokalnej estymacji wartości ROS występujący podczas wyznaczania tempa wypowiedzi przez algorytmy ROS_{PR}, ROS_{PVD} i ROS_{REC}.

W dalszej części tego rozdziału zbadano skuteczność estymacji chwilowej etykiety tempa mowy wykorzystywanej w metodzie B (podrozdział 3.4). Skuteczność estymacji klasy tempa obliczono osobno dla każdej ramki analizowanego sygnału. W tab. 5.8 przedstawiono osiągnięte rezultaty. Algorytm ROS_{PVD} uzyskał najwyższą skuteczność detekcji wolnego tempa wypowiedzi (83,22%), a algorytm ROS_{REC} najwyższą skuteczność detekcji szybkiego ROS (81,29%). Skuteczność detekcji obu klas tempa obliczona dla algorytm ROS_{PR} jest zbliżona do najlepszych wyników osiągniętych przez dwa pozostałe algorytmy (ROS_{REC} i ROS_{PVD}). Przedstawione rezultaty pokazują, iż pomimo wysokiej wartości MSE uzyskanej przez algorytm ROS_{PR}, skuteczność detekcji chwilowej wartości klasy ROS osiągana przez ten algorytm jest wysoka. W tym miejscu należy podkreślić, iż przedstawione skuteczności detekcji klasy ROS nie są obciążone błędem wynikającym z faktu, iż wartość progu ROS_{th} została wyznaczona na podstawie analizy tych samych nagrań, które użyto do wyznaczenia skuteczność opracowanego algorytmu estymacji. Jest tak, ponieważ etykiety ramek oznaczonych na podstawie ręcznej indeksacji nagrań, zostały

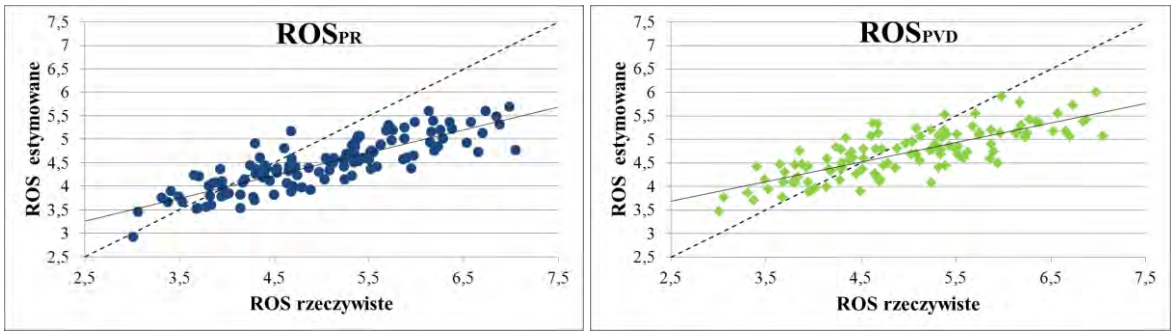
wyznaczone także na podstawie tego progu. Dlatego wartość ROS_{th} jest używana jedynie w celu określenia rzeczywistej klasy ROS i może zostać zdefiniowana w dowolny sposób.

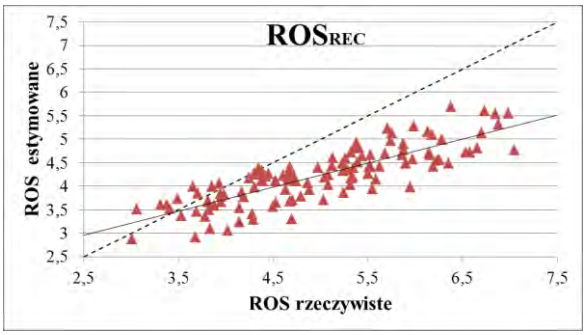
Tab. 5.8 Skuteczność rozpoznawania lokalnej klasy tempa wypowiedzi uzyskana przez algorytmy ROS_{PR}, ROS_{PVD} i ROS_{REC}.

	Skuteczność		**Liczba ramek**	
	wolna	**szybka**	**wolna**	**szybka**
ROS_{PR}	79,39	75,13	12873	6935
ROS_{PVD}	**83,22**	60,95		
ROS_{REC}	68,97	**81,29**		

5.3.2. Estymacja globalnej wartości tempa wypowiedzi

Skuteczność estymacji globalnej wartości tempa wypowiedzi została oceniona na podstawie analizy 120 nagrań. Dla każdego nagrania, rzeczywistą wartość ROS obliczono jako stosunek liczby samogłosek występujących w nagraniu (wartość wyznaczona na podstawie transkrypcji) do długość danego nagrania. Relację pomiędzy rzeczywistą i estymowaną wartością ROS przedstawiono na rys. 5.4. Podobnie jak w przypadku lokalnej estymacji tak i tu algorytm ROS_{PR} uzyskał najwyższą wartość współczynnika korelacji pomiędzy tymi wartościami (0,85). Algorytm ROS_{REC} osiągną niższą wartość współczynnika korelacji (0,83), a algorytm ROS_{PVD} najniższą wartość korelacji (0,79). Wyniki te mogą zostać porównane z wynikami przedstawionymi w tab. 2.5 (pp. 2.3.4). Jedynie algorytm zaproponowany przez Pellegrino *et al.* [139] dla języków: Chińskiego (0,88), Japońskiego (0,88) i Hinduskiego (0,91) osiągną wartość korelacji wyższą niż opracowany algorytm ROS_{PR}. W pozostałych przypadkach wartości współczynnika korelacji sprawozdawane w literaturze są niższe niż wartości współczynnika korelacji obliczone dla estymacji ROS w przypadku mowy wypowiadanej w języku Polskim.





Rys. 5.4 Relacja pomiędzy rzeczywistą i estymowaną globalną wartością ROS uzyskana dla algorytmów ROS_{PR}, ROS_{PVD} i ROS_{REC}.

5.4 Określenie złożoności obliczeniowej opracowanych metod modyfikacji sygnału

Opracowane metody modyfikacji czasu trwania mowy powinny umożliwiać dokonanie przetwarzania sygnału wejściowego w czasie rzeczywistym. W celu określenia ich złożoności obliczeniowej dla każdego z opracowanych algorytmów detekcji jak również dla metody B i C wyznaczono wartość współczynnika RTF(ang. *real-time factor*), określającego możliwości operowanie algorytmu w czasie rzeczywistym. Wartość tego współczynnika jest bezpośrednio zależna od platformy sprzętowej oraz od implementacji. RTF obliczana jest zgodnie ze wzorem [102]:

$$\mathrm{RTF} = \frac{t_p}{t_{\mathrm{zdarzenia}}} \tag{5.1}$$

gdzie t_p oznacza czas potrzebny do przetworzenia określonego fragmentu sygnału, a $t_{\mathrm{zdarzenia}}$ – czas trwania tego fragmentu. Jeżeli wartość RTF jest mniejsza od 1, to oceniany algorytm może przetwarzać sygnał w czasie rzeczywistym na danej platformie sprzętowej. Eksperymenty związane z oceną wydajności algorytmu zostały przeprowadzone na komputerze przenośnym wyposażonym w procesor Intel Core i7-263QM CPU 2,00 GHz, 8GB RAM i pracującym pod 64 bitowym systemem operacyjnym Windows 7 Home Premium. Oceniono wydajność algorytmów zaimplementowanych w środowisku MATLAB.

W tab. 5.9 umieszczono wartości RTF osiągnięte przez poszczególne algorytmy detekcji oraz przez metody modyfikacji czasu trwania sygnału (B i C). Eksperyment przeprowadzono dla stałej wartości współczynnika skali równego 1,75 (metoda B) i stałej wartości komfortowego $ROS_o = 4{,}2$ samogłosek/s (metoda C). Wszystkie przebadane algorytmy detekcji jaki i zaproponowane metody TSM umożliwiają przetwarzanie sygnału w czasie rzeczywistym (wszystkie wartość RTF są mniejsze od 1). Uwzględniono tu trzy różne algorytmy VRD (PR_{det}, PVD_{det} i REC_{det}). Tak, jak można było oczekiwać, algorytm

PR_{det} jest najbardziej wymagającym obliczeniowo algorytmem VRD (RTF = 0,0359), a najmniej wymagającym obliczeniowo algorytm VRD jest algorytm PVD_{det}. Należy też zauważyć, iż algorytmy detekcji stanowią znacznie mniejsze wyzwanie pod względem obliczeniowym niż sama procedura TSM oparta na algorytmie SOLA. W zależności od metody (B i C) wartość RTF dla TSM wyniosła odpowiednio 0,101 i 0,0862. Różnice w wartościach współczynnika RTF pomiędzy metodami związane są z różnymi chwilowymi wartościami współczynnika skali wykorzystywanymi przez obie metody.

Tab. 5.9 Wartości współczynnika RTF wyznaczone dla algorytmów składowych opracowanych metod TSM.

		RTF		
VAD		0,0376		
ROS		0,006		
TSM	**metoda B**	0,101		
	metoda C	0,0862		
VRD		**PR**	**PVD**	**REC**
		0,0359	**0,019**	0,0264
metoda B		0,1805	**0,1636**	0,171
metoda C		0,1657	**0,1488**	0,1562

5.5 Analiza opóźnień wprowadzanych przez opracowaną metodę modyfikacji sygnału

Podczas badań ocenie poddano mowę zmodyfikowaną za pomocą metody B oraz mowę zmodyfikowaną z wykorzystaniem algorytmu zaproponowanego przez Nejime *et al.* [121](pp. 2.1.6). Nie oceniano tu metody C, ponieważ, jak to zostało pokazane w rozdziale 4 rozprawy, metoda ta pozwala na uzyskanie niewielkiej poprawy rozumienia mowy przez dzieci z pogorszoną rozdzielczością czasową słuchu, a u osób starszych nie wykryto istotnych statystycznie różnic w rozumieniu mowy oryginalnej i zmodyfikowanej za pomocą tej metody. Algorytm opracowany przez Nejime *et al.* [121] został uwzględniony w teście, aby określić relację pomiędzy jakością oraz naturalnością mowy zmodyfikowanej za pomocą opracowanej metody TSM, a tymi parametrami osiąganymi przez inny algorytm znany w literaturze. Wybrano metodę Nejime *et al.* [121], ponieważ jej przeznaczenie jest takie samo jak przeznaczeniem metody B (wspomaganie rozumienia mowy przez osoby z pogorszoną rozdzielczością czasową słuchu) i dodatkowo przetwarzanie sygnału odbywa się tu w czasie rzeczywistym. Należy także zauważyć, iż podobnymi założeniami odnośnie grupy docelowej kierowali się Nakamura *et al.* [118]. Dodatkowo obie metody modyfikacji czasu trwania sygnału mowy (opracowane przez Nejime *et al.* i Nakamura *et al.*) bazują na tym samym algorytmie TSM (TDHS)

wykonującym nierównomierną modyfikację struktury czasowej polegająca na spowalnianiu głosek dźwięcznych z wyższymi wartościami współczynnika skali niż głosek bezdźwięcznych. Dlatego uznano, iż możliwy jest wybór jednej z tych metod jako reprezentanta. Do celów badawczych algorytm opracowany przez Nejime *et al.* został zaimplementowany w środowisku MATLAB zgodnie z jego szczegółowym opisem dostępnym w referacie [121]. W dalszej części tego rozdziału metoda opracowana przez Nejime *et al.* będzie nazywana metodą R (referencyjną).

Opracowana metoda (B), została zaprojektowana w taki sposób by umożliwiała wykonanie procedury spowalniania sygnału rejestrowanego przez mikrofon i równoczesne jego odtwarzanie. W naturalny sposób powoduje to powstawanie opóźnienia pomiędzy sygnałem wejściowym (rejestrowanym przez mikrofon), a sygnałem wyjściowym (spowolnionym sygnałem odtwarzanym na słuchawkach). Dlatego celowe jest zbadanie relacji pomiędzy wartością współczynnika skali, a opóźnieniem pomiędzy sygnałem wejściowym i wyjściowym.

W tym celu korzystając z opracowanej metody B spowolniono 120 nagranych wypowiedzi. Przetwarzanie wykonano dla czterech wartości współczynnika skali (α_o = 1,25, 1,33, 1,5 i 1,75). Dla każdej wartości α_o wyznaczono średnią wartość *brutto* współczynnika skali ($\mu(\alpha_{brutto})$), gdzie α_{brutto} zdefiniowano jako stosunek czasu trwania nagrania zmodyfikowanego do czasu trwania nagrania wejściowego. Wyznaczono także średnią wartości *netto* współczynnika ($\mu(\alpha_{netto})$), gdzie α_{netto} oznacza stosunek czasu trwania sygnału mowy w nagraniu zmodyfikowanym do czasu trwania mowy w sygnale wejściowym. Nazewnictwo to jest pochodną sposobu definiowania ROS (podrozdział 2.3). Należy zauważyć, iż inaczej niż w estymacji ROS, wartość *netto* współczynnika skali jest zawsze większa bądź równa wartości *brutto* (a nie mniejsza bądź równa). W tab. 5.10 przedstawiono średnie wartości współczynników skali oraz średnie wartości opóźnienia ($\mu(t_{wej})$) (wyrażone w sekundach), wyznaczone dla mowy wypowiadanej w tempie wolnym, średnim i szybkim. W ostatniej kolumnie tabeli umieszczono średni czas trwania nagrań należących do danej grupy tempa. Dla wszystkich wartości α_o, $\mu(\alpha_{netto})$ jest większe niż $\mu(\alpha_{brutto})$, i obie wartości są mniejsze niż α_o. Różnice pomiędzy średnią wartością *netto* i średnią wartością *brutto* współczynnika skali są większe w przypadku mowy wypowiadanej w tempie wolnym, i dodatkowo wartości te są mniejsze dla mowy wypowiadanej w tempie średnim i szybkim. Przyczyny takiej sytuacji są dwie:

- dla mowy wypowiadanej w tempie wolnym występuje większa liczba fragmentów zawierających ciszę spowodowaną artykulacją (oddech, pauza) stąd większe różnice pomiędzy średnimi wartościami *netto* i *brutto* współczynnika skali,
- metoda B dokonuje adaptacji chwilowej wartości α w zależności od aktualnego tempa mowy wejściowej, w taki sposób by spowalniać mowę wolną w mniejszym stopniu niż mowę wypowiadaną w tempie szybkim (stąd wyższe średnie wartości współczynników skali w przypadku mowy wypowiadanej w szybszym tempie).

Skutkiem omówionych powyżej relacji jest fakt, iż pomimo dłuższego czasu trwania wypowiedzi, których ROS określane jest jako wolne, $\mu(o_t)$ zazwyczaj jest znacznie mniejsze niż w mowie wypowiadanej w tempie średnim i szybkim. Jedynie dla najwyższej wartości α_o, $\mu(o_t)$ jest wyższe niż dla pozostałych wartości ROS. Jak można zauważyć, dla pierwszych dwóch wartości α_o (1,25, 1,33), opóźnienie pomiędzy sygnałem wejściowym i wyjściowym jest niewielkie (od 0,35 do 1,59 s), biorą pod uwagę średni czas trwania wypowiedzi (13,14 – 19,94 s). Dwie najwyższe wartości α_o (1,5, 1,75) powodują powstanie znacznego opóźnienia, szczególnie dla $\alpha_o = 1{,}75$. Można jednak założyć, iż w szczególnych sytuacjach (np. podczas zajęć szkolnych) takie opóźnienie nie będzie powodowało znacznego utrudnienia, ponieważ nauczyciel prowadzący zajęcia, dość często robi dłuższe pauzy, które będą powodowały, iż chwilowo powstałe opóźnienie będzie szybko redukowane dzięki eliminacji fragmentów ciszy przez opracowany algorytm.

Tab. 5.10 Średnie wartości współczynników skali oraz średnia różnica długości wypowiedzi uzyskane dla mowy szybkiej, średniej i wolnej spowolnionej za pomocą opracowanej metody B.

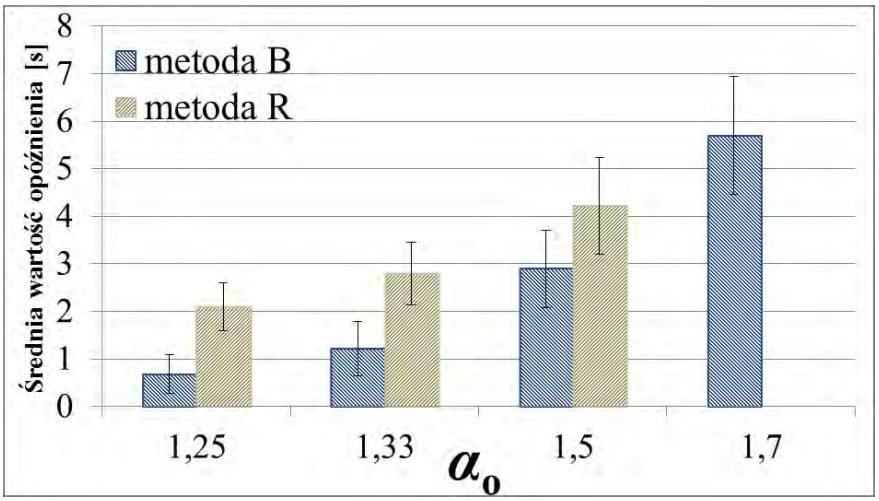
α_o	ROS	$\mu(\alpha_{netto})$	$\mu(\alpha_{brutto})$	$\mu(o_t)$ [s]	$\mu(t_{wej})$ [s]
1,75	szybkie	1,51	1,42	5,57	13,14
	średnie	1,48	1,37	5,75	15,82
	wolne	1,45	1,3	5,79	19,94
	średnia	1,48	1,36	5,70	16,30
1,5	szybkie	1,3	1,25	3,15	13,14
	średnie	1,27	1,20	2,99	15,82
	wolne	1,25	1,14	2,55	19,94
	średnia	1,27	1,2	2,90	16,30
1,33	szybkie	1,15	1,12	1,59	13,14
	średnie	1,13	1,08	1,28	15,82
	wolne	1,12	1,04	0,78	19,94
	średnia	1,14	1,08	1,22	16,30
1,25	szybkie	1,1	1,08	0,98	13,14
	średnie	1,09	1,05	0,71	15,82
	wolne	1,08	1,02	0,35	19,94
	średnia	1,09	1,05	0,68	16,30

W dalszej części eksperymentu zbadano opóźnienia oraz wyznaczono średnie wartości współczynników skali (*netto* i *brutto*), osiągane przez metodę R. Wyniki przedstawiono w tab. 5.11. Podobnie jak w metodzie B można tu zaobserwować, iż różnice pomiędzy średnimi wartościami *netto* i *brutto* współczynników skali są większe dla mowy wypowiadanej w tempie wolnym. Jednak różnie te są wyraźnie mniejsze niż dla opracowanej metody B. Skutkiem tego jest fakt, iż średnie wartości opóźnień pomiędzy sygnałem wejściowym i zmodyfikowanym rosną wraz z obniżaniem tempa mowy wejściowej. Omówione zależności wynikają z faktu, iż metoda zaproponowana przez Nejime *et al.* nie uzależnia procesu TSM od tempa mowy wejściowej.

Tab. 5.11 Średnie wartości współczynników skali oraz średnia różnica długości wypowiedzi uzyskane dla mowy szybkiej, średniej i wolnej spowolnionej za pomocą metody R.

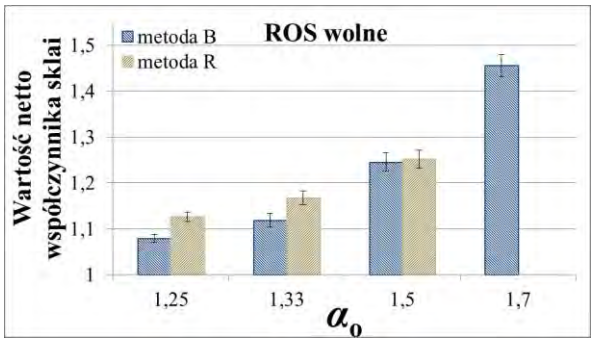
α_o	ROS	$\mu(\alpha_{netto})$	$\mu(\alpha_{brutto})$	$\mu(o_t)$ [s]	$\mu(t_{wej})$ [s]
1,5	szybkie	1,28	1,28	3,66	13,14
	średnie	1,27	1,26	4,17	15,82
	wolne	1,25	1,24	4,83	19,94
	średnia	1,27	1,26	4,22	16,30
1,33	szybkie	1,19	1,19	2,45	13,14
	średnie	1,18	1,18	2,78	15,82
	wolne	1,17	1,16	3,17	19,94
	średnia	1,18	1,17	2,80	16,30
1,25	szybkie	1,14	1,14	1,85	13,14
	średnie	1,13	1,13	2,09	15,82
	wolne	1,13	1,12	2,35	19,94
	średnia	1,14	1,13	2,10	16,30

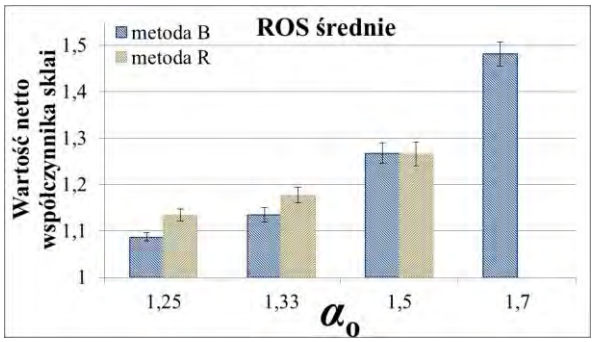
Na rys. 5.5 wykreślono wartości $\mu(o_t)$ uzyskane dla metod B i R. Można zauważyć, iż dla pierwszych dwóch wartości α_o różnice te są największe. Dla $\alpha_o = 1,5$ różnica pomiędzy opóźnieniem wprowadzanym przez metodę R a tym wprowadzanym przez metodę B wynosi 1,32 s. Jest to znacząca różnica biorąc po uwagę konieczność pracy w czasie rzeczywistym. Jest ona prawie równa różnicy pomiędzy $\mu(o_t)$ uzyskaną dla metody R i $\alpha_o = 1,5$, a $\mu(o_t)$ uzyskaną dla metod B i $\alpha_o = 1,75$ (1,48 s). Przedstawione wyniki pokazują, iż metoda B pozwala uzyskiwać mniejsze wartości opóźnień niż metoda R.

