Politechnika Wrocławska

DZIEDZINA: nauki inżynieryjno-techniczne

DYSCYPLINA: automatyka, elektronika, elektrotechnika i technologie kosmiczne

ROZPRAWA DOKTORSKA

Propozycje algorytmów parametryzacji, odpornych na zmienność tonu krtaniowego, dla automatycznego rozpoznawania mowy

Mgr inż. Stanisław M. Gmyrek

Promotor / Promotorzy:
prof. dr hab. inż. Ryszard Makowski

Promotor pomocniczy: dr inż. Robert Hossa

Słowa kluczowe: automatyczne rozpoznawanie mowy, odporna parametryzacja, gaussowski model rozkładów prawdopodobieństw, rozpoznawanie fonemów

WROCŁAW 2026

Streszczenie

Projektowanie systemów Automatycznego Rozpoznawania Mowy (ARM) wymaga uwzględnienia losowości i zmienności sygnału mowy, wynikającej z różnorodnych czynników, których negatywny wpływ powoduje obniżenie efektywności działania tych systemów. Szczególną uwagę należy poświęcić zmienności wewnątrzosobniczej oraz międzyosobniczej mówców. Systemy ARM bazują na widmach fragmentów mowy, gdyż słuch jest analizatorem widma. Okresowość pobudzenia oraz zmienność częstotliwości tonu krtaniowego prowadzą do degradacji tych widm, a w konsekwencji do zwiększenia wariancji estymatorów współczynników cepstralnych i pogorszenia wyników rozpoznawania. W celu zminimalizowania negatywnego wpływu różnych czynników stosuje się rozmaite techniki kompensacyjne, takie jak: grupowanie mówców, adaptacja modeli akustycznych, normalizacja wartości wektorów obserwacji oraz odporna parametryzacja.

Głównym celem badawczym niniejszej pracy było opracowanie algorytmów odpornej parametryzacji ukierunkowanych na kompensację negatywnego wpływu zmienności częstotliwości tonu krtaniowego i poprawę jakości rozpoznawania.

W ramach pracy zaproponowano dwa autorskie algorytmy odpornej parametryzacji. Pierwszy z nich, metoda PS-HFCC (ang. *Pitch-Synchronized Human Factor Cepstral Coefficients*), stanowi uogólnienie klasycznej metody parametryzacji cepstralnej poprzez analizę sygnału w ramkach o zmiennej długości, synchronizowanej z okresem tonu krtaniowego. Drugi algorytm bazuje na filtracji odwrotnej IAIF (ang. *Iterative Adaptive Inverse Filtering*), realizowanej z wykorzystaniem algorytmu LPC (ang. *Linear Predictive Coding*). Filtracja odwrotna pozwala na estymację modułu transmitancji kanału głosowego, który staje się podstawą parametryzacji.

W badaniach zastosowano powszechnie wykorzystywany w systemach ARM statystyczny model akustyczny postaci GMM (ang. *Gaussian Mixture Model*). Model ten opisuje wartości współczynników cepstralnych każdej jednostki fonetycznej jako ważoną sumę rozkładów Gaussa, określoną za pomocą wektorów wartości oczekiwanych, macierzy kowariancji oraz współczynników wagowych. Wyznaczanie elementów GMM realizowano w oparciu o klasyczny algorytm Bauma-Welcha, stanowiący algorytm estymacji typu EM (ang. *Expectation-Maximization*).

Jako miary skuteczności zaproponowanych metod przyjęto: (i) odchylenia standardowe współrzędnych wektorów obserwacji, (ii) dywergencję Kullbacka-Leiblera (stanowiącą miarę odległości pomiędzy wielowymiarowymi rozkładami prawdopodobieństw GMM) oraz (iii) wskaźnik błędów rozpoznawania pojedynczych ramek sygnału – FER (ang. *Frame Error Rate*). Wyniki analizy wskazują, że zaproponowane przez autora metody przewyższają pod względem skuteczności klasyczne po-

dejście HFCC, zapewniając lepsze rozpoznawanie najczęściej występujących fonemów dźwięcznych i kompensują negatywny wpływ zmienności częstotliwości tonu krtaniowego na współczynniki parametryzacji.