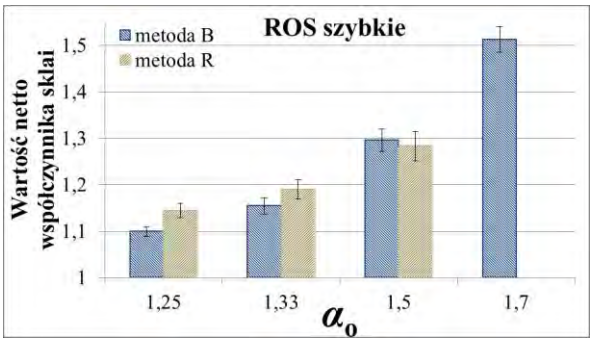


Rys. 5.5 Porównanie średniej wartości opóźnień uzyskanych przez metody B i R (wyniki uśrednione dla trzech wartości tempa mowy: wolnego, średniego i szybkiego).

Pozostaje pytanie, czy uzyskane różnice w wartościach opóźnień wynikają jedynie z przedstawionych powyżej różnic w metodach, czy też przyczyna jest inna. W celu odpowiedzenia na to pytanie porównano średnie wartości *netto* współczynników skali osiągnięte przez obie metody. Na rys. 5.6 przedstawiono to porównanie dla mowy wypowiadanej w tempie wolnym, średnim i szybkim. Można zauważyć, iż jedynie dla $\alpha_o = 1{,}5$ różnice pomiędzy $\mu(\alpha_{netto})$ jest znikoma (od 0 do 0,02), w pozostałych przypadkach ($\alpha_o = 1{,}25$ i 1,33) różnica ta jest znacząca (od 0,4 do 0,5). Stąd znacząco niższa wartość $\mu(o_t)$ osiągnięta przez metody B dla $\alpha_o = 1{,}25$ i 1,33 nie jest jedynie wynikiem wprowadzenia zależności w doborze chwilowej wartości α od aktualnego tempa mowy, lecz również powstaje z powodu używania niższych średnich wartości α_{netto} w tej metodzie.







Rys. 5.6 Porównanie średniej wartości *netto* współczynników skali uzyskane przez metody B i R (wyniki uzyskane dla mowy wypowiadanej w tempie: wolnym, średnim i szybkim).

5.6 Ocena jakości mowy spowolnionej

W ramach badań nad subiektywną oceną jakość mowy spowolnionej za pomocą opracowanych metod, przeprowadzono serię testów odsłuchowych. Procedura testowa miała umożliwić określenie jakości oraz naturalności mowy zmodyfikowanej. Jakość sygnału w trakcie badań oceniana była za pomocą skali DCR (ang. *Degradation Category Rating*), która jest rekomendowana przez normę ITU-T P.800 [60] jako skala pozwalająca na wykrycie niewielkich różnic w jakości mowy. Skala ta została opracowana w celu badania jakości kodeków mowy [17]. Ocena mowy zmodyfikowanej odbywała się poprzez porównanie mowy przetworzonej z nagraniem referencyjnym. Zgodnie z normą ITU-T P.800 skala DCR jest 5 stopniowa, a znaczenie poszczególnych wartości jest następujące:

1. Zniekształcenia są bardzo dokuczliwe
2. Zniekształcenia są dokuczliwe
3. Zniekształcenia są nieznacznie dokuczliwe
4. Zniekształcenia są słyszalne, ale niedokuczliwe
5. Zniekształcenia są niesłyszalne

Wartością określającą jakość mowy w danych warunkach (np. dla określonej metody modyfikacji sygnału) jest parametr DMOS (ang. *Degradation Mean Opinion Score*) wyznaczony jako średnia wartość ocen wszystkich osób biorących udział w teście.

Naturalność mowy spowolnionej oceniono korzystając ze skali ACR (ang. *Absolute Category Rating*) opracowanej do celów oceny kodeków mowy wykazujących znaczące różnice w jakości [60]. Podczas testów słuchacz ma za zadanie ocenić określony parametr (konkretnie naturalność) korzystając z pięciostopniowej skali:

1. Bardzo niska

2. Niska
3. Dostateczna
4. Wysoka
5. Bardzo wysoka

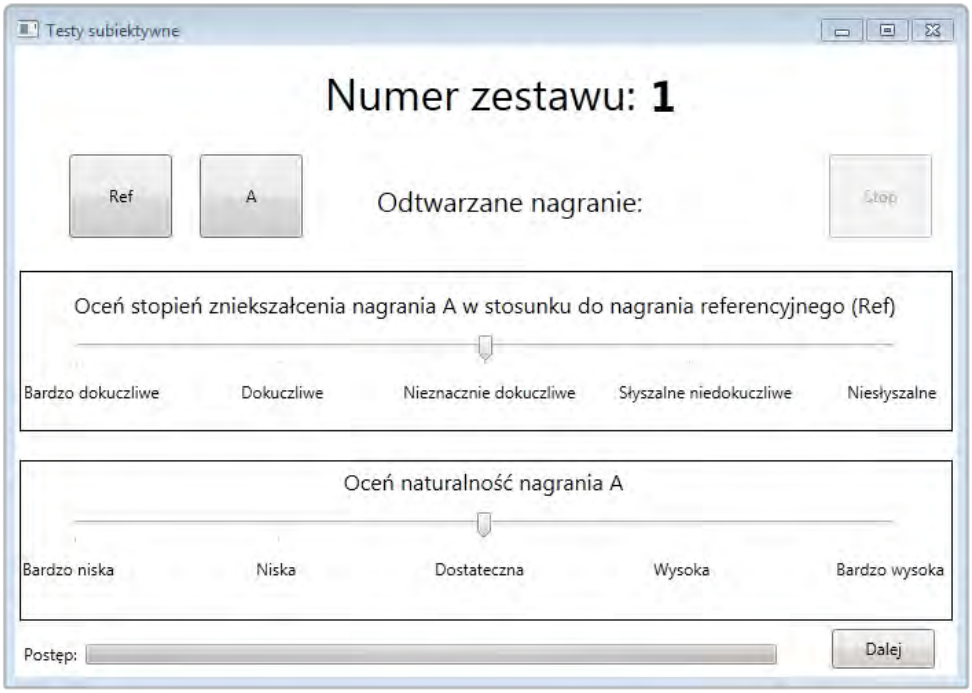
Parametrem opisującym wynik testu (przy danych warunkach) jest MOS (ang. *Mean Opition Score*). MOS jest odpowiednikiem parametru DMOS i obliczany jest jako wartość średnia ocen uzyskanych (w danych warunkach) dla wszystkich osób biorących udział w badaniu.

5.6.1. Metodologia badań

Podczas badań ocenie poddano mowę zmodyfikowaną za metodą B oraz mowę zmodyfikowaną z wykorzystaniem algorytmu zaproponowanego przez Nejime *et al.* Zgodnie z zaleceniami zawartymi w normie ITU-T P.800 [60] podczas testów modyfikacji poddano mowę zarejesrtowaną przez mówcę męskiego i żeńskiego. Jednak mimo, iż w normie zaleca się wykorzystanie mowy pochodzącej od czterech mówców (2 męskich i 2 żeńskich) w badaniach użyto jedynie dwóch mówców. Było to uzasadnione, ponieważ ocena 4 różnych głosów wiązałaby się ze znaczącym wydłużeniem trwania testu (do około 20–40 minut). Ocenę każdej z metod (B i R) wykonano dla trzech różnych wartość współczynnika skali: 1,25, 1,33, 1,5. Wartości te zostały wybrane, ponieważ metoda R pozwala na użycie wyłącznie tych wartości α_o (jest to związane z modyfikacją wprowadzoną przez Nejime *et al.* w algorytmie TDHS, która ma na celu zwiększenie jakości mowy spowolnionej). Dodatkowo w metodzie B ocenie podano także mowę spowolnioną z użyciem $\alpha_o = 1{,}75$. Wartość ta została wybrana, ponieważ była ona wykorzystywana w rozprawie podczas badań wpływu opracowanych metod modyfikacji sygnału mowy na jej rozumienie przez osoby z pogorszoną czasową rozdzielczością słuchu (rozdział 4). Dodatkowo, użyto także dwóch różnych prędkości mowy wejściowej: wolnej i szybkiej. Miało to na celu zbadanie wpływu tempa mowy wejściowej na jakość i naturalność mowy zmodyfikowanej. Każdy zestaw nagrań zawierał także dwa nagrania nazwane kontrolnymi. Zawierały one zamiast mowy zmodyfikowanej mowę niezmodyfikowaną wypowiadaną w tempie wolnym. Słuchacz nie wiedział o istnieniu tego typu nagrań i oceniał mowę wypowiadaną w tempie wolnym myśląc, że jest to mowa zmodyfikowana przez algorytm. Zestawy kontrolne zostały dodane zgodnie z sugestiami Combescure *et al.* [17] w celu wykrycia tego, czy słuchacze są wiarygodni. W efekcie

pełen zestaw testowy składał się z 30 nagrań (28 nagrań ocenianych i dwóch nagrań kontrolnych).

Badania przeprowadzono z wykorzystaniem aplikacji opracowanej zgodnie z sugestiami przedstawionymi przez Bech i Zacharova. [10]. Interfejs graficzny tej aplikacji zamieszczono na rys. 5.7. Każdej osobie biorącej udział w testach przedstawiono sposób działania aplikacji oraz przekazano ustnie instrukcję związane ze sposobem oceny nagrań. Badania przeprowadzono w cichym pomieszczeniu. Odsłuch nagrań odbywał się poprzez słuchawki. Poziom odsłuchu był zawsze ten sam, ale każdy ze słuchaczy, w razie potrzeby, mógł przed rozpoczęciem badania dostosować go tak, by słyszeć mowę na komfortowym poziomie. Podczas testów wykorzystano kluczowanie swobodne (każde nagranie w zestawie mogło być odtwarzane wielokrotnie). Czas badania dla jednego słuchacza wahał się w przedziale od 10 do 20 minut. Na płycie dołączonej do rozprawy zamieszczono nagrania wykorzystane podczas testów a w załączniku nr 6 opis zawartości katalogu z nagraniami.



Rys. 5.7 Interfejs graficzny aplikacji wykorzystywanej podczas testów.

5.6.2. Analiza wyników

W badaniach wzięło udział 29 osób. Spośród nich wyłoniono grupę „ekspertów”. Znalazły się w niej osoby, które w teście oceniły jakość i naturalność nagrań niezmodyfikowanych na wysoką. Reguły klasyfikacji słuchacza do grupy „ekspertów” były następujące, słuchacz ocenił:

- zarówno naturalność jak i jakość obu niezmodyfikowanych nagrań na 5 lub,
- ocenił naturalność i jakość jednego z nagrań na 5, a oceny drugiego nagrania nie były niższe niż 4 bądź,
- w obu nagraniach jeden z parametrów oceniono na 5 a drugi na 4.

Wybór osób, które uznały jakość nagrania testowego na 4 jako „ekspertów" było związane z faktem, iż oba nagrania zarejestrowano osobno (nagranie referencyjne: mowa wypowiadana szybko; nagranie oceniane: mowa wypowiadana wolno). Jako że nagrania wykonano za pomocą typowego mikrofonu komputerowego, charakterystyka szumu pojawiającego się w nagraniu mogła się nieznacznie różnić pomiędzy nagraniami, a to mogło prowadzić do obniżenia oceny jakości. Podobnie przy ocenie naturalności, podczas kolejnego czytania tej samej frazy, lektor mógł wypowiedzieć zdanie w nieco inny sposób niż poprzednio, co prowadziło do obniżenia oceny naturalności. W grupie osób biorących udział w badaniach 23 osoby (na 29) spełniły powyższe wymagania, i podczas analizy tylko oceny tych osób były brane pod uwagę.

Analizę uzyskanych wyników przeprowadzono niezależnie dla obu parametrów (jakości i naturalności). Podczas szacowania istotności różnic ocen nie wykorzystano wieloczynnikowego testu ANOVA[15], ponieważ rozkład ocen pośród różnych czynników nie był normalny (rozkład zbadano za pomocą testu Shapiro-Wilka) a liczba „ekspertów" była mniejsza niż 30, więc nie można było założyć rozkładu normalnego danych opierając się na centralnym twierdzeniu granicznym (ang. *Central Limit Theorem* – CLT). Możliwa była jednak analiza różnic rozkładów ocen za pomocą nieparametrycznego testu Friedmana, który jest odpowiednikiem dwuczynnikowego testu ANOVA lub jednoczynnikowego testu ANOVA z powtórzeniami.

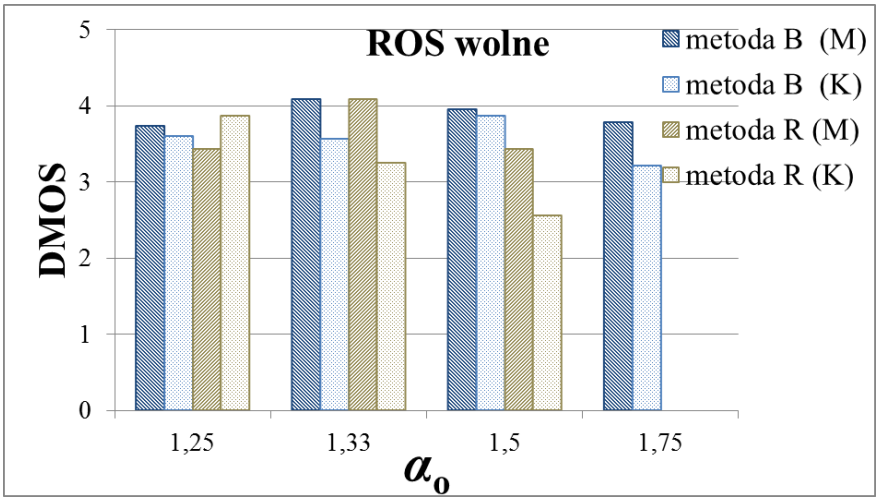
Ocena jakości sygnału mowy poddanego modyfikacji tempa

Na rys. 5.8 i 5.9 przedstawiono oceny jakości uzyskane dla mowy wypowiadanej w tempie wolnym i szybkim. Litery M i K umieszczone w nawiasach obok nazwy metody TSM oznaczają odpowiednio mówcę męskiego albo żeńskiego. Analizując uzyskane wyniki dla mowy wypowiadanej w tempie wolnym można zauważyć, iż oceny osiągane dla mówcy męskiego są prawie zawsze wyższe od ocen mówcy żeńskiego. Wyjątek stanowi jedynie ocena osiągnięta przez metodę R dla $\alpha_o = 1{,}25$. Porównując natomiast

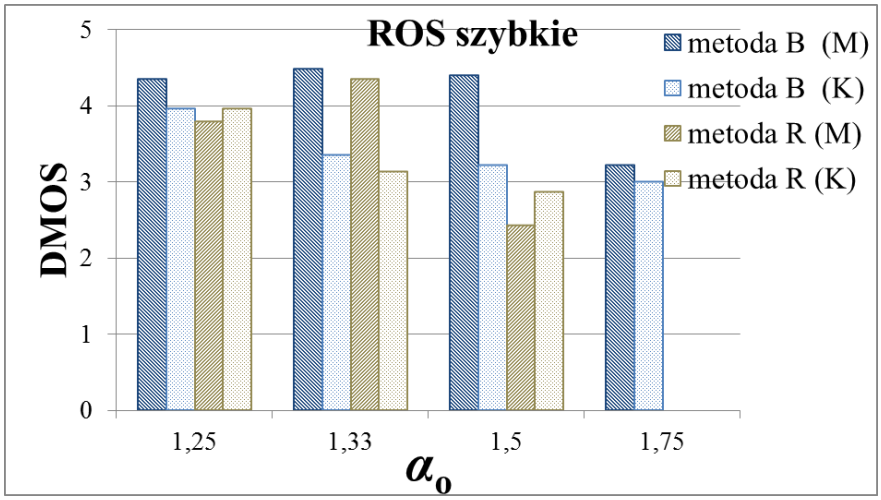
[15] W tym badaniu należałoby użyć czteroczynnikowego testu ANOVA, gdzie czynnikami są: płeć mówcy, tempo mowy wejściowej, rodzaj algorytmu spowalniania oraz współczynnik skali.

wartości ocen uzyskane metodą B i R dla mówcy męskiego i osobno dla mówcy żeńskiego widać, iż metoda B została prawie zawsze wyżej oceniona niż metoda R niezależnie od wartości α_o. Można było tu oczekiwać sytuacji, w której wraz ze wzrostem wartości współczynnika skali wartości ocen zaczną maleć (jest to typowe zjawisko dla wszystkich metod TSM), jednak tutaj oceny uzyskane dla $\alpha_o = 1{,}33$ są wyższe, niż dla $\alpha_o = 1{,}25$. Należy także zauważyć, iż najwyższe różnice ocen pomiędzy algorytmami widoczne są przy $\alpha_o = 1{,}5$ (na korzyść metody B), a przy $\alpha_o = 1{,}75$ metoda B została oceniona wyżej niż metoda R dla $\alpha_o = 1{,}5$.

Analizując wartości ocen uzyskane dla mowy wypowiadanej w tempie szybkim (rys. 5.9) można zaobserwować podobne relacje pomiędzy ocenami, co dla mowy wypowiadanej w tempie wolnym. Jednak przy tym ROS różnice ocen pomiędzy mówcą męskim i żeńskim w metodzie B są znacznie wyższe, niż ma to miejsce w sytuacji, gdy fraza wypowiadana jest tempie wolnym (rys. 5.8).



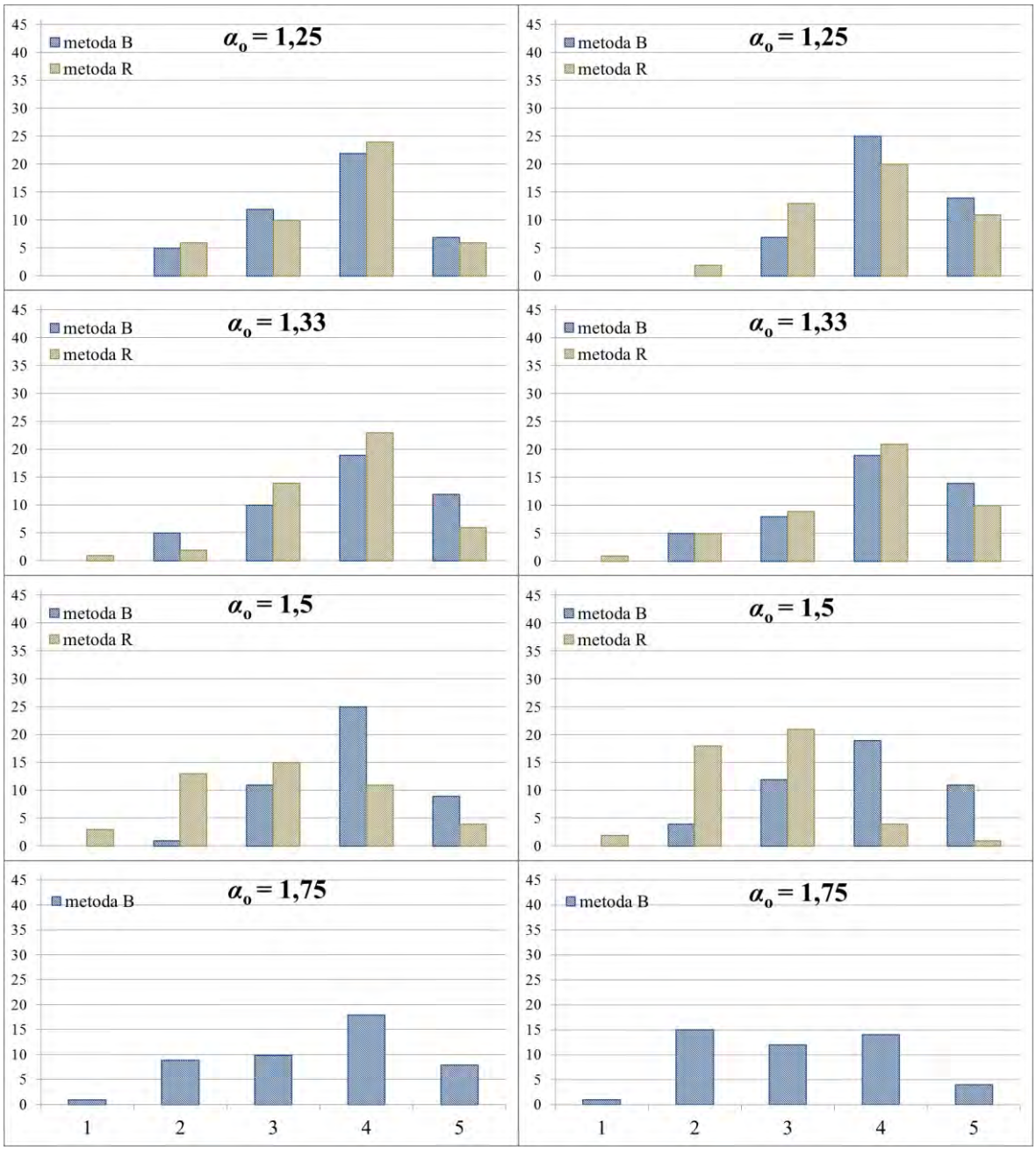
Rys. 5.8 Oceny DMOS uzyskane dla mowy wypowiadanej w tempie wolnym.



Rys. 5.9 Oceny DMOS uzyskane dla mowy wypowiadanej w tempie szybkim.

Na rys. 5.10 zamieszczono histogramy rozkładów ocen jakości wyrażonych w skali DCR. Przedstawione wyniki uzyskano dla różnych wartości tempa mowy wejściowej i $\alpha_o = \{1{,}25, 1{,}33, 1{,}5, 1{,}75\}$. Jak można zaobserwować, dla dwóch najniższych wartości α_o dominującą oceną jest 4. Należy zauważyć, iż metoda B uzyskała więcej ocen 5 oraz mniejszą liczbę ocen 2 niż metoda R. Tendencja ta jest widoczna zarówno dla mowy wypowiadanej w tempie wolny jak i szybkim, przy czym dla $\alpha_o = 1{,}33$ różnice te są większe, niż dla $\alpha_o = 1{,}25$. Dla $\alpha_o = 1{,}5$ oceny metody B mają rozkład podobny do rozkładu uzyskanego dla $\alpha_o = 1{,}33$ (z dominującą oceną 4 dla obu temp wypowiedzi), a rozkład ocen dla metody R jest przesunięty w kierunku wartości środkowej (3). Można tu zaobserwować znaczną liczbę ocen równych 2. Metoda B dla $\alpha_o = 1{,}75$ została oceniona znacznie wyższej, gdy modyfikowana wypowiedź miała tempo wolne (dominująca ocena wynosiła 4). Mowa wypowiadana w tempie szybkim osiągnęła rozkład ocen zbliżony do rozkładu normalnego z dominującą wartością równą 2. Jest on zbliżony do rozkładu ocen uzyskanych przez metodę R w przypadku wolnego ROS i $\alpha_o = 1{,}5$.