Praca zawiera pięć rozdziałów. We wstępie omówiono podstawowe zagadnienia związane z funkcjonowaniem systemów ARM oraz określono cel, tezę i założenia badawcze niniejszej pracy. W rozdziale drugim dokonano przeglądu literatury dotyczącej metod parametryzacji sygnału mowy, opisano działanie konwencjonalnych i współczesnych systemów ARM oraz omówiono wspomniane wcześniej techniki kompensacyjne. W rozdziale trzecim przybliżono główny problem badawczy i zaprezentowano autorskie algorytmy odpornej parametryzacji. W rozdziale czwartym przeprowadzono badania i analizę zaproponowanych metod na sygnałach modelowych i rzeczywistych oraz przedstawiono wyniki rozpoznawania. Rozdział piąty stanowi podsumowanie wraz z wnioskami końcowymi.

Słowa kluczowe: automatyczne rozpoznawanie mowy, odporna parametryzacja, gaussowski model rozkładów prawdopodobieństw, rozpoznawanie fonemów

Abstract

The design of Automatic Speech Recognition (ASR) systems requires consideration of speech signal variability due to a variety of factors that reduce the effectiveness of these systems. Particular attention must be paid to the intra-personal and inter-personal variability of speakers. ASR systems rely on the spectral representations of speech segments, since the human auditory system functions as a spectrum analyzer. The periodicity of the glottal excitation and the variability of the fundamental frequency lead to degradation of that spectra and, consequently, to an increase in the variance of cepstral coefficient estimators, which significantly worsens recognition results. Various compensatory techniques, such as: speaker clustering, acoustic model adaptation, normalisation of observation vectors values and robust parameterisation are used to minimise the negative impact of various factors.

The primary research objective of this study was to develop robust parameterization algorithms to compensate for the adverse effects of variability in the speech signal's fundamental frequency and to improve recognition performance.

Two own algorithms for robust parametrization were proposed in this work. The first one, the PS-HFCC (Pitch-Synchronised Human Factor Cepstral Coefficients) method, is a generalisation of classical cepstral methods, enriched with a variable signal frame length analysis, whose size is synchronised with the pitch period. The second algorithm is based on Iterative Adaptive Inverse Filtering (IAIF), implemented using the Linear Predictive Coding (LPC) method. Inverse filtering allows

the estimation of the vocal tract transfer function, which forms the basis for parameterisation.

The study used a statistical acoustic model commonly used in ASR systems - the Gaussian Mixture Model (GMM). This model describes each phonetic unit as a weighted sum of Gaussian distributions specified by expectation value vectors, covariance matrices and weighting factors. The determination of these parameters was realised using the classical Baum-Welch algorithm, which is an EM (Expectation-Maximisation) estimation method.

The standard deviations of the observation vector coordinates, the Kullback-Leibler divergence (which is a measure of the distance between the multivariate GMM probability distributions), and the Frame Error Rate (FER) of the recognition of single signal frames were used as performance measures of the proposed methods. The results of the analysis indicate that the robust parametrization methods proposed in this paper exceed the classical HFCC approach in terms of performance, providing better recognition of the most frequently occurring voiced phonemes and compensating for the negative effect of fundamental frequency on the parametrization coefficients.

This dissertation is organized into five chapters. The introduction discusses the basic issues related to the functioning of ASR systems and defines the objective, thesis, and research assumptions of this work. Chapter two contains a review of the literature on speech signal parameterization methods, a description of the operation of conventional and modern ASR systems, and presents the aforementioned compensation techniques. The third chapter outlines the main research problem and presents the author's algorithms for robust parameterization. The fourth chapter analyzes the proposed methods on model and real signals and presents the recognition results. The fifth chapter is a summary with final conclusions.

Keywords: automatic speech recognition, robust parametrization, gaussian mixture model, phoneme recognition

Spis treści

Spis rysunków **7**
Spis tabel **10**
Wykaz ważniejszych skrótów stosowanych w pracy **12**
1. Wstęp **15**
1.1. Klasyfikacja systemów ARM 15
1.2. Cele i teza pracy 16
1.2.1. Układ pracy 17
2. Podstawy teoretyczne i przegląd literaturowy algorytmów parametryzacji sygnałów systemach ARM **18**
2.1. Schemat konwencjonalnego systemu ARM 18
2.2. Model sygnału mowy 20
2.3. Parametryzacja sygnału 22
2.3.1. Odporna parametryzacja 23
2.3.2. Schemat parametryzacji HFCC 27
2.3.3. Wyznaczenie wektora obserwacji 30
2.4. Architektury nowoczesnych systemów ARM 32
3. Problem badawczy i autorskie rozwiązania w zakresie odpornej parametryzacji **35**
3.1. Wpływ częstotliwości tonu krtaniowego na współczynniki parametryzacji . . . 35
3.2. Własne rozwiązania problemu korekcji widma ramki sygnału 38
3.2.1. Zastosowanie zmiennej długości ramki sygnału 38
3.2.2. Metoda wykorzystująca filtrację odwrotną 40
3.2.3. Podsumowanie 42
4. Badania skuteczności algorytmów i jakości rozpoznawania **43**
4.1. Metodologia badań 43
4.1.1. Sygnały modelowe i ich charakterystyka 43
4.1.2. Baza sygnałów rzeczywistych 46
4.1.3. Statystyczny model fonemów 48
4.2. Miary skuteczności rozpoznawania 50
4.2.1. Odchylenia standardowe współrzędnych wektora obserwacji 50
4.2.2. Odległości pomiędzy rozkładami prawdopodobieństwa GMM 50

4.2.3. Wskaźnik błędnie rozpoznanych ramek 52
4.3. Wyniki badań i wnioski 52
4.3.1. Przykładowe wyniki na sygnałach modelowych 52
4.3.2. Przykładowe wyniki dla sygnałów rzeczywistych 55
4.3.3. Globalna analiza błędów 57
4.4. Podsumowanie 74
5. Podsumowanie 76
Literatura 79
A. Estymacja częstotliwości tonu krtaniowego 88
A.1. Klasyfikacja algorytmów estymacji 89
A.2. Algorytmy korelacyjne estymacji częstotliwości podstawowej 90
B. Opis pozostałych algorytmów wykorzystywanych w pracy 92
B.1. Detekcja dźwięcznych ramek sygnału mowy 92
B.2. Algorytm filtracji odwrotnej IAIF (ang. Iterative Adaptive Inverse Filtering) . . 92
C. Błędy rozpoznawania par fonemów 94

Spis rysunków

2.1. Szczegółowy schemat architektury konwencjonalnych systemów ARM wykorzystujących model akustyczny AM, model językowy LM i słownik PM. 18
2.2. Sygnał pobudzenia w postaci ciągu impulsów tonu krtaniowego według modelu Rosenberga 21
2.3. Moduł transmitancji kanału głosowego z formantami odpowiadającymi samogłosce a. 21
2.4. Fragment podsystemu akustycznego z odporną parametryzacją w systemie ARM. . 24
2.5. Fragment podsystemu akustycznego z grupowaniem mówców w systemie ARM. . . 25
2.6. Fragment podsystemu akustycznego z normalizacją wartości wektora parametrów w systemie ARM. 26
2.7. Fragment podsystemu akustycznego z adaptacją modelu statystycznego w systemie ARM. 27
2.8. Schemat obliczania współczynników cepstralnych. 28
2.9. Bank filtrów wykorzystywany w stosowanej parametryzacji HFCC 30
2.10. Schemat architektury End-To-End systemów ARM. Blok ekstrakcji cech jest w niektórych przypadkach zintegrowany wewnątrz modelu DNN, jednak najczęściej nawet w wewnętrznych reprezentacjach modeli głębokich dokonywana jest parametryzacja sygnału. 33
3.1. Widma amplitudowe ramek sygnału mowy zawierających fonem a z naniesionymi w tle bankami filtrów stosowanymi w parametryzacji HFCC. Częstotliwość tonu krtaniowego wynosi: a) $F_0 = 130$ Hz, b) $F_0 = 195$ Hz. 36
3.2. Widma melowe poszczególnych ramek sygnału mowy zawierających fonem a. Częstotliwość tonu krtaniowego wynosi: a) $F_0 = 130$ Hz, b) $F_0 = 195$ Hz. 37
3.3. Cepstra poszczególnych ramek sygnału mowy zawierających fonem a. Częstotliwość tonu krtaniowego wynosi: a) $F_0 = 130$ Hz, b) $F_0 = 195$ Hz. 37
3.4. Schemat obliczania współczynników parametryzacji PS-HFCC 38
3.5. Ilustracja wpływu synchronizacji długości ramki z okresem podstawowym na widma sygnałów rzeczywistych: a) przebieg czasowy sygnału w ramce z zaznaczonymi okresami T_1, T_2, T_3, b) widmo ramki sygnału z widocznymi zafalowaniami, c) widmo okresu T_2, d) widmo okresu T_3. 39
3.6. Schemat wyznaczania współczynników parametryzacji IF-HFCC. 41