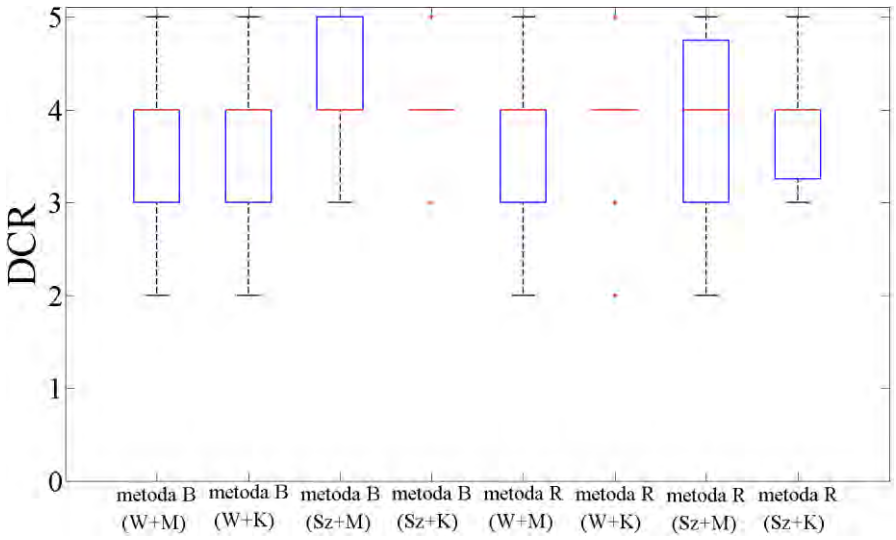
W celu sprawdzenia tego, czy uzyskane różnice ocen pomiędzy metodami B i R są istotne statystycznie, wykonano analizę różnic rozkładów uzyskanych w ramach tych samych wartości α_o. Analizę wykonano za pomocą testu Friedmana. Hipotezą zerową testu jest zgodność rozkładu prawdopodobieństw wartości DMOS w ramach różnych wariantów metod modyfikacji czasu trwania sygnału (B, R), różnego ROS mowy wejściowej oraz różnych mówców (M i K).



Rys. 5.10 Histogramy wartości DCR uzyskane dla B i R. W lewej kolumnie umieszczono rozkłady ocen mowy wypowiadanej w tempie wolnym, a w prawej kolumnie – mowy wypowiadanej w tempie szybkim.

Na rys. 5.11 przedstawiono wykres „ramka-wąsy" ocen jakości sygnału mowy zmodyfikowanej uzyskanych przy $\alpha_o = 1{,}25$. Można zauważyć, iż wartość mediany wszystkich wariantów wynosi 4 (czerwona linia). Znaczenie symboli w nawiasach obok nazwy metody jest następujące: mówca K – żeński, M – męski; tempo mowy wejściowej W – wolne, Sz – szybkie. Obliczona wartość statystyki testu Friedmana dla tych danych wyniosła $\chi^2(7) = 19{,}94$ i $p = 0{,}01$. Ponieważ osiągnięty poziom istotności jest mniejszy od założonego poziomu istotności p_i, należy odrzucić hipotezę zerową testu mówiącą iż wszystkie rozkłady prawdopodobieństw ocen uzyskane dla $\alpha_o = 1{,}25$ są równe. W związku

z tym należy przyjąć hipotezę alternatywną mówiącą, iż rozkłady prawdopodobieństw ocen różnią się przynajmniej w jednej parze.



Rys. 5.11 Wykres „ramka-wąsy" ocen jakości uzyskany dla mowy zmodyfikowanej przy $\alpha_o = 1{,}25$.

W celu wykrycia par, w których rozkłady prawdopodobieństw ocen różnią się pomiędzy sobą, wykonano nieparametryczny test *post hoc* będący odpowiednikiem parametrycznego testu Fishera (LSD). Krytyczna wartość testu wyniosła $\chi^2(7)_{cv} = 27{,}19$. W tab. 5.12 umieszczono wartości statystyk obliczone dla kolejnych par. Pogrubioną czcionką oznaczono wartości przekraczające krytyczną wartość statystyki ($\chi^2(7)_{cv}$). Właśnie te pary mają różne rozkłady prawdopodobieństw ocen jakości. Analizę wyników przeprowadzono tak, by zaobserwować różnice występujące w ramach jednej metody przy jednym wspólnym czynniku (tempo lub płeć mówcy)[16] oraz różnice występujące pomiędzy metodami przy tych samych warunkach (ten sam mówca i to samo tempo wypowiedzi)[17]. Ten sposób analizy wyników testu *post hoc* zastosowano także w dalszej części rozdział (podczas interpretacji wyników uzyskanych dla kolejnych wartości α_o) oraz podczas analizy wyników oceny naturalności mowy zmodyfikowanej. I tak w metodzie B różnice w rozkładach prawdopodobieństw ocen wystąpiły w następujących parach: (M+W)·(M+Sz), (K+Sz)·(M+Sz). Oznacza to iż jakość mowy zmodyfikowanej za pomocą metody B dla $\alpha_o = 1{,}25$ i mówcy męskiego zależna jest od tempa mowy wejściowej a w przypadku mowy szybkiej jakość jest różna dla mówcy męskiego i żeńskiego. Natomiast nie istnieją statystycznie istotne różnice w ocenie jakości mowy zmodyfikowanej z użyciem metody R.

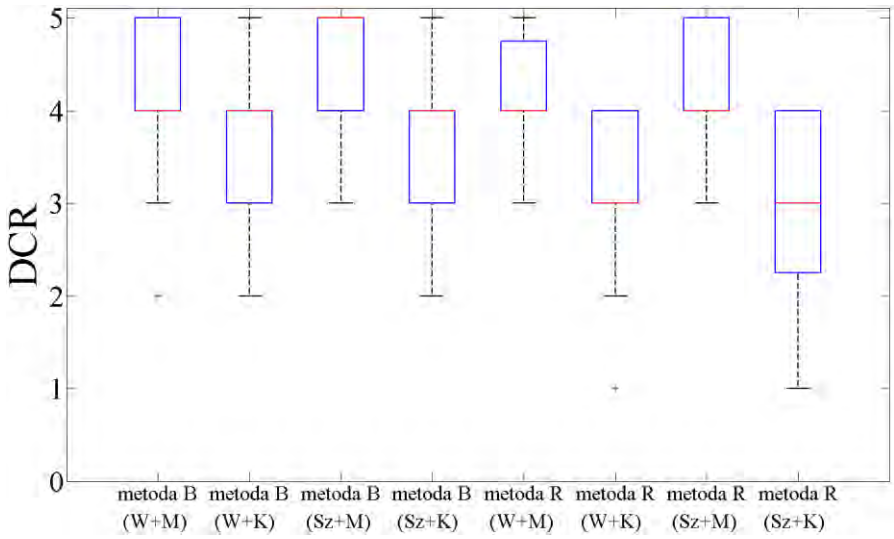
[16] Komórki tabeli odpowiadające tym parom wyróżniono pogrubioną linią ciągłą
[17] Komórki tabeli odpowiadające tym parom wyróżniono pogrubioną linią przerywaną

Można także zauważyć, iż dla mówcy męskiego i mowy szybkiej istnieje statystycznie istotna różnica w rozkładzie prawdopodobieństw ocen jakości pomiędzy metodą B (DMOS = 4,35) i R (DMOS = 3,78). Jak pokazują wyniki umieszczone w tab. 5.12 w pozostałych sytuacjach różnice w wartościach DMOS uzyskanych dla metody B i R nie są istotne statystycznie.

Tab. 5.12 Wyniki nieparametrycznego testu *post hoc* uzyskane dla mowy zmodyfikowanej przy $\alpha_o = 1{,}25$.

	met. B (M+W)	**met. B (K+W)**	**met. B (M+Sz)**	**met. B (K+Sz)**	**met. R (M+W)**	**met. R (K+W)**	**met. R (M+Sz)**	**met. R (K+Sz)**
met. B (M+W)	0	0	0	0	0	0	0	0
met. B (K+W)	10	0	0	0	0	0	0	0
met. B (M+Sz)	**38**	**48**	0	0	0	0	0	0
met. B (K+Sz)	10,50	20,50	**27,50**	0	0	0	0	0
met. R (M+W)	20	10	**58**	**30,50**	0	0	0	0
met. R (K+W)	7	17	**31**	3,50	27	0	0	0
met. R (M+Sz)	0,50	9,50	**38,50**	11	19,50	7,50	0	0
met. R (K+Sz)	11	21	27	0,50	**31**	4	11,50	0

Dla $\alpha_o = 1{,}33$ mediany ocen nie są jednakowe i przyjmują wartości z zakresu od 3 do 5 (rys. 5.12). Analiza różnic rozkładów prawdopodobieństw ocen za pomocą nieparametrycznego testu Friedmana pokazała, iż przynajmniej jeden rozkład prawdopodobieństw ocen jakości różni się od pozostałych ($\chi^2(7) = 61{,}68$; $p = 0$).



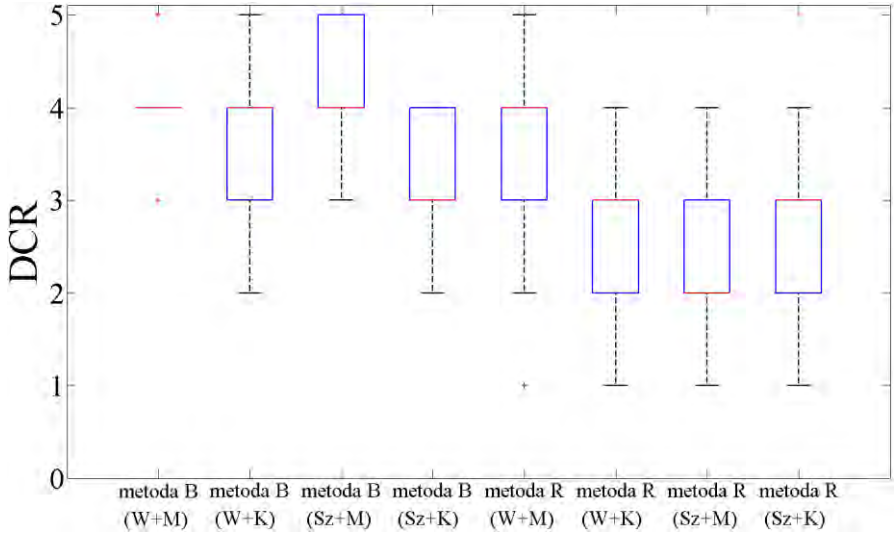
Rys. 5.12 Wykres „ramka-wąsy" ocen jakości uzyskany dla mowy zmodyfikowanej przy $\alpha_o = 1{,}33$.

Określenie par, które przyczyniły się do powstania różnic w rozkładach prawdopodobieństw ocen zostało wykonane za pomocą testu LSD. Wyniki testu umieszczono w tab. 5.13. Wartość krytyczna statystyki testu wyniosła $\chi^2(7)_{cv} = 22{,}92$. W przypadku metody B różnice w rozkładach prawdopodobieństw ocen wystąpiły w parach: (M+W)·(K+W) i (K+Sz)·(M+Sz). Oznacza to, że oceny jakości nie zależą od tempa mowy wejściowej i są różne dla mówcy męskiego i żeńskiego. Podobna sytuacja występuje w metodzie R. Dodatkowo, nie można zaobserwować istotnych statystycznie różnic w ocenach jakości sygnału pomiędzy metodami B i R.

Tab. 5.13 Wyniki nieparametrycznego testu *post hoc* uzyskane dla mowy zmodyfikowanej przy $\alpha_o = 1{,}33$.

	met. B (M+W)	**met. B (K+W)**	**met. B (M+Sz)**	**met. B (K+Sz)**	**met. R (M+W)**	**met. R (K+W)**	**met. R (M+Sz)**	**met. R (K+Sz)**
met. B (M+W)	0	0	0	0	0	0	0	0
met. B (K+W)	**32**	0	0	0	0	0	0	0
met. B (M+Sz)	19,50	**51,50**	0	0	0	0	0	0
met. B (K+Sz)	**48,50**	16,50	**68**	0	0	0	0	0
met. R (M+W)	3,50	**28,50**	**23**	**45**	0	0	0	0
met. R (K+W)	**49**	17	**68,50**	0,50	**45,50**	0	0	0
met. R (M+Sz)	17,50	**49,50**	2	**66**	21	**66,50**	0	0
met. R (K+Sz)	**52**	20	**71,50**	3,50	**48,50**	3	**69,50**	0

Ostatnia analizowana wartość α_o wynosiła 1,5. Na rys. 5.13 przedstawiono wykres „ramka-wąsy" ilustrujący rozkład ocen jakości w skali DCR uzyskany dla tej wartości współczynnika skali. Mediana ocen zawiera się w przedziale od 2 do 4. Obliczona wartość statystyki testu Friedmana wyniosła: $\chi^2(7) = 89{,}68$; $p = 0$. Dlatego należy odrzucić hipotezę zerową testu mówiącą o równości rozkładów prawdopodobieństw ocen i przyjąć hipotezę alternatywną (brak równości rozkładów prawdopodobieństw ocen).



Rys. 5.13 Wykres „ramka-wąsy" ocen jakości uzyskany dla mowy zmodyfikowanej przy $\alpha_o = 1{,}5$.

W tab. 5.14 umieszczono wyniki testu *post hoc*. Wartość krytyczna testu wyniosła $\chi^2(7)_{cv} = 20{,}75$. Jak można zauważyć jakość mowy zmodyfikowanej przy użyciu metody B jest zależna od płci mówcy wtedy, gdy jest ona wypowiadana w tempie szybkim. Dodatkowo występują także różnice w ocenach jakości pomiędzy mową wypowiadaną w tempie szybki i wolnym (zarówno dla mówcy męskiego jak i dla mówcy żeńskiego). Natomiast metoda R uzyskuje istotne statystycznie różnice, w ocenach w zależności od płci mówcy, a także w sytuacji, gdy mówcą jest mężczyzna w zależności od tempa mowy wejściowej. Porównując jakość mowy zmodyfikowanej obu metodami, widoczna jest istotna statystycznie różnica w wartościach ocen jakości w przypadku wszystkich par. Dla wszystkich par wyższe oceny uzyskała metoda B (rys. 5.8 i 5.9), przy czym największe różnice widoczne są dla mowy wypowiadanej w tempie wolnym przez kobietę (DMOS = 3,87 i 2,56) oraz mowy wypowiadanej w tempie szybkim przez mężczyznę (DMOS = 4,39 i 2,43).

Tab. 5.14 Wyniki nieparametrycznego testu *post hoc* uzyskane dla mowy zmodyfikowanej przy $\alpha_o = 1{,}5$.

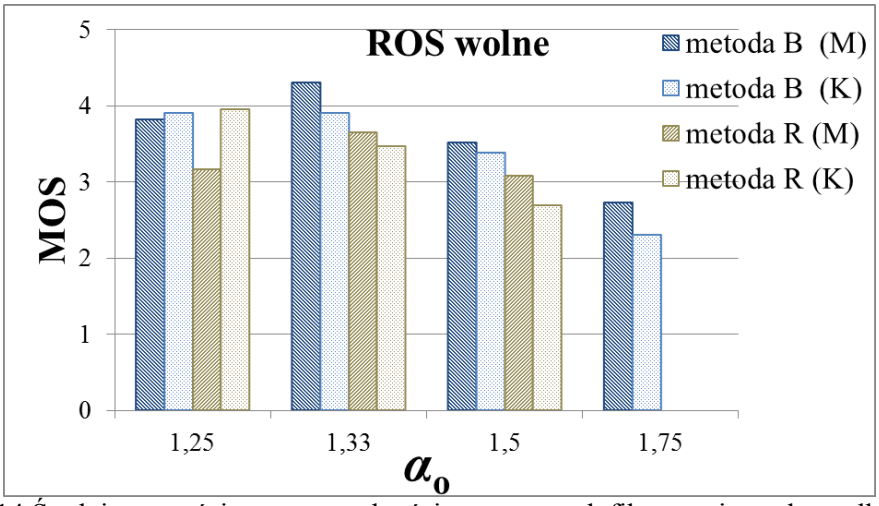
	met. B (M+W)	**met. B (K+W)**	**met. B (M+Sz)**	**met. B (K+Sz)**	**met. R (M+W)**	**met. R (K+W)**	**met. R (M+Sz)**	**met. R (K+Sz)**
met. B (M+W)	0	0	0	0	0	0	0	0
met. B (K+W)	0	0	0	0	0	0	0	0
met. B (M+Sz)	**24**	**24**	0	0	0	0	0	0
met. B (K+Sz)	**37**	**37**	**61**	0	0	0	0	0
met. R (M+W)	**27**	**27**	**51**	10	0	0	0	0
met. R (K+W)	**74,50**	**74,50**	**98,50**	**37,50**	**47,50**	0	0	0
met. R (M+Sz)	**85,50**	**85,50**	**109,50**	**48,50**	**58,50**	11	0	0
met. R (K+Sz)	**60**	**60**	**84**	**23**	**33**	14,50	**25,50**	0

W celu ogólnego porównania jakości obu metod wyznaczono także całkowitą jakość mowy spowolnionej przez każdą z nich. Wartość tą obliczono jako średnią wszystkich uzyskanych ocen danej metody niezależnie od tempa mowy wejściowej, płci mówcy oraz wartości współczynnika skali. Wartości te dla metody B i R wyniosły odpowiednio: 3,9 i 3,4. Widoczna jest znacznie wyższa ocena jakości uzyskana przez metodę B. Na podstawie przeprowadzonej analizy statystycznej należy przyjąć, iż opracowana metoda pozwala na uzyskanie znacznie wyższej jakości mowy spowolnionej dla wyższych wartości α_o. Dodatkowo, całkowita wartość DMOS jest bliska 4, która w skali DCR oznacza pojawianie się w sygnale zmodyfikowanym niedokuczliwych zniekształceń. Należy także zauważyć, iż jakość mowy zmodyfikowanej niekiedy zależna jest od płci mówcy oraz od tempa mowy wejściowej.

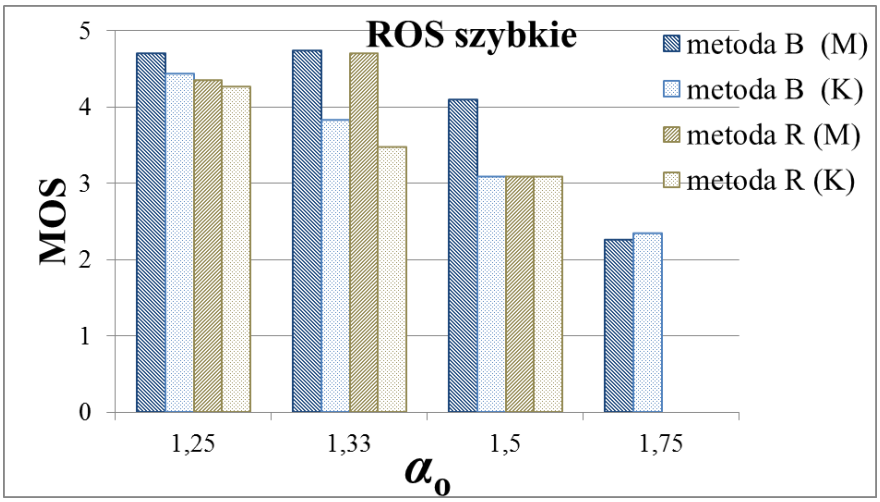
Ocena naturalności sygnału mowy poddanego modyfikacji tempa

W dalszej części niniejszego rozdziału analizie poddano wyniki ocen subiektywnych naturalności mowy spowolnionej. Na rys. 5.14 i 5.15 umieszczono oceny uzyskane dla mowy wypowiadanej w tempie wolnym i szybkim. Widoczna jest tendencja polegającą na obniżeniu oceny naturalności mowy spowolnionej wraz ze wzrostem wartości α_o. Może to być związane z tym, iż osoby biorące udział w badaniach zaczynały oceniać mowę spowolnioną z wyższymi wartościami α_o jako mało naturalną, ponieważ jej tempo było znacznie niższe niż tempo typowej wolnej wypowiedzi. Pewną regułą jest także to, iż

ocena naturalności spowolnionego głosu męskiego prawie zawsze jest wyższa niż oceny naturalności spowolnionego głosu żeńskiego. Należy także zauważyć, iż zmodyfikowana mowa wypowiadana w tempie szybkim jest oceniana wyżej niż zmodyfikowana mowa wypowiadana w tempie wolnym. Ta obserwacja zdaje się potwierdzać założenie dotyczące przyczyny obniżenia ocen naturalności pojawiającego się wraz ze wzrostem wartości α_o. Ostatnią istotną różnicą w wartościach ocen naturalności mowy spowolnionej jest różnica pojawiająca się pomiędzy ocenami metod B i R. Prawie w każdym z tych wariantów opracowana metoda uzyskuje wyższe wartości oceny naturalności niż metoda R.



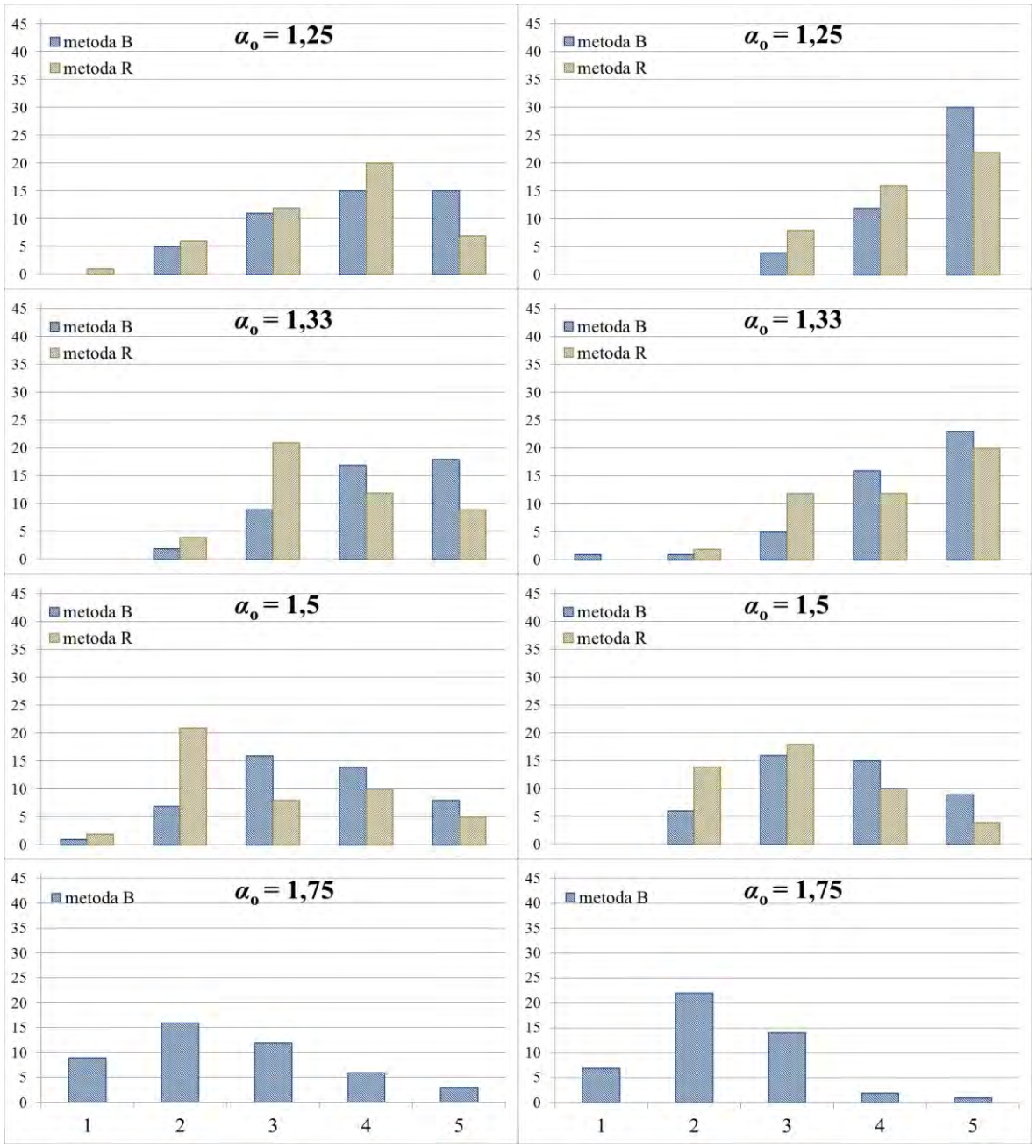
Rys. 5.14 Średnie wartości ocen naturalności mowy zmodyfikowanej uzyskane dla mowy wypowiadanej w tempie wolnym.



Rys. 5.15 Średnie wartości ocen naturalności mowy zmodyfikowanej uzyskane dla mowy wypowiadanej w tempie szybkim.

W celu dokładnej analizy różnic w wartościach ocen naturalności występujących pomiędzy analizowanymi metodami TSM na rys. 5.16 przedstawiono histogramy uzyskane dla mowy wypowiadanej w tempie wolnym i szybkim. Można zauważyć, iż dla α_o mieszczącego się w zakresie od 1,25 do 1,5 środek ciężkości ocen naturalności dla

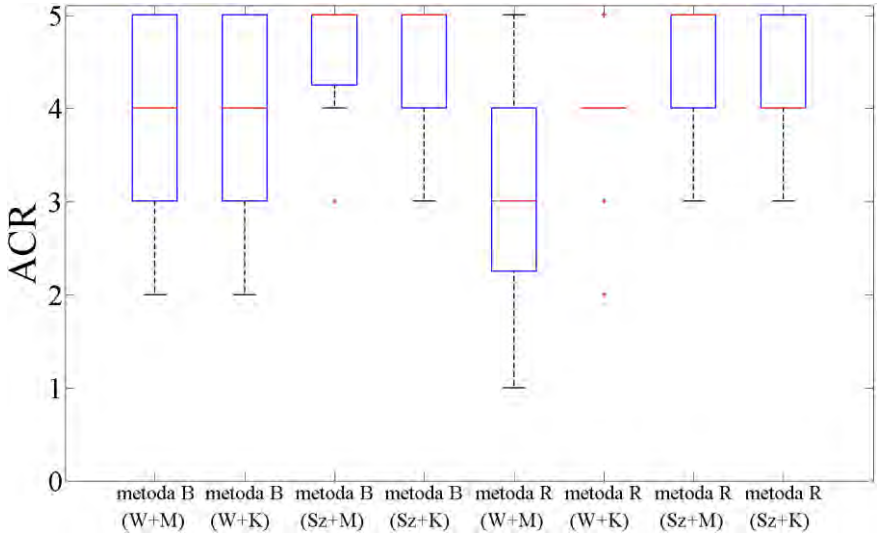
metody B przesunięty jest w stronę wyższych wartości. Przy czym dla $\alpha_o = 1{,}25$ i 1,33 dominującą oceną była 5 (bardzo wysoka naturalność). Metoda R dla tych samych wartości współczynnika skali uzyskała dominująca ocenę równą 5 jedynie dla mowy wypowiadanej w tempie szybkim.



Rys. 5.16 Histogramy ocen naturalności mowy uzyskane dla metod B i R. W lewej kolumnie umieszczono rozkłady ocen mowy wypowiadanej w tempie wolnym, a w prawej kolumnie mowy – wypowiadanej w tempie szybkim.

W dalszej części wykonano analizę statystyczną uzyskanych wyników dla $\alpha_o = \{1{,}25, 1{,}33, 1{,}5\}$. Na rys. 5.17 umieszczono wykres „ramka-wąsy" dla ocen naturalności mowy zmodyfikowanej dla $\alpha_o = 1{,}25$. Wartości mediany ocen mieszczą się w przedziale od 3 do

5. Wartość statystyki testu Friedmana pokazała, iż istnieją statystycznie istotne różnice w rozkładach prawdopodobieństw ocen naturalności ($\chi^2(7) = 53{,}24$; $p = 0$).



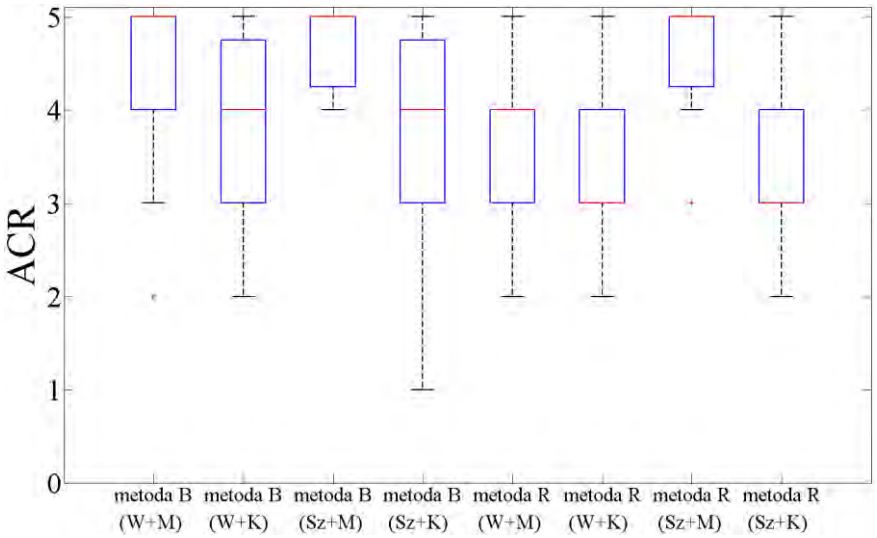
Rys. 5.17 Wykres „ramka-wąsy" dla ocen naturalności uzyskany dla mowy zmodyfikowanej przy $\alpha_o = 1{,}25$.

W celu określenia czynnika wpływającego na powstawanie różnic tych rozkładów wykonano test *post hoc*. Krytyczna wartość statystyki testu wyniosła $\chi^2(7)_{cv} = 23{,}60$. W tab. 5.15 umieszczono wartości statystyk. Na podstawie uzyskanych wyników można stwierdzić, iż naturalność mowy zmodyfikowanej metodą B zależna jest od tempa mowy wejściowej i jest niezależna od płci mówcy. Analizując wartości MOS przedstawione na rys. 5.14 i 5.15 widać, iż naturalność mowy jest wyższa, gdy jest ona wypowiadana w tempie szybkim. W metodzie R naturalność zależy od tempa mowy wejściowej i dodatkowo od płci mówcy (jeżeli mowa wypowiadana jest w tempie wolnym). Należy tu podkreślić, iż ocena naturalności mowy zmodyfikowanej jest wyższa wtedy, gdy tempo mowy wejściowej jest szybkie. Natomiast różnice w ocenach pomiędzy metodami są istotne statystycznie jedynie, gdy mówca jest mężczyzną i mowa wypowiadana jest w tempie wolnym (metoda B MOS = 3,82; metoda R MOS = 3,17).

Tab. 5.15 Wyniki nieparametrycznego testu *post hoc* uzyskane dla mowy zmodyfikowane przy $\alpha_o = 1{,}25$.

	met. B (M+W)	**met. B (K+W)**	**met. B (M+Sz)**	**met. B (K+Sz)**	**met. R (M+W)**	**met. R (K+W)**	**met. R (M+Sz)**	**met. R (K+Sz)**
met. B (M+W)	0	0	0	0	0	0	0	0
met. B (K+W)	1,50	0	0	0	0	0	0	0
met. B (M+Sz)	**51**	**49,50**	0	0	0	0	0	0
met. B (K+Sz)	**32**	**30,50**	19	0	0	0	0	0
met. R (M+W)	**39,50**	**41**	**90,50**	**71,50**	0	0	0	0
met. R (K+W)	1,50	3	**52,50**	**33,50**	**38**	0	0	0
met. R (M+Sz)	**29**	**27,50**	22	3	**68,50**	**30,50**	0	0
met. R (K+Sz)	23,50	22	**27,50**	8,50	**63**	**25**	5,50	0

Na rys. 5.18 umieszczono wykres „ramka-wąsy" ocen naturalności mowy zmodyfikowanej przy $\alpha_o = 1{,}33$. Widać, iż mediany wartości ACR są mocno zróżnicowane (zakres od 3 do 5), przy czym wyższe wartości uzyskano w przypadku metody B. Wartości statystki testu Friedmana pokazały, iż przynajmniej jeden z czynników wpływa na brak równości rozkładów prawdopodobieństw ocen naturalności ($\chi^2(7) = 65{,}24$; $p = 0$).



Rys. 5.18 Wykres „ramka-wąsy" dla ocen naturalności uzyskany dla mowy zmodyfikowanej przy $\alpha_o = 1{,}33$.

W dalszej części rozprawy określono czynniki powodujące powstanie różnic (test *post hoc*). Wartość krytyczna statystyki testu wyniosła $\chi^2(7)_{cv} = 22{,}08$, a wyniki testu umieszczono w tab. 5.16. Można zaobserwować, iż metoda B powoduje powstanie różnicy

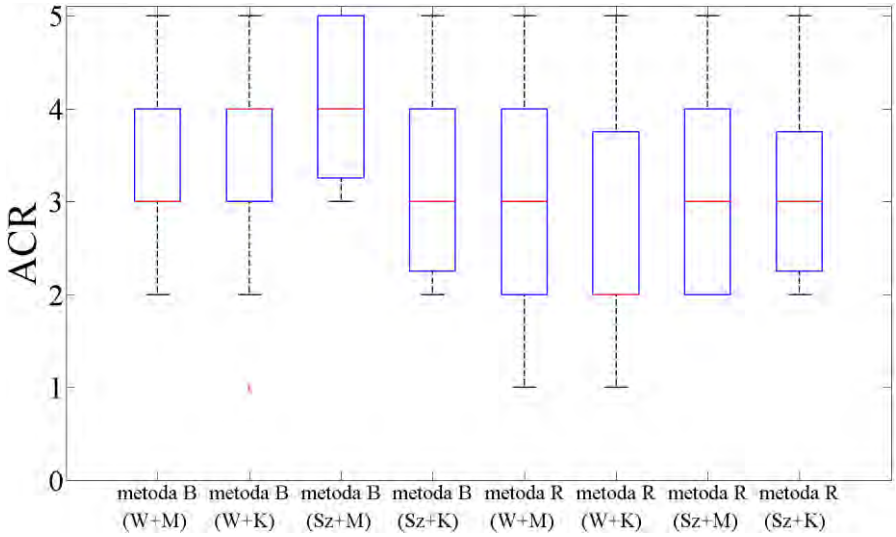
w ocenach naturalności mowy wypowiadanej przez mężczyznę i kobietę. Dodatkowo u mówcy męskiego różnice występują także w zależności od tempa mowy wejściowej. Naturalność zmodyfikowanej mowy męskiej oceniana jest wyżej niż naturalność zmodyfikowanej mowy żeńskiej. Podobna relacja widoczna jest i dla mowy wypowiadanej w tempie szybki i wolnym (rys. 5.14 i 5.15). Natomiast metoda R powoduje powstanie statystycznie istotnych różnic w ocenie naturalności mowy wypowiadanej przez mężczyznę w tempie szybkim i wolnym oraz, dla mowy wypowiadanej w tempie szybkim, pomiędzy głosem męskim i żeńskim. Różnice ocen są tu takie same, jak dla metody B, czyli głos męski oraz mowa wypowiadana w tempie szybkim uzyskują wyższe oceny naturalności niż głos żeński i mowa wypowiadana w tempie wolnym. Analizując wpływ rodzaju metody modyfikacji czasu trwania sygnału na ocenę naturalności widać, iż jest on statystycznie istotny w sytuacji, gdy mowa wypowiadana jest w tempie wolnym. Różnice w rozkładach prawdopodobieństw ocen naturalności mowy zmodyfikowanej występując wtedy zarówno u mówcy męskiego jak i żeńskiego. Naturalność mowy zmodyfikowanej za pomocą metody B oceniana jest wyżej niż naturalność mowy zmodyfikowanej za pomocą metody R (głos męski MOS = 4,3 i 3,65; głos żeński MOS = 3,91 i 3,48).

Tab. 5.16 Wyniki nieparametrycznego testu *post hoc* uzyskane dla mowy zmodyfikowanej przy $\alpha_o = 1{,}33$.

	Met. B (M+W)	**met. B (K+W)**	**met. B (M+Sz)**	**met. B (K+Sz)**	**met. R (M+W)**	**met. R (K+W)**	**met. R (M+Sz)**	**met. R (K+Sz)**
met. B (M+W)	0	0	0	0	0	0	0	0
met. B (K+W)	**26,50**	0	0	0	0	0	0	0
met. B (M+Sz)	**25,50**	**52**	0	0	0	0	0	0
met. B (K+Sz)	**29,50**	3	**55**	0	0	0	0	0
met. R (M+W)	**40**	13,50	**65,50**	10,50	0	0	0	0
met. R (K+W)	**51**	**24,50**	**76,50**	21,50	11	0	0	0
met. R (M+Sz)	20,50	**47**	5	**50**	**60,50**	**71,50**	0	0
met. R (K+Sz)	**51**	**24,50**	**76,50**	21,50	11	0	**71,50**	0

Ostatnia analizowana wartość α_o była równa 1,5. Rys. 5.19 przedstawia wykres „ramka-wąsy” wykonany dla ocen naturalności mowy zmodyfikowanej za pomocą metod B i R. Widoczne jest znaczne zróżnicowanie ocen naturalności w zależności od

uwzględnianych czynników (wartości median ocen mieszczą się w zakresie od 2 do 4). W celu zbadania, czy istnieje różnica przynajmniej w jednym z rozkładów prawdopodobieństw ocen naturalności, wykonano test Friedmana. Wyniki testu pokazały różność rozkładów prawdopodobieństw przynajmniej w jednej z analizowanych par ($\chi^2(7) = 38{,}11$; $p = 0$).



Rys. 5.19 Wykres „ramka-wąsy" ocen naturalności uzyskany dla mowy zmodyfikowanej przy $\alpha_o = 1{,}5$.

Dalsza analiza za pomocą testu *post hoc* pozwoliła wyznaczyć pary powodujące powstanie różnic w rozkładach prawdopodobieństw ocen naturalności. W tab. 5.17 umieszczono wyniki testu LSD (obliczona wartość krytyczna statystyki testu równa była 25,72). Różnice w ocenach naturalności mowy zmodyfikowanej za pomocą metody B są statystycznie istotne wtedy, gdy mówcą jest mężczyzna, a mowa wypowiadana jest w tempie wolnym i szybkim. Wyżej oceniana jest naturalność mowy zmodyfikowanej i wypowiadanej w tempie szybkim. Dodatkowo, przy tym tempie mowy istnieją różnice pomiędzy ocenami naturalności mowy wypowiadanej przez mężczyznę i kobietę (MOS = 4,09 i 3,09). Oceny naturalności mowy zmodyfikowanej metodą R różnią się w sposób istotny statystycznie w sytuacji, gdy fraza wypowiadana jest w tempie wolnym przez kobietę. Porównując oceny pomiędzy metoda (B i R) istotne statystycznie różnice widoczne są w sytuacji, gdy fraza wypowiadana jest przez kobietę w tempie wolnym (MOS = 3,69 i 2,69) lub przez mężczyznę w tempie szybkim (MOS = 4,09 i 3,09).

Tab. 5.17 Wyniki nieparametrycznego testu *post hoc* uzyskane dla mowy zmodyfikowanej przy $\alpha_o = 1{,}5$.

	Met. B (M+W)	**met. B (K+W)**	**met. B (M+Sz)**	**met. B (K+Sz)**	**met. R (M+W)**	**met. R (K+W)**	**met. R (M+Sz)**	**met. R (K+Sz)**
met. B (M+W)	0	0	0	0	0	0	0	0
met. B (K+W)	2	0	0	0	0	0	0	0
met. B (M+Sz)	**33,50**	**35,50**	0	0	0	0	0	0
met. B (K+Sz)	24,50	22,50	**58**	0	0	0	0	0
met. R (M+W)	21	19	**54,50**	3,50	0	0	0	0
met. R (K+W)	**47,50**	**45,50**	**81**	23	**26,50**	0	0	0
met. R (M+Sz)	22,50	20,50	**56**	2	1,50	25	0	0
met. R (K+Sz)	24	22	**57,50**	0,50	3	23,50	1,50	0

W celu bezpośredniego porównania ocen naturalności mowy zmodyfikowanej przy użyciu obu algorytmów obliczono także wartości średnie wszystkich ocen uzyskanych dla mowy przetworzonej przez obie oceniane metody modyfikacji. Średnie wartości ocen naturalności wyniosły odpowiedni 4,0 (metoda B) i 3,6 (metoda R). Widoczna jest znaczna różnica w całkowitych wartościach ocen naturalności pomiędzy metodami. Powody różnic tych wartości już przedstawiono powyżej (szczegółowa analiza statystyczna wartości ocen w tym podpunkcie). Jako że metoda B uzyskała średnią wartość oceny równą 4, co w skali ACR oznacza wysoką jakość jak również w wielu sytuacjach naturalność mowy zmodyfikowanej opracowaną metodą jest wyższa, niż dla mowy spowolnionej z wykorzystaniem metody R, można przyjąć, iż naturalność mowy zmodyfikowanej za pomocą tej metody jest wysoka.

5.7 Wnioski

Przeprowadzona analizy opracowanych metod detekcji i modyfikacji sygnału mowy wykazała, iż zapewniają one wysoką skuteczność detekcji oraz równie wysoką jakość i naturalność mowy przetworzonej. W testach detektorów pokazano, iż opracowany algorytm VAD umożliwia skuteczną detekcję ramek zawierających mowę przy równocześnie małym prawdopodobieństwie fałszywych alarmów w warunkach, gdy SNR ≥ 10 dB. Zaproponowany algorytm VRD uzyskuje małą częstość błędów wyrażoną

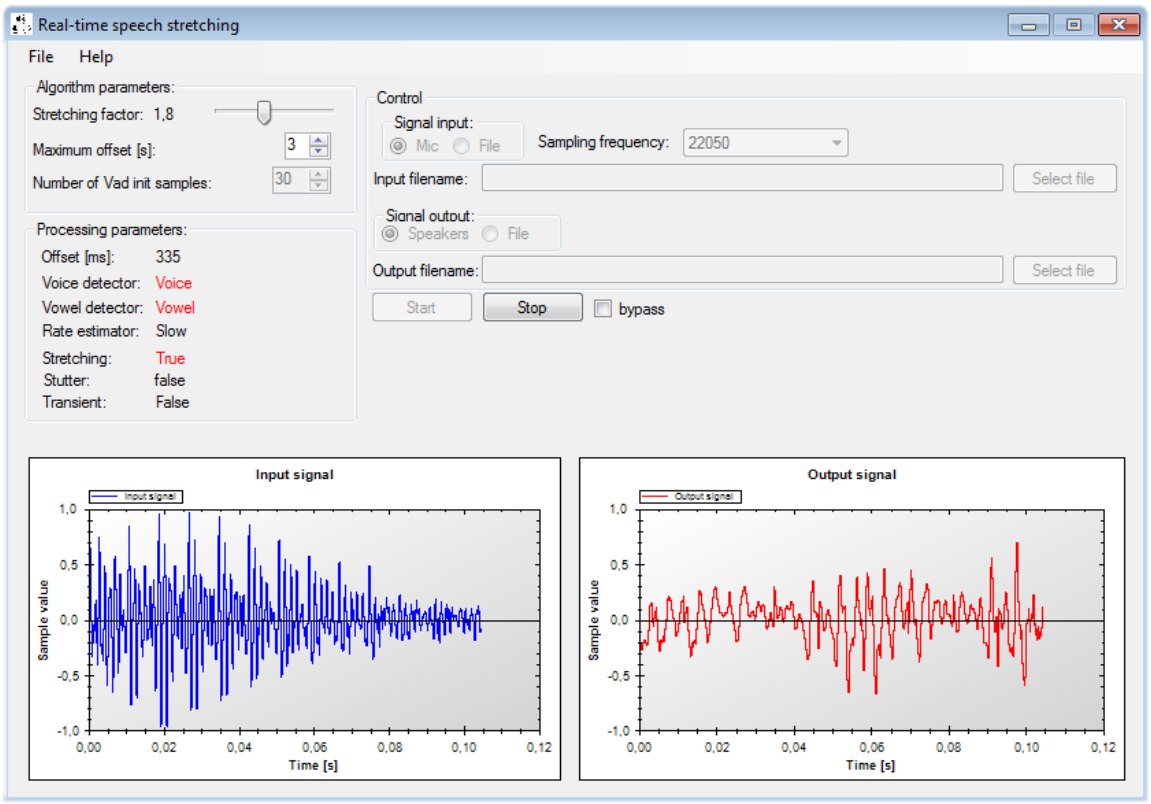
miarą VER i zadawalającą skuteczność detekcji samogłosek. Dodatkowo, operuje on w czasie rzeczywistym, tzn. ze zwłoką nieprzekraczającą 50 ms. Skuteczność estymacji kategorii ROS jest wysoka, przez co wykorzystanie jej w metodzie B pozwala na trafny dobór chwilowych wartości współczynnika skali.

Testy subiektywne wykazały, iż zarówno jakość jak i naturalność mowy spowolnionej za pomocą metody B jest wysoka (średnia ocena jakości wyniosła 3,9, a średnia ocena naturalności 4) i w zależności od analizowanych czynników często wyższa od jakości i naturalność mowy zmodyfikowanej za pomocą metody R. Istotność statystyczna różnic oceny jakości i naturalności mowy zmodyfikowanej została wykazana poprzez analizę statystyczną z użyciem testu Friedmana. Ważny jest także fakt, iż opóźnienie pomiędzy sygnałem wejściowym, a sygnałem spowolnionym wprowadzane przez metodę B jest niewielkie i znacznie mniejsze niż opóźnienie wprowadzane przez metodę R.

6 Opracowane oprogramowanie

Opracowana w ramach rozprawy metoda modyfikacji czasu trwania sygnału w czasie rzeczywistym została zaimplementowana w formie biblioteki programistycznej w języku c++. Dodatkowo powstały także trzy samodzielne aplikacje wykorzystujące opracowaną bibliotekę. Dwie z nich autor rozprawy opracował w ramach prac w projekcie TYPOSZEREG. Trzecia aplikacja została zaimplementowana przez zespół osób biorących udział w projekcie.

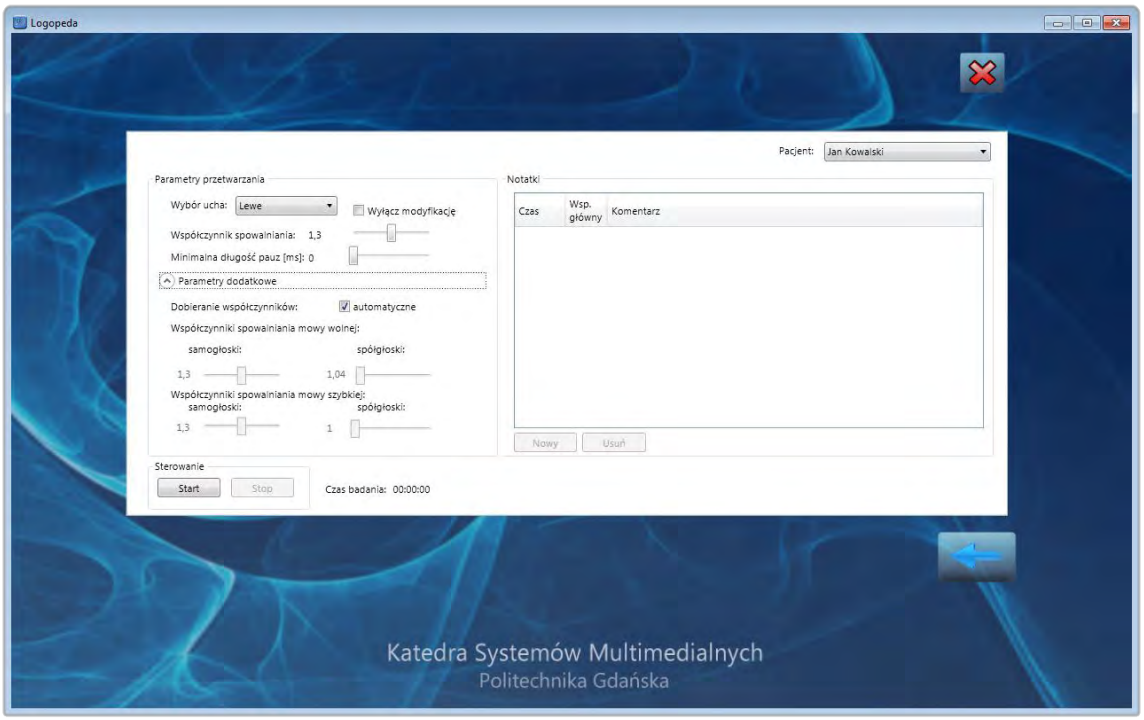
Pierwsza z aplikacji demonstruje sposób działania opracowanych metod analizy i modyfikacji sygnału mowy. Umożliwia ona spowalnianie w czasie rzeczywistym sygnału rejestrowanego przez mikrofon podłączony do komputera. W oknie aplikacji wyświetlane są aktualne wartości wyjść poszczególnych detektorów. Pozwala ona także odtwarzanie mowy zapisane w pliku i jej spowalnianie w czasie rzeczywistym. Na rys. 6.1 umieszczono zrzut ekranu przedstawiający interfejs użytkownika opracowanej aplikacji.



Rys. 6.1 Interfejs użytkownika aplikacji demonstrującej możliwości opracowanych metod modyfikacji sygnału mowy.

Druga aplikacja powstała w celu przeprowadzenia badań wpływu opracowanej metody modyfikacji sygnału mowy na osoby z różnego typu zaburzeniami m.in. osób

jąkających się, dzieci z dysleksją oraz osób z afazją. Była ona wykorzystywana przez lekarzy logopedów oraz pacjentów na Uniwersytecie Marii Curie-Skłodowskiej w Lublinie w okresie od stycznia do czerwca 2012. Spotkała się ona z dobrym przyjęciem i pozwoliła wykazać, iż także w analizowanych grupach słuchaczy opracowana metoda pozwala na prowadzenie skutecznego treningu słuchowego. W aplikacji udostępniono trzy główne funkcjonalności: zarządzanie bazą pacjentów, badanie oraz trening. Zarządzanie danymi pacjentów odbywa się poprzez bazę danych, w której oprócz danych pacjenta, zapisywane są także wyniki badań oraz treningu. W trakcie badania lekarz dobiera wartość współczynnika skali, który będzie wykorzystywany podczas treningu. Pacjent ma dostęp jedynie do opcji treningu, który polega na słuchaniu spowolnionej wersji własnej mowy. W aplikacji udostępniono dodatkowe parametry algorytmu modyfikacja czasu trwania takie jak, wartości współczynników skali wykorzystywane dla samogłosek i spółgłosek oraz wartości współczynników skali używane dla mowy wolnej i szybkiej. Lekarz może podczas badania w czasie rzeczywistym dostosowywać wartości współczynników i obserwować reakcję pacjenta. Dodatkowo możliwa jest także modyfikacja sygnału jedynie w jednym kanale (spowalnianie mowy np. w lewym uchu i odtwarzanie mowy oryginalnej w prawym). Ta funkcjonalność jest istotna z punktu widzenia grupy docelowej tej aplikacji, ponieważ umożliwia ona kształtowanie profilu lateralizacji. Na rys. 6.2 umieszczono zrzut ekranu demonstrujący interfejs użytkownika aplikacji.



Rys. 6.2 Interfejs użytkownika aplikacji opracowanej do celów badawczych.

Ostatnia aplikacja zaimplementowana została na urządzenia mobilne z system operacyjnym iOS. Umożliwia ona spowalnianie sygnału mowy rejestrowanej przez mikrofon wbudowany w aparat telefoniczny i odsłuchanie jej w czasie rzeczywistym przez słuchawki. Taka implementacja pozwala wykorzystać opracowaną metodę osobom z pogorszoną rozdzielczością czasową np. podczas rozmowy lub podczas zajęć lekcyjnych. Na rys. 6.3 przedstawiono interfejs użytkownika aplikacji mobilnej.

Rys. 6.3 Interfejs użytkownika aplikacji stworzonej na urządzenia mobilne.

7 Podsumowanie i wnioski

W rozprawie przedstawiono metody modyfikacji sygnału mowy stworzone w celu wspierania rozumienia mowy przez osoby z pogorszoną rozdzielczością czasową słuchu. Opracowane algorytmy detekcji samogłosek i estymacji ROS rozszerzają dotychczasowy stan wiedzy o możliwość operowania w czasie rzeczywistym przy zachowaniu wysokiej skuteczności oraz dokładności. Dodatkowo zaproponowana metoda modyfikacji czasu trwania sygnału mowy (metoda B) stanowi realizację postulowanych przez Coyla *et al.* [22] metod nierównomiernej modyfikacji czasu trwania sygnału mowy i rozszerzają je poprzez wprowadzenie uzależnienie sposobu modyfikacji sygnału od tempa mowy wejściowej. Druga z zaproponowanych metod (metoda C) jest autorskim rozwiązaniem i wprowadza nowy sposób doboru tempa mowy spowolnionej poprzez zastąpienie wartości współczynnika skali wartością ROS_o. W efekcie na wyjściu algorytmu uzyskuje się mowę o określonym tempie.

Testy rozumienia mowy spowolnionej przeprowadzone z udziałem dwóch grup słuchaczy (dzieci głuchych i osób starszych) pokazały, iż istnieją różnice w stopniu rozumieniu mowy zmodyfikowanej za pomocą opracowanych metod a rozumieniem mowy niezmodyfikowanej. Różnice te są istotne statystycznie jedynie w przypadku mowy wypowiadanej w tempie szybkim ($ROS^{PP}_{szybkie}$ i $ROS^{P}_{szybkie}$). W przypadku dzieci głuchych z pogorszoną rozdzielczością czasową słuchu wszystkie z przebadanych metod modyfikacji sygnału mowy, miały statystycznie istotny wpływ na poprawę jej rozumienia (wyniki testu Friedmana i testu *post hoc*). Najwyższy wzrost rozumienia mowy zaobserwowano w przypadku metody B (27,78%). Metoda A, polegająca na równomiernym spowalnianiu sygnału mowy, umożliwiała osiągnięcie niewiele słabszej poprawy w rozumieniu wypowiedzi (27,64%). Najsłabsze rezultaty uzyskano stosując metodę C (17,5%). Mogło to być związane z faktem, iż podczas testu wykorzystywano zdania o krótkim czasie trwania (2–3s), a estymator tempa wypowiedzi wykorzystywany w metodzie C używa 1,5s przedziału czasu w celu wyznaczenia aktualnego tempa mowy. Dlatego chwilowa wartość współczynnika skali jest dopasowywana do aktualnego tempa mowy z opóźnieniem równym 1,5s. Na początku pracy algorytmu, kiedy tempo mowy jest nieznane (przed upłynięciem pierwszego 1,5s przedziału czasu) założono, iż chwilowa wartość ROS wynosi 4 samogłoski/s (przeciętna wartość dla mowy wolnej). Następnie wartość współczynnika skali zaczyna szybko dążyć do założonej wartości ROS_d, przez co

w sygnale mowy pojawia się efekt nienaturalnej zmiany tempa. Dlatego metoda C powinna być stosowana głównie do spowolniania długich wypowiedzi. Dodatkowo wykazano, iż w przypadku dzieci głuchych z pogorszoną rozdzielczością czasową słuchu, istnieje korelacja pomiędzy wartością progu TCT_{50} i rozumieniem szybko wypowiadanej mowy (współczynnik korelacji –0,62). Pokazano także, iż istnieje korelację pomiędzy stopniem poprawy rozumienia mowy a wartością progu TCT_{50} wyznaczona dla metody A (współczynnik korelacji –0,59). Ostatnie spostrzeżenia pozwalają zakładać, iż do pewnego stopnia, możliwe jest oszacowanie, na podstawie wartości progu TCT_{50}, zysku jaki przyniesie zastosowanie równomiernego spowalniania sygnału mowy.

W przypadku osób starszych z pogorszoną rozdzielczością czasową słuchu jedynie metoda B pozwoliła osiągnąć istotną statystycznie poprawę rozumienia mowy (18,67%). Wynik ten jest spójny z wynikami badań Nejime *et al.* [122], w których zbadano wpływ metody spowalniania mowy opracowanej przez tych badaczy, na rozumienie mowy przez osoby starsze z pogorszoną rozdzielczością czasową słuchu (zmierzoną za pomocą testu RGDT). Nejime *et al.* wykazali istotną statycznie poprawę rozumienia mowy uzyskaną dzięki zastosowaniu ich metody nierównomiernej modyfikacji czasu trwania sygnału mowy (metoda ta była zbliżona do opracowanej w ramach tej rozprawy metody B).

W związku z tym, iż w przypadku obu badanych grup słuchaczy zaobserwowano istotną statystycznie poprawę współczynnika rozumienia (PRM) u osób z pogorszoną rozdzielczością czasową słuchu, a istotność wystąpiła w przypadku mowy spowolnionej z wykorzystaniem metody B i C (wyłącznie dzieci głuche) będących metodami nierównomiernej i uzależnionej od tempa wypowiedzi modyfikacji czasu trwania sygnału mowy, pierwsza teza rozprawy mówiąca, iż:

Zastosowanie nierównomiernej i zależnej od tempa wypowiedzi, modyfikacji czasu trwania mowy, powoduje wzrost współczynnika zrozumiałości u osoby o pogorszonej rozdzielczości czasowej słuchu.
została potwierdzona.

W dalszej części eksperymentów zbadano skuteczność opracowanych metod. Przeprowadzone analizy pokazały, iż opracowana metoda modyfikacji czasu trwania sygnału mowy (metoda B) osiąga wysokie oceny jakości oraz naturalności mowy dla współczynników skali zawierających się w zakresie od 1,25 do 1,5. W przypadku $\alpha_o = 1{,}75$ uzyskano dość niską ocenę naturalności mowy. Najprawdopodobniej jest to związane z

faktem, iż mowa zmodyfikowana z wykorzystaniem tak wysokiej wartości współczynnika skali wydawała się słuchaczom nienaturalnie wolna, mimo, iż brzmiała naturalnie pod względem prozodii. W pewny sensie obserwację tą potwierdzają wyniki oceny jakości uzyskane dla tej wartości α_o, gdzie jakość mowy spowolnionej została oceniona na podobnym lub wyższym poziomie co jakość mowy zmodyfikowanej za pomocą metody R. Dodatkowo prawie w każdej sytuacji metoda B została wyżej oceniona niż metoda R będąca pewnego rodzaju referencją.

Statystyczna analiza różnic rozkładu prawdopodobieństwa ocen nie wykazała jednoznacznego trendu w ocenach związanego z analizowanymi czynnikami (tempem mowy wejściowej, płcią mówcy, metodą modyfikacji). Widoczne są jednak pewnego rodzaju tendencję. I tak w przypadku oceny jakości niezależnie od użytej wartości α_o istotne statystycznie okazały się różnice pomiędzy oceną jakości głosu męskiego i żeńskiego w sytuacji, gdy mowa wypowiadana była w tempie szybkim. Zależność ta jest prawdziwa zarówno dla metody B jak i R[18] (tab. 5.10–5.12). Dodatkowo jakość mowy zależy także od płci mówcy, jeżeli mowa jest wypowiadana w tempie wolnym a modyfikacja wykonywana jest za pomocą metody R, gdzie α_o jest równa 1,33 lub 1,5. W przypadku metody B widoczna jest natomiast różnica w ocenach jakości mowy wypowiadanej przez mężczyznę w tempie szybkim i wolnym (dla $\alpha_o = 1{,}25$ i 1,5), a spowolniona mowa wypowiadana w tempie szybkim oceniana jest wyżej niż spowolniona mowa wypowiadana w tempie wolnym. Istotna statystycznie różnica ocen widoczna jest także w przypadku mowy wypowiadanej przez kobietę dla $\alpha_o = 1{,}5$. Wyższa ocena jakości mowy wypowiadanej w tempie szybkim jest o tyle istotna, iż opracowana metoda jest przeznaczona do modyfikacji mowy szybkiej. Dodatkowo jak zostało pokazane w rozdziale 4 rozprawy, właśnie w przypadku szybkiej mowy zastosowanie opracowanej metody modyfikacji tempa wypowiedzi pozwala osiągnąć istotną statystycznie poprawę jej rozumienia przez osoby z pogorszona rozdzielczością czasową słuchu. Ostatnią powtarzającą się, i potwierdzoną poprzez analizę statystyczną, różnicą dotyczącą oceny jakości występującą pomiędzy metodami B i R, jest wyższa ocena jakości mowy wypowiadanej w tempie szybkim przez mężczyznę w sytuacji gdy modyfikację wykonano za pomocą metody B.

Analizując trendy w ocenach naturalności mowy zmodyfikowanej widać istotne statystycznie różnie w ocenach naturalności mowy wypowiadanej w tempie wolnym i

[18] Dla tej metody różnice nie są istotne statystycznie jedynie w przypadku $\alpha_o = 1{,}25$.

szybkim (niezależnie od użytej wartości α_o) w przypadku, gdy mówcą jest mężczyzna. Zależność ta jest prawdziwa zarówno dla metody B jak i R[19], a naturalność jest wyższa gdy mowa wypowiadana jest w tempie szybkim (tab. 5.13–5.15). Należy też zwrócić uwagę na istotne statystycznie różnice w ocenach naturalności szybkiej mowy zmodyfikowanej za pomocą metody B w zależności od płci mówcy dla α_o = 1,33 i 1,5. W przypadku metody R różnice są istotne statystycznie dla α_o = 1,25 i 1,5. Różnice pomiędzy ocenami naturalności mowy zmodyfikowanej za pomocą analizowanych metod (B i R) są istotne statystycznie jedynie dla α_o = 1,25 i 1,33 i wolnej mowy męskiej oraz dla α_o = 1,33 i 1,5 i wolnej mowy żeńskiej.

Przeprowadzone eksperymenty pokazały, iż zarówno jakość jak i naturalność mowy zmodyfikowanej z wykorzystaniem opracowanej metody B jest większości przypadków wyższa niż w przypadku metody R. Dodatkowo średnie wartości obu ocen w przypadku tych parametrów są wysokie (MOS = 3,9 i DMOS = 4), co świadczy o wysokiej jakości oraz naturalności mowy zmodyfikowanej z wykorzystaniem opracowanej metody. W związku z powyższym druga teza rozprawy mówiąca, iż:

Opracowana metoda modyfikacji tempa mowy w czasie rzeczywistym, zapewnia wysoką jakość i naturalność subiektywnie odbieranej wypowiedzi.
została wykazana.

Opracowana metoda B została zaimplementowana w formie biblioteki programistycznej, co pozwoliło na jej wykorzystanie w trzech aplikacjach których opis znalazł się w rozprawie. Szczególnie użyteczna wydaje się aplikacja mobilna, która może być wykorzystana przez osoby z pogorszoną rozdzielczością czasową słuchu np. podczas rozmów lub zajęć szkolnych. Dodatkowo możliwe jest także wykorzystanie tej metody w wielu innych zastosowaniach np. do spowalniania mowy podczas: rozmowy telefonicznej, lub podczas oglądania telewizji.

Należy zauważyć, iż w przyszłości skuteczność opracowanych rozwiązań powinna zostać oceniona w warunkach rzeczywistych, np. podczas zajęć lekcyjnych. Pozwoliłoby to na znalezienie odpowiedzi na kilka pytań, które nie zostały postawione w rozprawie. Uwzględniając możliwość korzystania z mobilnej implementacji metody można by rozważyć następujące aspekty: czy słuchanie mowy spowolnionej nie skutkowałoby wprowadzeniem pewnego rodzaju bariery polegającej na odsunięciu słuchacza od

[19] Dla tej metody różnice nie są istotne statystycznie jedynie w przypadku α_o=1,5.

aktualnych wydarzeń? Jaki jest wpływ braku synchronizacji ust mówcy z dźwiękiem słyszanym w słuchawkach? Jak duże opóźnienie pojawi się w sytuacji, gdy algorytm będzie działał np. jedną godzinę lekcyjną bez przerwy? Czy korzystanie z algorytmu nie będzie powodowało dyskomfortu u użytkownika? Przestawione aspekty pokazują, iż pomimo tego, że w rozprawie wykazano użyteczność opracowanych metod, docelowa implementacja musi zostać szczegółowo przebadana, gdyż korzystny wpływ opracowanej metody może zostać obniżony przez czynniki zewnętrzne nieuwzględnione podczas eksperymentów prowadzonych w ramach rozprawy.

Podziękowania

Autor niniejszej rozprawy składa podziękowania Promotorowi pracy – prof. dr. hab. inż. Andrzejowi Czyżewskiemu za opiekę naukową.

Autor dziękuje również Kolegom z zespołu KSM biorącym udział w pracach realizowanych w ramach projektu TYPOSZEREG za ich pomoc w trakcie prac badawczych, i wspólne prace wdrożeniowe, zaś Najbliższym za dobre słowo i wyrozumiałość, przede wszystkim w końcowym etapie pisania pracy.

Autor pragnie także podziękować Pracownikom i Pacjentom Specjalistycznego Ośrodka Diagnozy i Rehabilitacji Dzieci i Młodzieży z Wadą Słuchu Polskiego Związku Głuchych w Gdańsku oraz Domu Opieki Kościoła Chrześcijan Baptystów „Nasz Dom" za możliwość przeprowadzenia badań z udziałem ich podopiecznych.

W okresie od 1 grudnia 2009 do 31 lipca 2010 praca nad rozprawą była finansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego. Projekt systemowy Województwa Pomorskiego pn. „InnoDoktorant – stypendia dla doktorantów, II edycja.

Od 1 sierpnia 2010 opisane eksperymenty oraz prace implementacyjne wykonano w ramach projektu No. POIG.01.03.01-22-017/08, o nazwie „Opracowanie typoszeregu komputerowych interfejsów multimodalnych oraz ich wdrożenie w zastosowaniach edukacyjnych, medycznych, w obronności i przemyśle". Projekt finansowany jest przez Europejski fundusz rozwoju regionalnego oraz przez budżet państwa.

8 Bibliografia

[1] A. Abel et al., *Maximising Audiovisual Correlation with Automatic Lip Tracking and Vowel Based Segmentation*. Berlin: Springer Verlag, 2009, vol. 5707/2009.

[2] A. Acero, "A Mixed-Excitation Frequency Domain Model for Time-Scale Pitch-Scale Modification of Speech," in *Proc. of the Int. Conf. on Spoken Language Processing*, 1998.

[3] American Speech-Language-Hearing Association, "(Central) auditory processing disorders - the role of the audiologist," 2005.

[4] R. Andre-Obrecht, "A new statistical approach for automatic speech segmentation," *IEEE Transactions on Acoustic Speech and Signal Processing*, vol. 36, no. 1, pp. 29-40, January 1988.

[5] B. Arons, "Techniques, Perception, and Applications of Time-Compressed Speech," in *American Voice I/O Society*, 1992, pp. 169 - 177.

[6] ASHA, "Central Auditory Processing Current Status of research and implication for clinical Practice," *Am. J. Audiology*, vol. 2, 1996.

[7] B. Atal and L. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," vol. 24, no. 3, pp. 201 - 212, 1976.

[8] M. Baig, S. Masud, and M. Awais, "Support Vector Machine based Voice Activity Detection," in *International Symposium on Intelligent Signal Processing and Communication Systems (ISPAC2006)*, Tottori, 2006.

[9] Ch. Bartels and J. Bilmes, "Use of Syllable Nuclei Locations to Improve ASR," in *Proc. IEEE Automatic Speech Recognition and Understanding (ASRU)*, Kyoto, 2007.

[10] S. Bech and N. Zacharov, *Perceptual Audio Evaluation – Theory, Method and Application*.: The Atrium, Southern Gate, Chichester, West Sussex, England: John Wiley & Sons, 2006.

[11] British Society of Audiology, "Auditory processing disorders," 2006.

[12] A. Camarena-Ibarrola and E. Chavez, "Using a new Discretization of the Fourier Transform to Discriminate Voiced From Unvoiced Speech," in *ENC 2006*, 2006, pp. 127-134.

[13] A. Chalamandaris, P. Tsiakoulis, and S. Karabetsos, "An Efficient and Robust Pitch Marking Algorithm on the Speech Waveform for TD-PSOLA," in *IEEE International Conference on Signal and Image Processing Applications*, 2009.

[14] S-H Chen, Sh-H Chen, and B. R. Chang, "A Support Vector Machine Based Voice Activity Detection Algorithm for AMR-WB Speech Codec System," in *Second International Conference on Innovative Computing, Information and Control, 2007. ICICIC '07*, Kumamoto, 2007.

[15] G. D. Chermak and F. E. Musiek, *Central Auditory Processing Disorders - New Perspectives*. San Diego: Singular Publishing Group Inc, 1997.

[16] Wai C. Chu and Khosrow Lashkari, "Energy-Based Nonuniform Time-Scale Compression of Audio Signals," *IEEE Transactions on Consumer Electronics*, vol. 49, no. 1, pp. 183-

187, June 2003.

[17] P. Combescure, A. Le Guyader, and A. Gilloire, "Quality evaluation of speech coded at 32 kbit/s by means of degradation category ratings," in *Proc. ICASSP 82 (International Conference on Acoustics,Speech and Signal Processing)*, vol. 2, Paryż, 1982.

[18] W. J. Conover, *Practical Nonparametric Statistics (3rd edition)*.: John Wiley & Sons, 1999.

[19] J. C. Cooper and G. A. Gates, "Hearing in the elderly: the Framingham cohort, 1983-1985. Part I. Basic audiometric test results," *Ear and Hearing*, vol. 12, no. 5, pp. 304-311, October 1991.

[20] D. Cournapeau, T. Kawahara, K. Mase, and T. Toriyama, "Voice Activity Detector Based on Enhanced Cumulant of LPC Residual and On-line EM Algorithm," in *INTERSPEECH 2006 - ICSLP*, 2006, pp. 1201-1204.

[21] M. Covell, M. Withgott, and M. Slaney, "Mach1: Nonuniform Time-Scale Modification of Speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1998.

[22] E Coyle et al., "Time-scale modification as a speech therapy tool for children with verbal apraxia," in *5th Intl. Conf. Disability, Virtual Reality and Assoc. Tech.*, Oxford, 2004, pp. 247-252.

[23] E. Coyle et al., "Time-scale modification as a speech therapy tool for children with verbal apraxia," in *5th Intl. Conf. Disability, Virtual Reality and Assoc. Tech.*, Oxford, 2004, pp. 247-252.

[24] C. C. Crandell and J. J. Smaldino, "Classroom Acoustics for Children With Normal Hearing and With Hearing Impairment," *Language, Speech, and Hearing Services in Schools*, vol. 31, pp. 362–370, October 2000.

[25] A. Czyżewski et al., "Typoszereg komputerowych interfejsów multimodalnych," in *w Materiałach Krajowego Sympozjum Telekomunikacji i Teleinformatyki KSTiT 2012*, Warszawa, 2012.

[26] L. Daben and F. Kubala, "A cross-channel modeling approach for automatic segmentation of conversational telephone speech," in *IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU '03*, 2003, pp. 333-338.

[27] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabicword recognition in continuously spoken sentences," *IEEE Transactions on Acoustic, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, August 1980.

[28] A. Davis, S. Nordholm, and R. Togneri, "Statistical Voice Activity Detection Using Low-Variance Spectrum Estimation and an Adaptive Threshold," *IEEE Transactions On Audio, Speech, And Language Processing*, vol. 14, no. 2, March 2006.

[29] N. H. De Jong, "Praat script to detect syllable nuclei and measure speech rate automatically," *Behavior research methods*, vol. 41, no. 2, pp. 385-390, 2009.

[30] M. Demol , K. Struyve, W. Verhelst, H. Paulussen, and P. Verhoeve, "Efficient non-uniform time-scaling of speech with WSOLA for CALL applications," in *InSTIL/ICALL 2004 Symp. Comput. Assisted Learn., NLP Speech Technol. Adv. Lang. Learn. Syst.*, Venice, 2004.

[31] M. Demol, W. Verhelst, K. Struye, and P. Verhoeve, "Efficient Non-Uniform Time-Scaling of Speech with WSOLA," in *Speech and Computers (SPECOM)*, 2005.

[32] N. Derakhshan, A. Akbari, and A. Ayatollahi, "Noise power spectrum estimation using constrained variance spectral smoothing and minima tracking," *Speech Communication*, vol. 51, no. 11, pp. 1098-1113, November 2009.

[33] M. Dolson, "The Phase Vocoder: a Tutorial," *Computer Music Journal*, vol. 10, no. 4, pp. 14-27, 1986.

[34] D. M. Domitz and R. L. Schow, "A new CAPD battery – multiple auditory processing assessment: factor analysis and comparisons with SCAN," *American Journal of Audiology*, vol. 9, pp. 101-111, October 2000.

[35] O. Donnellan, E. Jung, and E. Coyle, "Speech-Adaptive Time-Scale Modification for Computer Assisted Language-Learning," in *Third IEEE International Conference on Advanced Learning Technologies (ICALT'03)*, 2003, pp. 165-169.

[36] D. Dorran, Audio Time-Scale Modification, 2005, PhD Thesis.

[37] D. Dorran, E. Coyle, and R. Lawlor, "Audio time-scale modification using a hybrid time-frequency domain approach," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2005, pp. 279 - 282.

[38] D. Dorran, R. Lawlor, and E. Coyle, "A comparison of time-domain time-scale modification algorithms," in *120th Convention of Audio Engineering Society*, Paryż, 2006.

[39] D. Dorran, R. Lawlor, and E. Coyle, "A Hybrid Time—Frequency Domain Approach to Audio Time-Scale Modification," *Journal of the Audio Engineering Society*, vol. 54, no. 1/2, pp. 21-31, February 2006.

[40] T. Ebihar, Y. Ishikawa, Y. Kisuki, T. Sakamoto, and T. Has, "Speech Synthesis Software with Variable Speaking Rate and its Implementation on a 32-bit Microprocessor," in *19th IEEE International Conference on Consumer Electronics (ICCE 2000)*, Los Angeles, 2000.

[41] K. El-Maleh and P. Kabal, "Natural quality background noise coding using residual substitution," in *Proc. Eurospeech*, vol. 5, 1999, pp. 2359–2362.

[42] D. Enqing, L. Guizhong, Z. Yatong , and Z. Xiaodi, "Applying support vector machines to voice activity detection," in *6th International Conference on Signal Processing*, vol. 2, 2002, pp. 1124 – 1127.

[43] ETSI, Digital Cellular Telecommunications System (Phase 2+); Voice Activity Detector (VAD) for Adaptive Multi Rate (AMR) Speech Traffic Channels, GSM 06.94 v7.1.1, ETSI EN 301 708, 1999.

[44] ETSI, Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Advanced Front-End Feature Extraction Algorithm; Compression Algorithms, ETSI ES 202 050 Rec., 2002.

[45] G. Fairbanks, W. L. Everit, and R. P. Jaeger, "Method for time or frequency compression-expansion of speech," *Proffesional Group on Audio*, pp. 7-12, 1953.

[46] A. Fuente and B. McPherson, "Ośrodkowe procesy przetwarzania słuchowego: wprowadzenie i opis testów możliwych do zastosowania u pacjentów polskojęzycznych," *Audiologia i Foniatria*, vol. 6, no. 2, pp. 66-76, 2007.

[47] S. V. Gerven and F. Xie, "A comparative study of speech detection methods," in *Porc. Eurospeech*, vol. 3, 1997, pp. 1095-1098.

[48] J. M. Gorriz, J. Ramirez, J. C. Segura, and S. Hornill, "Voice Activity Detection using

Higher Order Statistics," *Lecture Notes in Computer Science*, vol. 3512, pp. 837-844, 2005.

[49] S. Gustafsson, R. Martin, P. Jax, and P. Vary, "A Psychoacoustic Approach to Combined Acoustic Echo Cancellation and Noise Reduction," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 245-256, July 2002.

[50] B. Hagerman, "Sentences for testing speech intelligibility in noise," *Scand. Sudiol.*, vol. 11, pp. 79-87, 1982.

[51] P. Hanna and M. Desainte-Catherine, "Time scale modification of noises using a spectral and statistical model," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03)*, vol. 6, 2003, pp. 181-184.

[52] D. Hao and Y. Tianren, "Endpoint detection based on mel-scale features and phoneme segmentation," in *7th International Conference on Signal Processing, ICSP '04*, vol. 1, 2004, pp. 667-670.

[53] J. A. Hartigan and M. A. Wong, "A K-Means Clustering Algorithm," *Applied Statistics*, vol. 28, no. 1, pp. 100-108, 1979.

[54] A. J. Hayte, "The Maximum Familywise Error Rate of Fisher's Least Significant Difference Test ," *Journal of the American Statistical Association*, vol. 81 , no. 396, pp. 1000–1004, 1986.

[55] K. K. Hong and S. L. Hwang, "Use of spectral autocorrelation in spectral envelope linear prediction for speech recognition," *IEEE Transactions on SAP*, vol. 7, no. 5, pp. 533-541, 1999.

[56] A. W. Howitt, "Automatic syllable detection for vowel landmarks," MIT, PhD Thesis 2000.

[57] J. D. Hoyt and H. Wechsler, "Detection of human speech in structured noise," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, Adelaide, 1994, pp. 237-240.

[58] Ch. H. Hsieh, T. Y. Feng, and P. Ch. Huang, "Energy-based VAD with grey magnitude spectral subtraction," *Speech Communication*, vol. 51, pp. 810–819, 2009.

[59] L. S. Huang and C.-H. Yang, "A novel apprpach to robust speech endpoint detection in car enviroment," in *ICASSP*, 2000, pp. 1751-1754.

[60] International Telecommunication Union (ITU-T), Serises P: Telephone Transmission Quality: Methods for objective and subjective assessment of quality, Sierpień 1996.

[61] ITU, Coding of speech at 8 kbit/s using conjugate structure algebraic code-excited linear-prediction (CS-ACELP) Annex B: A silence compression scheme, 1996.

[62] J. Jerger and F. Musiek, "Report of the Consensus Conference on the Diagnosis of Auditory Processing Disorders in School-Aged Children," *Journal of the American Association of Audiology*, vol. 11, no. 9, pp. 467-474, October 2000.

[63] Q.-H. Jo, J.-H. Chang, J. W. Shin, and N. S. Kim, "Statistical model-based voice activity detection using support vector machine," *IET Signal Processing*, vol. 3, no. 3, pp. 205-210, 2009.

[64] J.-C. Junqua, B. Mak, and B. Reaves, "A robust algorithm for word boundary detection in the presence of noise," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 3, July 1994.

[65] L. Karray and A. Martin, "Towards improving speech detection robustness for speech recognition in adverse environments," *Speech Communitation*, no. 3, pp. 261–276, 2003.

[66] R. W. Keith, "Standardization of the Time Compressed Sentence Test," *Journal of Educational Audiology*, pp. 15-20, 2002.

[67] T. Kinnunen, E. Chernenko, M. Tuononen, P. Fränti, and H. Li, "Voice Activity Detection Using MFCC Features and Support Vector Machine," in *Proc. Speech and Computer (SPECOM)*, vol. 2, 2007, pp. 556-561.

[68] B. W. Kleijn and K. K. Paliwal, *Speech Coding and Synthesis*. Nowy Jork: Elsevier Science Inc., 1995.

[69] A. M. Kondoz, *Digital Speech*. New York: John Wiley and Sons, 1999.

[70] M. Kos, "Noise Reduction Algorithm for Robust Speech Recognition Using Minimum Statistics Method and Neural Network VAD," in *14th International Workshop on Systems, Signals and Image Processing, 6th EURASIP Conference focused on Speech and Image Processing*, 2007, pp. 284-287.

[71] T. Kristjansson, S. Deligne, and P. Olsen, "Voicing Features for Robust Speech Detection," in *Eurospeech 2005*, 2005.

[76] A. Kupryjanow and A. Czyżewski, "A Method of Real-Time Non-Uniform Speech Stretching," in *ICETE 2011, CCIS 314*.: Springer, Heidelberg, 2012, pp. 362-373.

[72] A. Kupryjanow and A. Czyżewski, "A non-uniform real-time speech time-scale stretching method," in *SIGMAP 2011*, Sewilla, 2011.

[73] A. Kupryjanow and A. Czyżewski, "Improved method for real-time speech stretching," *Intelligent Decision Technologies (IDT) Journal*, 2012.

[77] A. Kupryjanow and A. Czyżewski, "Methods of Improving Speech Intelligibility for Listeners with Hearing Resolution Deficit," *Diagnostic Pathology*, vol. 7, no. 129, 2012.

[74] A. Kupryjanow and A. Czyżewski, "Porównanie metod modyfikacji czasu trwania sygnału mowy," in *14th International Symposium on Sound Engineering and Tonmeistering*, Wrocław, 2011.

[78] A. Kupryjanow and A. Czyżewski, "Realtime speech stretching for supporting hearing impaired schoolchildren," *Elektronika – Konstrukcje, Technologie, Zastosowania*, no. 3/2010, pp. 24-28, 2010.

[75] A. Kupryjanow and A. Czyżewski, "Real-time speech-rate modification experiments," in *Audio Engineering Society Convention Paper, preprint No. 8052*, Londyn, 2010.

[79] A. Kupryjanow and A. Czyżewski, "Sposób i system wspomagania rozumienia mowy," Zgłoszenie patentowe nr P.394202, Marzec 14, 2011.

[80] A. Kupryjanow and A. Czyżewski, "System wspomagania rozumienia mowy," Zgłoszenie patentowe nr P.396294, Wrzesień 12, 2011.

[81] A. Kupryjanow and A. Czyżewski, "Time-scale modification of speech signals for supporting hearing impaired schoolchildren," in *Proc. of International Conference NTAV/SPA, New Trends in Audio and Video, Signal Processing: Algorithms, Architectures*, Poznań, 2009, pp. 159 - 162.

[82] A. Kupryjanow and A. Czyżewski, "Zastosowanie spowalniania wypowiedzi w celu poprawy rozumienia mowy przez dzieci w szkole," in *XIII międzynarodowe sympozjum inżynierii*

i reżyserii dźwięku, Warszawa, Warszawa, 2009, pp. 81-87.

[83] A. Kupryjanow, P. Suchomski, P. Odya, and A. Czyżewski, "System Supporting Speech Perception in Special Educational Needs Schoolchildren," in *ICCHP 2012*, vol. 2, Linz, 2012, pp. 133-136.

[84] H. Kuwabara , "Acoustic and Perceptual Properties of Phonemes in Continuous Speech as a Function of Speaking Rate," in *Eurospeech 97*, 1997, pp. 1003-1006.

[85] J. Laroche, *Time and Pitch Scale Modification of Audio Signals*. Londyn: The Kluwer International Series in Engineering and Computer Science, 2002.

[86] J. Laroche and M. Dolson, "Improved Phase Vocoder for Time-Scale Modification of Audio," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 323-332, May 1999.

[87] B. Lawlor and A. D. Fagan, "A novel high quality efficient algorithm for time-scale modification of speech," in *Eurospeech, 6th Conference on Speech Communication and Technology, Budapest*, Budapest, 1999.

[88] H. H. Lee and C. K. Un, "A study of On-Off characteristics of conversational speech," *IEEE Transactions on Communications*, vol. 34, no. 6, pp. 630-637, June 1986.

[89] C. Y. Lin and J. S. Jang, "A two-phase pitch marking method for TD-PSOLA synthesis," in *INTERSPEECH*, 2004.

[90] Ch.-T. Lin, J.-Y. Lin, and G.-D. Wu, "A Robust Word Boundary Detection Algorithm for Variable Noise-Level Environment in Cars," *IEEE Transactions on Intelligent Transportation Systems*, vol. 3, no. 1, pp. 89-101, March 2002.

[91] Y. Li, T. Wang, H. Cui, and K. Tang, "Voice Activity Detection in Non-stationary Noise," in *IMACS Multiconference on "Computational Engineering in Systems Applications"(CESA)*, 2006.

[92] A. P. Lobo and P. C. Loizou, "Voiced/unvoiced speech discrimination in noise using Gabor atomic decomposition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003 (ICASSP '03)*, vol. 1, 2003, pp. 820-823.

[93] C. Lopes and F. Perdigao, "Speech event detection using SVM and NMD," in *Signal Processing and Its Applications, 2007. ISSPA 2007. 9th International Symposium on*, Sharjah, 2007, pp. 1-4.

[94] S. R. Mahadeva Prasanna, B. V. Sandeep Reddy, and P. Krishnamoorthy, "Vowel Onset Point Detection Using Source, Spectral Peaks, and Modulation Spectrum Energies," *IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING*, vol. 17, no. 4, pp. 556-565, May 2009.

[95] S. R. Mahadeva Prasanna and B. Yegnanarayana, "Detection of Vowel Onset Point Events using Excitation Information," in *INTERSPEECH 2005*, Lisbon, 2005, pp. 1133-1136.

[96] D. Malah, "Time-Domain Algorithms for Harmonic Bandwidth Reductiona and Time Scaling of Speech Signals," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, April 1979.

[97] M. Marzinzik and B. Kollmeier, "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 2, pp. 109-118, February 2002.

[98] M. P. Masquelier, "Management of auditory processing disorders," *Acta Otorhinolaryngol*

Belg., vol. 57, no. 4, pp. 301-310, 2003.

[99] J. W. Mauchly, "Significance Test for Sphericity of a Normal n-Variate Distribution," *The Annals of Mathematical Statistics*, vol. 11, no. 2, pp. 204–209, 1940.

[100] P. Mermelstein, "Automatic Segmentation of Speech into Syllabic Units," *The Journal of the Acoustical Society of America*, vol. 58, pp. 880-883, 1975.

[101] N. Mirghafori, E. Fosler, and N. Morgan, "Towards Robustness to Fast Speech in ASR," in *Proc. ICASSP'96*, 1996, pp. 335-338.

[102] M. H. Moattar and M. M. Homayo, "A Weighted Feature Voting Approach for Robust and Real-Time Voice Activity Detection," *ETRI Journal*, vol. 33, no. 1, pp. 99-109, February 2011.

[103] M. H. Moattar and M. M. Homayounpour, "A simple but efficient real-time voice activity detection algorithm," in *17th European Signal Processing Conference (EUSIPCO 2009)*, Glasgow, 2009.

[104] M. Moattar, M. Homayounpour, and N. Kalantari, "A new approach for robust realtime voice activity detection using spectral pattern," in *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2010.

[105] S. Molau, M. Pitz, R. Schluter, and H. Ney, "Computing Mel-Frequency Cepstral Coefficients on the Power Spectrum," in *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. (ICASSP '01)*, vol. 1, Salt Lake City, 2001, pp. 73 - 76.

[106] N. Morgan and E. Fosler-Lussier, "Combining Multiple Estimators of Speaking Rate," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1998)*, Washington, 1998, pp. 729-732.

[107] N. Morgan, E. Fosler, and N. Mirghafori, "Speech recognition using online estimation of speaking rate," in *Eurospeech*, vol. 4, Rhodes, 1997, pp. 2079-2082.

[108] E. Moulines and F. Charpentier, "Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones," *Speech Communication*, vol. 9, no. 5/6, pp. 453-467, 1990.

[109] E. Moulines and J. Laroche, "Non-parametric techniques for pitch-scale and time-scale modification of speech," *Speech Communication*, vol. 16, no. 2, pp. 175-205, 1995.

[110] I. Mporas, A. Lazaridis, T. Ganchev, and N. Fakotakis, "Using Hybrid HMM-based Speech Segmentation to Improve Synthetic Speech Quality," in *13th Panhellenic Conference on Informatics*, 2009, pp. 118 - 122.

[111] N. B. Muluk, F. Yalcınkaya, and R. W. Keith, "Random gap detection test and random gap detection test-expanded: Results in children with previous language delay in early childhood," *Auris Nasus Larynx*, pp. 6–13, 2011.

[112] R. Muralishankar, R. V. Prasad, S. Vijay, and H. N. Shanka, "Order Statistics for Voice Activity Detection in VoIP," in *IEEE International Conference on Communications (ICC)*, 2010, pp. 1-6.

[113] F. E. Musiek and G. D. Chermak, *Handbool of (Central) Auditory Processing Disorder*.: Plural Publishing, 2007, vol. 1.

[114] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika, "The OGI multilanguage telephone speech corpus," in *ICSLP*, 1992, pp. 895-898.

[115] S. S. Nagarajan et al., "Speech Modifications Algorithms Used for Training Language Learning-Impaired Children," *IEEE Transactions on Rehabilitation Engineering*, vol. 6, no. 3, pp. 257-268, September 1998.

[116] A. Nakamura, N. Seiyama, A. Imai, T. Takagi, and E.i Miyasak, "A New Approach to Compensate Degeneration of Speech Intelligibility for Elderly Listeners," *IEEE Transactions on Broadcasting*, vol. 42, no. 3, pp. 285-293, November 1996.

[117] A. Nakamura, N. Seiyama, T. Takagi, and E. Miyasaka, "Real Time Speech Rate Converting System For Elderly People," in *IEEE Int. Conf. Acoust., Speech, Signal Proc (ICASSP)*, vol. 11, Adelaide, 1994, pp. 225-228.

[118] A. Nakamura, N. Seiyama, T. Takagi, and E. Miyasaka, "Real Time Speech Rate Converting System For Elderly People," in *IEEE Int. Conf. Acoust., Speech, Signal Proc (ICASSP)*, vol. 11, Adelaide, 1994, pp. 225-228.

[119] S. Narayanan and D. Wang, "Speech rate estimation via temporal correlation and selected sub-band correlation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005, pp. 413 - 416.

[120] J. Natvig, S. Hansen, and J. De Brito, "Speech processing in the paneuropean digital mobile radio system (GSM) – system overview," in *Proc. IEEE Global Telecommunications Conference (IEEE GLOBECOM 1989)*, 1989, pp. 1060–1064.

[121] Y. Nejime et al., "A portable digital speech-rate converter for elderly hearing-impaired listeners," in *16th Annual International Conference of the IEEE*, 1994, pp. 271-271.

[122] Y. Nejime, T. Aritsuka, T. Imamura, T. Ifukube, and J. Matsushima, "A portable digital speech-rate converter for hearing impairment," *IEEE Trans. Rehabil. Eng*, vol. 4, no. 2, pp. 73-83, 1996.

[123] Y Nejime and B. C. J. Moore, "Evaluation of the effect of speech-rate slowing on speech intelligibility in noise using a simulation of cochlear hearing loss," *Journal Acoustical Society of America*, vol. 103, no. 1, pp. 572-576, 1998.

[124] E. Nemer, R. Goubran, and S. Mahmoud, "Robust vad using hos in the lpc residual domain," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 217-231, March 2001.

[125] S. Ohno and H. Fujisaki, "A method for quantitative analysis of the local speech rate," in *In Proceedings of Eurospeech'1995*, vol. 1, Madrid, 1995, pp. 421 - 424.

[126] A. Ouzounov, "A Robust Feature for Speech Detection," *Cybernetics and Information Technologies*, vol. 4, no. 2, pp. 3-14, 2004.

[127] A. Ouzounov, "An Experimental Comparative Study of Three Robust Features for Speech Detection," *Cybernetics And Information Technologies*, vol. 5, no. 2, 2005.

[128] D. Owens, P. E. Campbell, A. Liddell, C. Deplacido, and M. Wolters, "Random Gap Detection Test : A Useful Measure of Auditory Ageing ?," in *Proc. European Federation on Audiology*, Heidelberg, 2007.

[129] E. Ozimek, D. Kutzner, and P. Libiszewski, "Zrozumiałość zdaniowa mowy przyśpieszonej," *Biuletyn Polskiego Stowarzyszenia Protetyków Słuchu*, vol. 42, no. 2/11, pp. 8-9, 2011.

[130] E. Ozimek, D. Kutzner, P. Libiszewski, A. Warzybok, and J. Kociński, "The new polish tests fo speech intelligibility measurements ," *Int. J. Audiol.*, vol. 49, no. 6, pp. 444-454, June 2010.

[131] E. Ozimek, P. Libiszewski, and D. Kutzner, "Polski Pediatryczny Test Zdaniowy do pomiarów zrozumiałości mowy prezentowanej na tle szumu," *Biuletyn Polskiego Stowarzyszenia Protetyków Słuchu*, vol. 40, no. 4/10, pp. 9-13, 2010.

[132] R. Padmanabhan, S. H. K. Parthasarathi, and M. A. Hema , "A pattern recognition approach to VAD using modified group delay," in *Proc. of 14th National conference on Communications*, Bombay, 2008, pp. 432-437.

[133] D. S. Pallet et al., "WSJ-CSR Benchmark Test Results," in *AREAS Spoken Language System Technology Workshop*, New Jersey, 1994.

[134] S.H.K. Parthasarathi , R. Padmanabhan, and M. A. Hema, "Voice Activity Detection using Group Delay Processing on Buffered Short-term," in *National Conference on Communications: NCC-2007*, 2007.

[135] F. Pellegrino and R. Andre-Obrecht, "An unsupervised approach to language identification," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 1999. Proceedings*, vol. 2, Toulouse, 1999, pp. 833 - 836.

[136] F. Pellegrino and R. Andre-Obrecht, "Automatic language identification: an alternative approach to phonetic modeling," *Signal Processing*, vol. 80, pp. 1231-1244, 2000.

[137] F. Pellegrino and R. Andre-Obrecht, "From vocalic detection to automatic emergence of vowel systems," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-97*, vol. 3, 1997, pp. 1651-1654.

[138] F. Pellegrino, J.-H. Chauchat, R. Rakotomalala, and J. Farinas, "Can automatically extracted rhythmic units discriminate among languages?," in *In Proc. of International Conference on Speech Prosody*, 2002, pp. 563-566.

[139] F. Pellegrino, J. Farinas, and J.-L. Rouas, "Automatic estimation of speaking rate in multilingual spontaneous speech," in *In Proc. of International Conference on Speech Prosody*, 2004, pp. 517–520.

[140] D. Peng, "An adaptive soft voice activity detector for automatic speech recognition system," in *Information, Communications and Signal Processing (ICICS) 2011 8th International Conference on*, Singapur, 2011, pp. 1-5.

[141] T. Pfau and G. Ruske, "Estimating the speaking rate by vowel detections," in *ICASSP 98*, Tagungsband, Seattle, 1998, pp. 945-948.

[142] H. R. Pfitzinger, "Local speaking rate as a combination of syllable and phone rate," in *Proceeding of ICSLP 1998*, 1998.

[143] H. R. Pfitzinger, "Two approaches to speech rate estimation," in *Proceedings of SST'96*, Adelaide, 1996, pp. 421-426.

[144] Phonak Hearing Systems, "EduLink: Improves speech understanding in noisy classrooms," *Field study news*, pp. 1-2, May 2004.

[145] R. Plomp and A. M. Mimpen, "Improving the Reliability of Testing the Speech Reception Threshold for Sentences," *Audiology*, vol. 18, no. 1, pp. 43-52, 1979.

[146] P. Pollak and J. Rajnoha, "Long Recording Segmentation Based on Simple Power Voice Activity Detection with Adaptive Threshold and Post-Processing," in *SPECOM'2009*, St. Petersburg, 2009.

[147] M. P. Portnoff, "Time-Scale Modification of Speech Based on Short-Time Fourier Analysis," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-29, no. 3,

pp. 374-390, June 1981.

[148] V. K. Prasad, T. Nagarajan, and H. A. Murthy, "Automatic segmentation of continuous speech using minimum phase group delay functions," *Speech Communication*, vol. 43, no. 3-4, pp. 429-446, 2004.

[149] V. R. Prasad et al., "Comparison of voice activity detection algorithms for VoIP," in *Seventh International Symposium on Computers and Communications, Proceedings. ISCC 2002*, 2002, pp. 530-535.

[150] L. R. Rabiner and M. R. Sambur, "An Algorithm for Determining the Enpoints of Isolated Utterances," *The Bell System Technical Journal*, vol. 54, no. 2, pp. 297-315, February 1975.

[151] L. R. Rabiner and M. Sambur, "Voiced-unvoiced-silence detection using the Itakura LPC distance measure," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '77*, New Jersey, 1977, pp. 323 - 326.

[152] L. R. Rabiner and R. W. Schafer, *Introduction to Digital Speech Processing*., 2007.

[153] J. Ramirez, J. C. Segura, J. M. Gorriz, and L. Garcia, "Improved Voice Activity Detection Using Contextual Multiple Hypothesis Testing for Robust Speech Recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 8, pp. 2177-2189, November 2007.

[154] F. Ringeval and M. Chetouani, "A vowel based approach for acted emotion recognition," in *INTERSPEECH'2008*, 2008, pp. 2763-2766.

[155] F. Ringeval and M. Chetouani, "Exploiting a Vowel Based Approach for Acted Emotion Recognition," in *In COST 2102 Workshop*, Patras, 2007, pp. 243-254.

[156] J.-L. Rouas, J. Farinas, F. Pellegrino, and R. Andre-Obrecht, "Rhythmic unit extraction and modelling for automatic language identification," *Speech Communication*, vol. 47, pp. 436-456, 2005.

[157] S. Roucos and A. Wilgus, "High quality time-scale modification for speech. Acoustics, Speech, and Signal Processing," in *IEEE International Conference on ICASSP '85*, 1985, pp. 493-496.

[158] K. Sakhnov, E. Verteletskaya, and B. Simak, "Approach for Energy-Based Voice Detector with Adaptive Scaling Factor," *IAENG International Journal of Computer Science*, vol. 36, no. 4, November 2009.

[159] A. Sangwan et al., "Second and Third Order Adaptable Threshold for VAD in VoIP," in *6th International Conference on Signal Processing. ICSP'02*, vol. 2, Beijing, 2002, pp. 1693-1696.

[160] P.-A. Savard, P. Gournay, and R. Lefebvre, "Hybrid Time-Scale Modification of Audio," in *122 th Convention of the Audio Engineering Society*, vol. Paper no. 7133, Vienna, 2007.

[161] M. H. Savoji, "A robust algorithm for accurate end pointing of speech," *Speech Communication*, pp. 45–60, 1989.

[162] A. Senderski, "Diagnostyka centralnych zaburzeń przetwarzania słuchowego. Algorytm postępowania diagnostycznego," Warszawa, 2002.

[163] C. Shahnaz, W.,P. Zhu, and M. O. Ahmad, "A Bifeature Voiced / Unvoiced Discrimination Algorithm for Speech Signals in the Presense of Noise," *Computer Engineering*, pp. 89-92, 2007.

[164] C. E. Shannon, "A Mathematical Theory of Communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, July 1948.

[165] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples) ," *Biometrika*, vol. 52 , no. 3-4, pp. 591–611, 1965.

[166] M. Sharma and R. Mammone, ""Blind" speech segmentation: automatic segmentation of speech without linguistic knowledge," in *Fourth International Conference on Spoken Language, ICSLP 96*, vol. 2, Philadelphia, 1996, pp. 1237-1240.

[167] J. Shen, J. Hung, and L. Lee, "Robust Entropy-based Endpoint Detection for Speech Recognition in Noisy Environment," in *International Conference on Spoken Language Processing*, Sydney, 1998.

[168] M. A. Siegler and R. M. Stern, "On the effects of speech rate in large vocabulary speech recognition systems," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 1, Detroit, 1995, pp. 612-615.

[169] S. Skorik and F. Berthommier, "On a Cepstrum-Based Speech Detectro Robust to White Noise," in *CoRR*, vol. cs.CL/0010014, 2000.

[170] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1-3, 1999.

[171] M. Spitzer, *Jak uczy się mózg (tytuł oryginalny: Lernen. Gehirnforschung und die Schule des Lebens)*, Jagodzińska Maria, Ed. Warszawa: PWN, 2007, Tłumaczenie z języka niemieckiego.

[172] P. Tallal et al., "Language Comprehension in Language-Learning Impaired Children Improved with Acoustically Modified Speech," *Science*, vol. 271, pp. 81-84, January 1996.

[173] TIA, TDMA minimum performance standards for discontinuous transmission operation of mobile stations, 1998.

[174] Z. Tuske, P. Mihajlik, Z. Tobler, and T. Fegyo, "Robust Voice Activity Detection Based on the Entropy of Noise-Suppressed Spectrum," in *INTERSPEECH*, 2005, pp. 245-248.

[175] R. M. Uchanski, A. E. Geers, and A. Protopapas, "Intelligibility of Modified Speech for Young Listeners With Normal and Impaired Hearing," *Journal of Speech, Language, and Hearing Research*, vol. 45, pp. 1027-1038, October 2002.

[176] V. N. Vapnik, *Estimation of Dependencies Based on Empirical Data*. New York: Springer Verlag New York, 1982.

[177] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag New York, 1995.

[178] J. P. Verhasselt and J.-P. Martens, "A Fast and Reliable Rate of Speech Detector," in *Proceedings of the Fourth International Conference on Spoken Language Processing*, vol. 4, 1996, pp. 2258-2261.

[179] W. Verhelst, "Overlap-add methods for time-scaling of speech," *Speech Communication*, vol. 30, pp. 207-221, 2000.

[180] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-93*, 1993.

[181] N. J. Versfeld and W. A. Dreschler, "The relationship between the intelligibility of time-compressed speech and speech in noise in young and elderly listeners," *J. Acoust. Soc. Am.*, vol. 111, no. 1, pp. 401-408, January 2002.

[182] M. Vondrasek and P. Pollak, "Methods for Speech SNR Estimation: Evaluation Tool and Analysis of VAD Dependency," *Journal: Radioengineering*, vol. 14, pp. 6-11, 2005.

[183] K.-Ch. Wang, "Robust Voice Activity Detection Based on Discrete Wavelet Transform," in *Proceedings of the 20th Conference on Computational Linguistics and Speech Processing*, 2008.

[184] D. Wang and S. S. Narayanan, "Robust speech rate estimation for spontaneous speech," *IEEE Transactions on audio, speech and language processing*, vol. 15, no. 8, October 2007.

[185] J. Weihing, "FM systems as a treatment for CAPD," *Hearing Journal*, vol. 58, no. 10, p. 74, October 2005.

[186] S. A. J. Wood, "What happens to vowels and consonants when we speak faster?," in *The Stockhom-Lund Phonetics Symposium*, Lund, 1973, pp. 8-40.

[187] B. F. Wu and K.-Ch. Wang, "An Adaptive Band-Partitioning Spectral Entropy Based Speech Detection In Realistic Noisy Environments," in *INTERSPEECH 2004 ICSLP*, vol. 2, 2004, pp. 957-960.

[188] B. F. Wu and K. Ch. Wang , "Noise Spectrum Estimation with Entropy-Based VAD in Non-stationary Environments," *IEICE Trans. Fundamentals*, vol. E89–A, no. 2, pp. 479-485, February 2006.

[189] J. Wu and X. L. Zhang, "An efficient voice activity detection algorithm by combining statistical model and energy detection," in *EURASIP Journal on Advances in Signal Processing*, 2011.

[190] Z. Xie and P. Niyogi, "Robust acoustic-based syllable detection," in *Proceedings of INTERSPEECH'2006*, 2006.

[191] F. Yalcınkaya, N. B. Muluk, A. Atas, and R. W. Keith, "Random Gap Detection Test and Random Gap Detection Test-Expanded results in children with auditory neuropathy," *International Journal of Pediatric Otorhinolaryngology*, vol. 73, pp. 1558–1563, 2009.

[192] X. Yang, B. Tan, J. Ding, J. Zhang, and J. Gong, "Comparative Study on Voice Activity Detection Algorithm," in *International Conference on Electrical and Control Engineering*, Wuhan, 2010, pp. 599 - 602.

[193] P. Yelamos, J. Ramirez, J. M. Gorriz, C. G. Puntonet, and J. C. Segura, "Speech Event Detection Using Support Vector Machines," in *International Conference on Computational Science 2006*, 2006, pp. 356-363.

[194] G. S. Ying, C. D. Mitchell, and L. H. Jamieson, "Endpoint Detection of Isolated Utterances Based on a Modified Teager Energy Measurement," in *ICASSP*, 1993, pp. 732-735.

[195] I. C. Yoo and D. Yook, "Robust Voice Activity Detection Using the Spectral Peaks of Vowel Sounds," *ETRI Journal*, vol. 4, pp. 451-453, August 2009.

[196] M. L. Young, "Recognizing and Treating Children with Central Auditory Processing Disorders," 1996.

[197] S. Zhang, "An energy-based adaptive voice detection approach," in *ICSP2006 Proceedings*, 2006.

[198] Y. Zhang and James R. Glass, "Speech rhytm guided syllable nuclei detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 3797-3800.

[199] J. Zheng, H. Franco, and A. Stolcke, "Modeling word-level rate-of-speech variation in large vocabulary conversational speech recognition," *Speech Comunnication*, vol. 41, pp. 273-285, 2003.

[200] J. Zheng, H. Franco, and A. Stolcke, "Rate of speech modeling for large vocabulary conversational speech recognition," , 2000.

[201] J. Zheng, H. Franco, F. Weng, A. Sankar, and H. Bratt, "Word-level rate-of-speech modeling using rate-specificphones and pronunciations," in *Proc. IEEE Int. Conf.Acoust. Speech Signal Processing*, vol. 3, Istanbul, 2000, pp. 1775–1778.

[202] B. Ziółko, S. Manandhar, and R. C. Wilson, "Phoneme segmentation of speech," in *The 18th International Conference on Pattern Recognition (ICPR'06)*, Hong Kong, 2006, pp. 282 - 285.

[203] U. Zolzer, *DAFX Digital Audio Effects*.: John Wiley & Sons, 2005.

9 Załączniki

9.1 Załącznik nr 1

Wypowiedź 1

Halo, halo witam państwa bardzo serdecznie z zatłoczonej ulicy Żwirka i Muchomorka, gdzie właśnie już dziś, za klika, może kilkanaście minut będziemy świadkami otwarcia nowego supermarketu. Jest słonecznie. Bezchmurne niebo wydaje się być brunatne, ale to zapewne wina moich nowych okularów fotochromowych.

Wypowiedź 2

Witam państwa bardzo serdecznie ze skoczni. Jest minus trzynaście stopni, a nasz Adaś już jest na belce. Już poprawił wąsy. Nowiutki kombinezon, nowiutkie gogle – ależ ta guma ściśnięta. Spoglądam gdzie są niemieccy zawodnicy…

Wypowiedź 3

Rozumiem, że chodzi panu o pewne trudności, z którymi miał do czynienia pewien pański poprzednik, który odbywa pewną karę, w pewnym miejscu odosobnienia, i pewnie chciałby pan tego uniknąć?

Wypowiedź 4

Witam bardzo serdecznie wszystkich zebranych na zlocie absolwentów naszego liceum. Jako dyrektor tej placówki mam zaszczyt zaprosić do nas naszego absolwenta. Przyjechał do nas z bardzo daleka. Przed wami Jan Kowalski. Brawa!

Wypowiedź 5

A teraz się obaj skupcie, bo ja wciągnę was w mój magiczny plan. Zsynchronizujcie zegarki. Jest yyy 20:55. Punktualnie o 21 zacznę wyrzucać z boiska zawodników drużyny przyjezdnej. Najpierw bramkarz, później czterech obrońców. Od tej chwili ta akcja jest tajna.

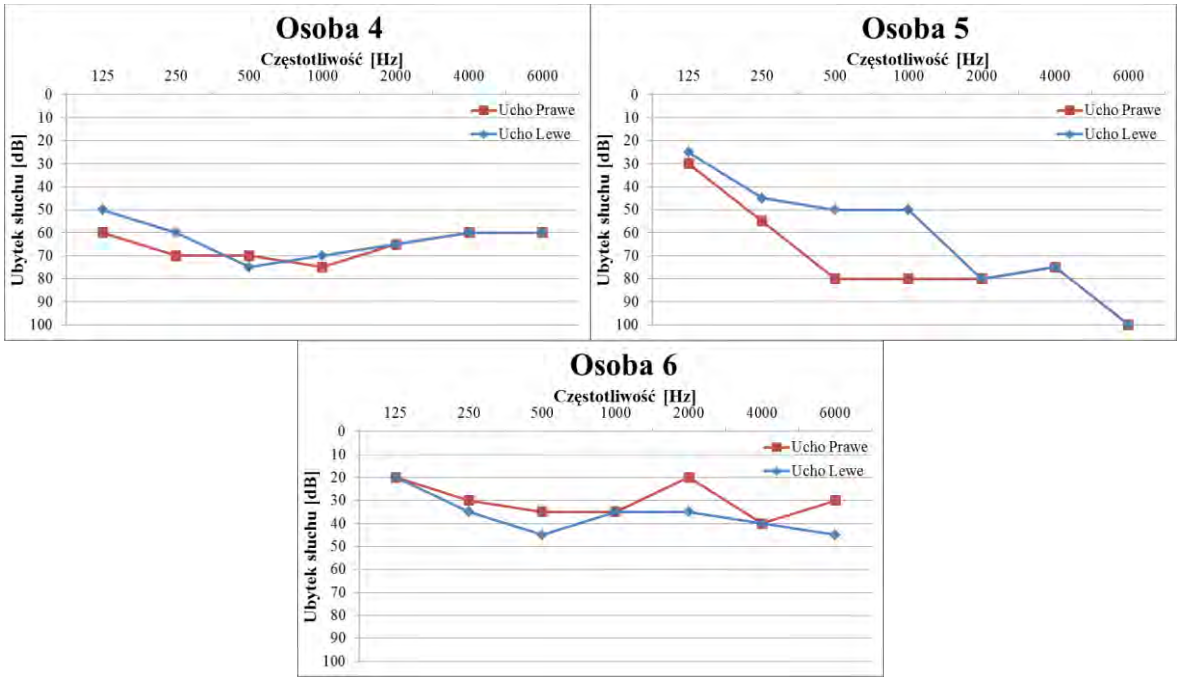
9.2 Załącznik nr 2

Tab. 9.1 Zbiór przykładowy zdań testu PTM oraz numery plików dźwiękowych przypisane do zdań wygenerowanych dla mowy wypowiadanej w tempie wolnym.

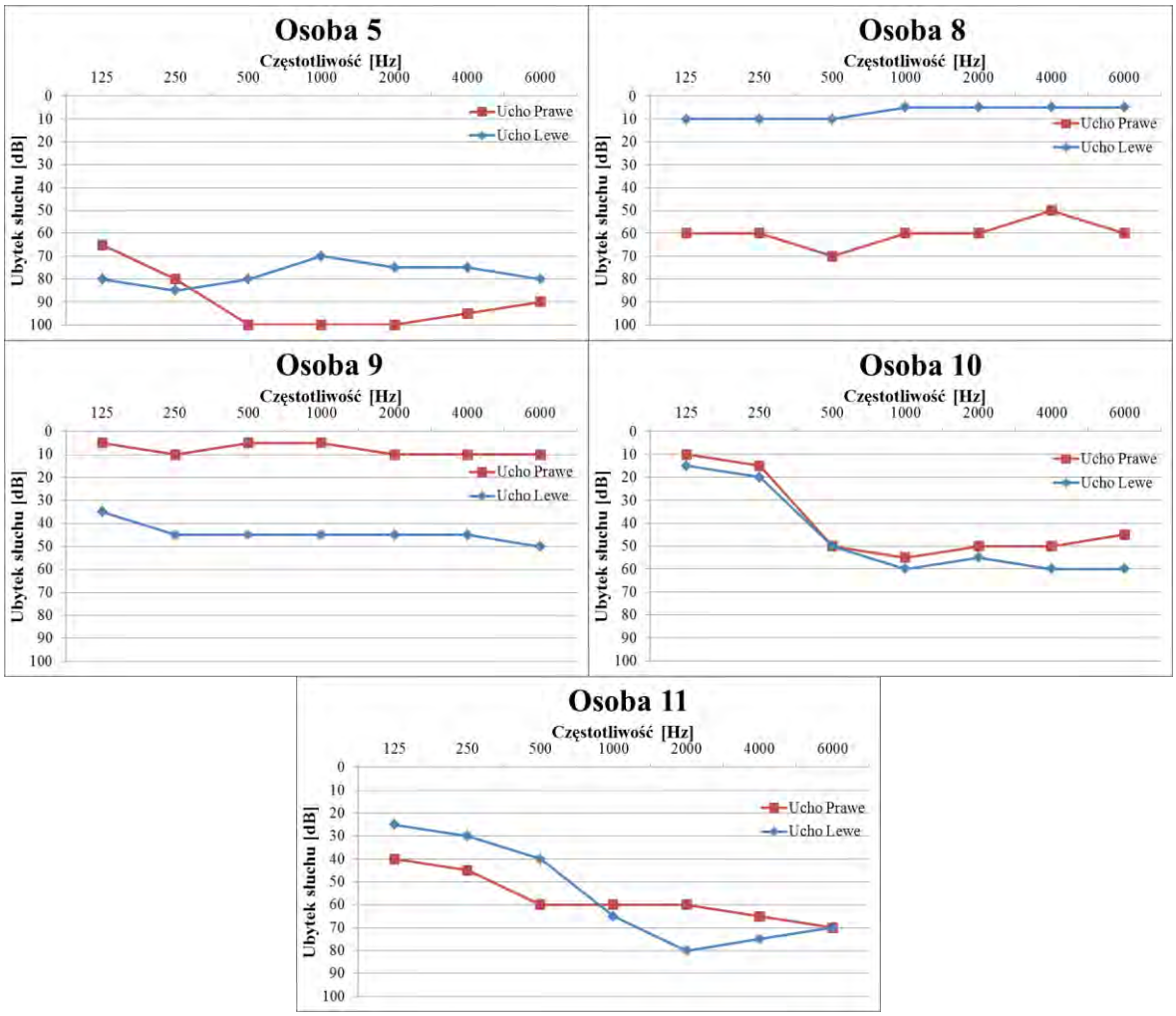
Numer zdania	Nazwa pliku	Treść zdania
1	ptm_01.wav	*Julia bierze siedem czarnych klocków*
2	ptm_02.wav	*Anna sprzeda sto drogich koszy*
3	ptm_03.wav	*Maria ma sześć nowych dzwonów*
4	ptm_04.wav	*Paweł daje pięć starych stołów*
5	ptm_05.wav	*Ewa kupi kilka pięknych okien*
6	ptm_06.wav	*Maciej nosi dużo białych soków*
7	ptm_07.wav	*Adam robi dziewięć tanich gazet*
8	ptm_08.wav	*Tomasz woli wiele żółtych toreb*
9	ptm_09.wav	*Zofia widzi tysiąc dobrych opon*
10	ptm_10.wav	*Michał wygra osiem dziwnych piłek*

Tab. 9.2 Zbiór przykładowy zdań testu PPTM oraz numery plików dźwiękowych przypisane do zdań wygenerowanych dla mowy wypowiadanej w tempie wolnym.

Numer zdania	Numer zbioru	Nazwa pliku	Treść zdania
1	1	pptm_1_1.wav	*Chłopiec otwiera samochód*
2		pptm_1_2.wav	*Chłopiec maluje szafę*
3		pptm_1_3.wav	*Babcia ogląda samochód*
4		pptm_1_4.wav	*Strażak otwiera dom*
5	2	pptm_2_1.wav	*Zając podlewa ogórki*
6		pptm_2_2.wav	*Miś podlewa ogórki*
7		pptm_2_3.wav	*Kot je wisienki*
8		pptm_2_4.wav	*Miś je ogórki*
9	3	pptm_3_1.wav	*Mama wiesza sweter*
10		pptm_3_2.wav	*Dziadek wiesza spodnie*
11		pptm_3_3.wav	*Tata pierze płaszcz*
12		pptm_3_4.wav	*Tata pierze koszulę*
13	4	pptm_4_1.wav	*Królewna nalewa mleko*
14		pptm_4_2.wav	*Królewna ma wodę*
15		pptm_4_3.wav	*Król miesza mleko*
16		pptm_4_4.wav	*Żołnierz nalewa herbatę*



Rys. 9.1 Audiogramy dzieci głuchych z pogorszoną rozdzielczością czasową słuchu.



Rys. 9.2 Audiogramy dzieci głuchych z normalną rozdzielczością czasową słuchu.

9.3 Załącznik nr 3

Szczegółowe wyniki analizy statystycznej.

Tab. 9.3 Wyniki testu Shapiro-Wilka dla dzieci głuchych z pogorszoną rozdzielczością czasową słuchu ($TCT_{50} < 5{,}71$ samogłosek/s).

tempo mowy	rodzaj modyfikacji	Wartość statystyki	Ilość stopni swobody	Istotność (p)
$ROS^{PP}_{szybkie}$	brak	0,78	5	**0,04**
	metoda A	0,81	5	0,08
	metoda B	0,94	5	0,73
	metoda C	0,90	5	0,41
$ROS^{PP}_{średnie}$	brak	0,97	5	0,93
	metoda A	0,91	5	0,46
	metoda B	0,92	5	0,50
	metoda C	0,96	5	0,84

Tab. 9.4 Wyniki testu Shapiro-Wilka dla dzieci głuchych z normalną rozdzielczością czasową słuchu ($TCT_{50} \geq 5{,}71$ samogłosek/s).

tempo mowy	rodzaj modyfikacji	Wartość statystyki	Ilość stopni swobody	Istotność (p)
$ROS^{PP}_{szybkie}$	brak	0,94	11	0,57
	metoda A	0,96	11	0,80
	metoda B	0,95	11	0,60
	metoda C	0,93	11	0,43
$ROS^{PP}_{średnie}$	brak	0,92	11	0,32
	metoda A	0,82	11	**0,02**
	metoda B	0,89	11	0,15
	metoda C	0,81	11	**0,01**

Tab. 9.5 Wyniki testu Mauchly'ego dla dzieci głuchych z pogorszoną rozdzielczością czasową słuchu ($TCT_{50} < 5{,}71$ samogłosek/s)

tempo mowy	Wartość statystyki χ^2	Ilość stopni swobody	Istotność (p)
$ROS^{PP}_{średnie}$	11,38	5	0,051

Tab. 9.6 Wyniki testu Mauchly'ego dla dzieci głuchych z normalną rozdzielczością czasową słuchu ($TCT50 \geq 5{,}71$ samogłosek/s)

tempo mowy	Wartość statystyki χ^2	Ilość stopni swobody	Istotność (p)
$ROS^{PP}_{szybkie}$	4,55	5	0,49

Tab. 9.7 Wyniki testu Shapiro-Wilka dla osób starszych z pogorszoną rozdzielczością czasową słuchu (TCT50 < 3,99 samogłosek/s).

tempo mowy	rodzaj modyfikacji	Wartość statystyki	Ilość stopni swobody	Istotność (p)
$ROS^{PP}_{szybkie}$	brak	0,90	5	0,39
	metoda A	0,89	5	0,32
	metoda B	0,96	5	0,82
	metoda C	0,96	5	0,82
$ROS^{PP}_{średnie}$	brak	0,95	5	0,77
	metoda A	0,91	5	0,49
	metoda B	0,94	5	0,67
	metoda C	0,86	5	0,20

Tab. 9.8 Wyniki testu Mauchly'ego dla osób starszych z pogorszoną rozdzielczością czasową słuchu (TCT50 < 3,99 samogłosek/s).

tempo mowy	Wartość statystyki χ^2	Ilość stopni swobody	Istotność (p)
$ROS^{PP}_{szybkie}$	6,51	5	0,27
$ROS^{PP}_{średnie}$	2,84	5	0,732

Tab. 9.9 Wyniki testu Mauchly'ego dla osób starszych z normalną rozdzielczością czasową słuchu (TCT50 ≥ 3,99 samogłosek/s).

tempo mowy	Wartość statystyki χ^2	Ilość stopni swobody	Istotność (p)
$ROS^{PP}_{szybkie}$	4,07	5	0,54
$ROS^{PP}_{średnie}$	7,28	5	0,20

Tab. 9.10 Wyniki testu Shapiro-Wilka dla osób starszych z pogorszoną rozdzielczością czasową słuchu (TCT50 ≥ 3,99 samogłosek/s).

tempo mowy	rodzaj modyfikacji	Wartość statystyki	Ilość stopni swobody	Istotność (p)
$ROS^{PP}_{szybkie}$	brak	0,96	11	0,80
	metoda A	0,87	11	0,07
	metoda B	0,90	11	0,22
	metoda C	0,89	11	0,17
$ROS^{PP}_{średnie}$	brak	0,86	11	0,06
	metoda A	0,94	11	0,58
	metoda B	0,93	11	0,47
	metoda C	0,89	11	0,13

9.4 Załącznik nr 4

Tab. 9.11 Poprawa rozumienia mowy zmodyfikowanej dla dzieci głuchych z pogorszoną rozdzielczością czasową słuchu (TCT_{50} < 5,71 samogłosek/s).

	$ROS^{PP}_{szybkie}$			$ROS^{PP}_{średnie}$		
L.p.	**metoda A**	**metoda B**	**metoda C**	**metoda A**	**metoda B**	**metoda C**
1 (IS)	16,67	33,33	12,50	3,33	10	–30,00
2 (IS)	16,67	23,33	0	3,33	0	–3,33
3 (AS)	43,33	26,67	30,00	16,67	16,67	13,33
4 (AS)	41,67	25,00	12,50	0	0	0
5 (AS)	43,33	50,00	33,33	6,67	10,00	–6,67
6 (AS)	4,17	8,33	16,67	–6,67	–16,67	–6,67
wartość średnia	27,64	27,78	17,50	3,89	3,33	–5,56
odchylenie standardowe	17,21	13,65	12,36	7,72	11,74	14,09

Tab. 9.12 Poprawa rozumienia mowy zmodyfikowanej dla dzieci głuchych z normalną rozdzielczością czasową słuchu (TCT50 ≥ 5,71 samogłosek/s).

	$ROS^{PP}_{szybkie}$			$ROS^{PP}_{średnie}$		
L.p.	**metoda A**	**metoda B**	**metoda C**	**metoda A**	**metoda B**	**metoda C**
1 (IS)	–8,34	–4,17	–8,34	–10,00	–13,33	3,33
2 (IS)	16,67	12,50	–8,33	0	–30,00	–20,00
3 (IS)	29,16	20,83	33,33	–3,33	10,00	13,33
4 (IS)	33,33	–4,17	25,00	6,67	0	0
5 (IS)	4,17	16,67	–8,33	–3,33	–16,67	13,33
6 (IS)	33,33	–12,50	12,50	10,00	10,00	3,33
7 (IS)	0	–20,83	12,50	–13,33	3,33	10,00
8(AS)	16,67	4,17	12,50	–3,33	16,67	20,00
9 (AS)	0	–8,33	0	–10,00	0	–6,67
10 (AS)	12,50	0	16,67	23,33	16,67	13,33
11 (AS)	0	–8,33	–16,67	0	–3,33	–6,67
wartość średnia	12,49999	–0,38	6,44	–0,30	–0,61	3,94
odchylenie standardowe	14,67235	12,84	15,85	10,48	14,59	11,72

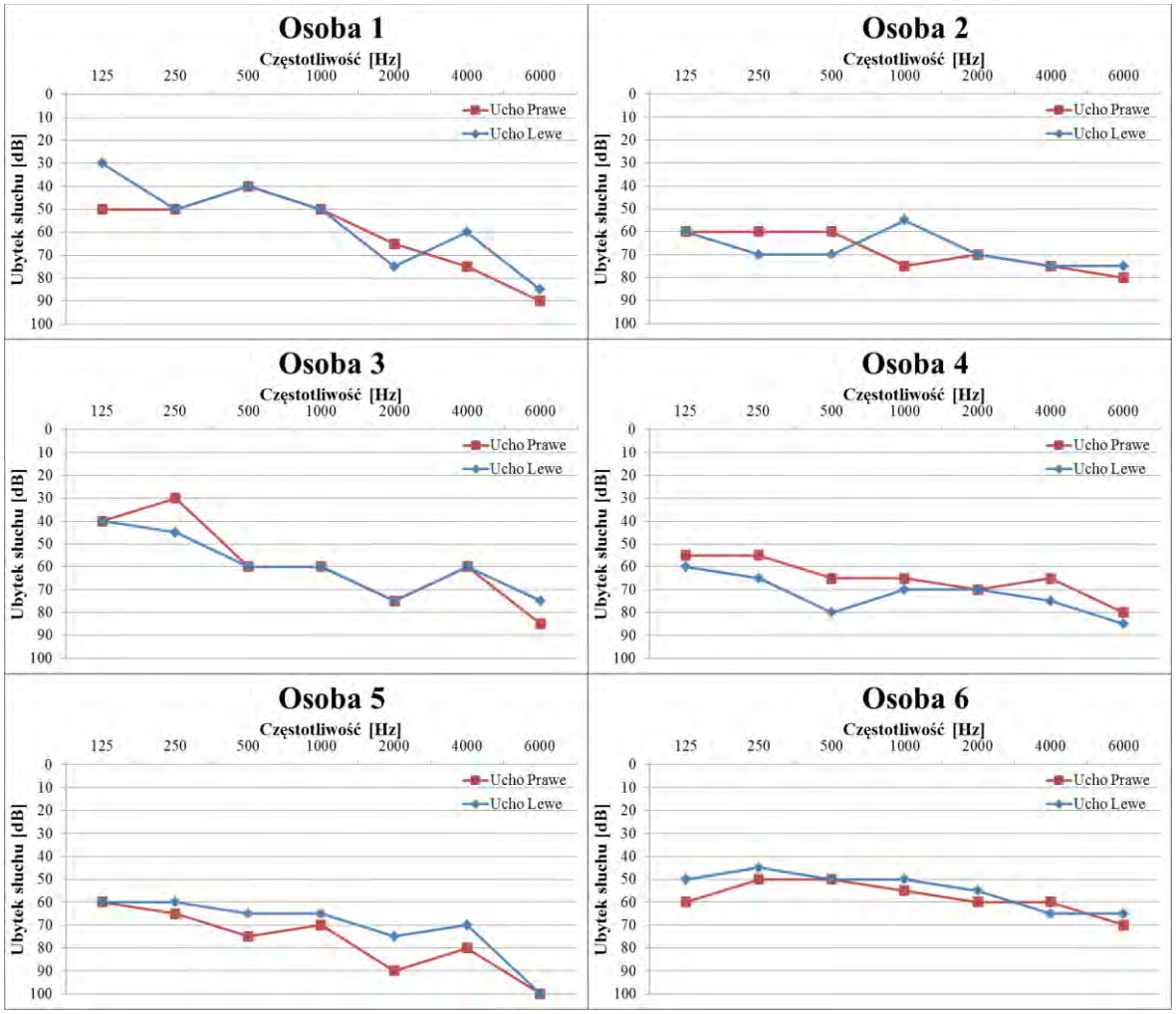
9.5 Załącznik nr 5

Tab. 9.13 Poprawa rozumienia mowy zmodyfikowanej dla osób starszych z pogorszoną rozdzielczością czasową słuchu ($TCT_{50} < 3{,}99$ samogłosek/s).

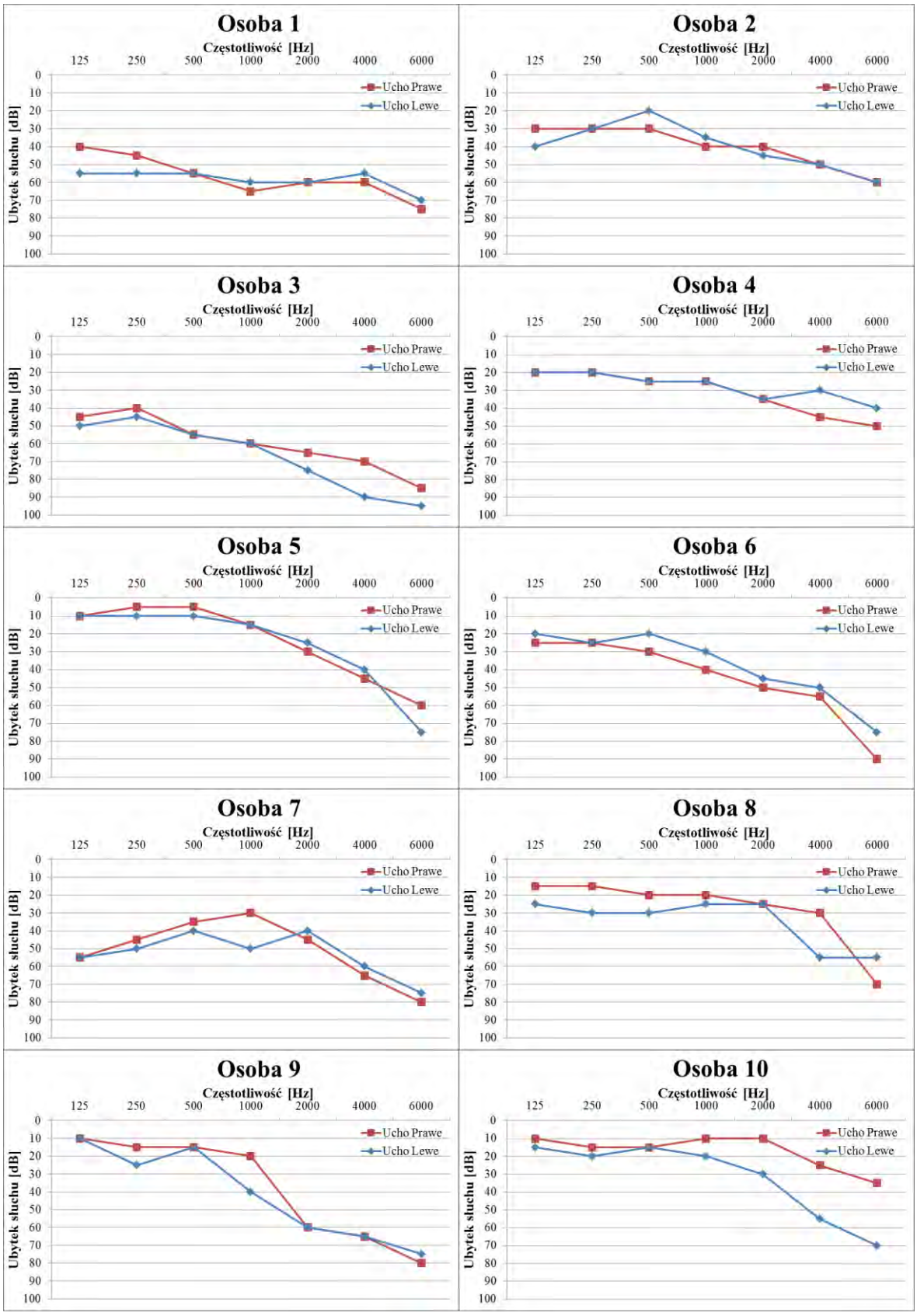
	$ROS^{P}_{szybkie}$			$ROS^{P}_{średnie}$		
L.p.	**metoda A**	**metoda B**	**metoda C**	**metoda A**	**metoda B**	**metoda C**
1	6	16	2	4	–4	4
2	–12	–6	–16	10	2	6
3	12	16	20	6	22	20
4	44	26	28	22	20	36
5(AS)	44	46	36	8	4	0
6	26	14	18	–4	–8	4
wartość średnia	20	18,67	14,67	7,67	6	11,67
odchylenie standardowe	22,23	17,01	18,83	8,52	12,39	13,76

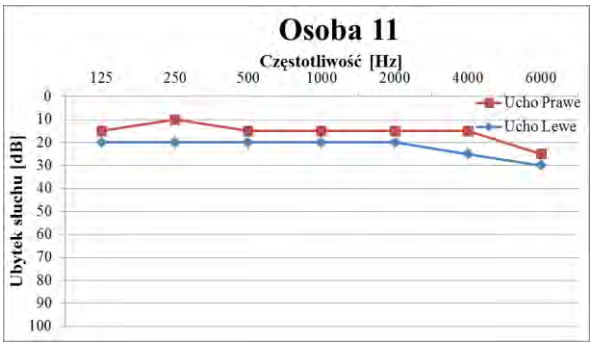
Tab. 9.14 Poprawa rozumienia mowy zmodyfikowanej dla osób starszych z normalną rozdzielczością czasową słuchu ($TCT_{50} \geq 3{,}99$ samogłosek/s).

	$ROS^{P}_{szybkie}$			$ROS^{P}_{średnie}$		
L,p,	**metoda A**	**metoda B**	**metoda C**	**metoda A**	**metoda B**	**metoda C**
1 (AS)	12	0	6	–4	–4	–2
2 (AS)	8	10	–8	2	2	–4
3	10	26	6	2	4	–2
4	6	2	8	–2	–2	2
5	4	10	12	2	4	6
6	6	2	–4	4	–4	–4
7	–22	0	2	4	4	–6
8	0	0	0	–2	–4	0
9	2	2	0	–2	0	0
10	0	0	0	–2	0	0
11	–2	4	6	4	4	2
wartość średnia	2,18	5,09	2,54	0,54	0,36	0,72
odchylenie standardowe	9,14	7,87	5,73	2,98	3,44	3,38



Rys. 9.4 Audiogramy osób starszych z pogorszoną rozdzielczością czasową słuchu.


Osoba 1
Osoba 2
Osoba 3
Osoba 4
Osoba 5
Osoba 6
Osoba 7
Osoba 8
Osoba 9
Osoba 10
Częstotliwość [Hz]
125 250 500 1000 2000 4000 6000
Ubytek słuchu [dB]
0 10 20 30 40 50 60 70 80 90 100
Ucho Prawe
Ucho Lewe



Rys. 9.4 Audiogramy osób starszych z normalną rozdzielczością czasową słuchu.

9.6 Załącznik nr 6

Na płycie dołączonej do rozprawy, w katalogu o nazwie „Załącznik nr 6” znajdują się przykładowe nagrania zmodyfikowane za pomocą metod B i R. Są to nagrania, które wykorzystano podczas testów subiektywnych opisanych w rozdziale 5. W podkatalogach o nazwach: „metoda B”, „metoda R” i „niezmodyfikowane” umieszczono odpowiednio nagrania zmodyfikowane za pomocą metody B, R i nagrania niezmodyfikowane. W podkatalogach odpowiadających obu metodom znajdują się także podkatalogi o nazwach „1.25”, „1.33”, „1.5” oraz „1.75”. Nazwy podkatalogów odpowiadają wartości α_o użytym do modyfikacji nagrań.