4.1. Impuls pobudzenia według modelu Rosenberga. 44
4.2. Przykładowy przebieg czasowy modelowego sygnału pobudzenia $x(n)$ o częstotliwości $F_0 = 120$ Hz. 44
4.3. Przykładowy moduł funkcji transmitancji modelowego kanału głosowego dla samogłoski a. 45
4.4. Przykładowy przebieg czasowy sygnału dla samogłoski a, wygenerowanego na podstawie modelu źródło-filtr Fanta. 46
4.5. Unormowane widma kilku ramek modelowego sygnału mowy zawierających fonem a wraz z naniesionym modułem transmitancji kanału głosowego (czerwona linia przerywana). 53
4.6. Przykładowe wyniki zastosowania odpornej parametryzacji PS-HFCC. Unormowane widma amplitudowe kilku ramek modelowego sygnału mowy wraz z wykreślonym modułem transmitancji kanału głosowego (czerwona linia przerywana). . . 53
4.7. Przykładowe wyniki zastosowania metody filtracji odwrotnej IAIF. Estymatory $H_{v2}(z)$ kilku ramek modelowego sygnału mowy wraz z wykreślonym modułem transmitancji kanału głosowego (czerwona linia przerywana). 54
4.8. Cepstra kilku ramek modelowego sygnału, wyznaczone z widm z rys. 4.5, uzyskane metodą HFCC 54
4.9. Cepstra kilku ramek modelowego sygnału wyznaczone metodą a) PS-HFCC, b) IF-HFCC. 54
4.10. Unormowane widma kilku ramek rzeczywistego sygnału mowy zawierających fonem e zaczerpnięte z nagrań głosu tego samego mówcy. 55
4.11. Przykłady wygładzonych widm kilku ramek rzeczywistego sygnału mowy zawierających fonem e, używane do parametryzacji PS-HFCC. 56
4.12. Przykładowe estymatory $H_{v2}(z)$ kilku ramek rzeczywistego sygnału mowy zawierających fonem e, używane do parametryzacji IF-HFCC. 56
4.13. Cepstra poszczególnych ramek sygnału mowy, zawierających fonem e, wyznaczone z widm zawierających okresowe powielenia z rys. 4.10. 57
4.14. Cepstra poszczególnych ramek sygnału mowy, zawierających fonem e, wyznaczone z widm z rys. 4.11 - 4.12 57
4.15. Porównanie wartości odchyleń standardowych $\sigma_{f,p}$ dla samogłoski e i trzech metod parametryzacji. 58
4.16. Porównanie wartości odchyleń standardowych $\sigma_{f,p}$ dla spółgłoski r i trzech metod parametryzacji. 58
4.17. Globalna miara odległości D_f dla samogłosek mowy polskiej przed korekcją (krzywa niebieska) i po korekcji (krzywa czarna i czerwona). 66
4.18. Globalna miara odległości D_f dla spółgłosek mowy polskiej przed korekcją (krzywa niebieska) i po korekcji (krzywa czarna i czerwona). 67

4.19. Globalny błąd FER wyrażony miarą E_f dla samogłosek mowy polskiej. Porównanie trzech metod parametryzacji dla grupy g_0, stanowiącej zbiór wszystkich mówców. . 68
4.20. Globalny błąd FER wyrażony miarą E_f dla spółgłosek mowy polskiej. Porównanie trzech metod parametryzacji grupy g_0, stanowiącej bazę wszystkich mówców. . . . 68
4.21. Globalny błąd FER wyrażony miarą E_f dla samogłosek mowy polskiej. Porównanie trzech metod parametryzacji dla mówców z grupy g_1. 69
4.22. Globalny błąd FER wyrażony miarą E_f dla spółgłosek mowy polskiej. Porównanie trzech metod parametryzacji dla mówców z grupy g_1. 69
4.23. Globalny błąd FER wyrażony miarą E_f dla samogłosek mowy polskiej. Porównanie trzech metod parametryzacji dla mówców z grupy g_2. 70
4.24. Globalny błąd FER wyrażony miarą E_f dla spółgłosek mowy polskiej. Porównanie trzech metod parametryzacji dla mówców z grupy g_2. 70
4.25. Globalny błąd FER wyrażony miarą E_f dla samogłosek mowy polskiej. Porównanie trzech metod parametryzacji dla mówców z grupy g_3. 70
4.26. Globalny błąd FER wyrażony miarą E_f dla spółgłosek mowy polskiej. Porównanie trzech metod parametryzacji dla mówców z grupy g_3. 71
4.27. Globalny błąd FER wyrażony miarą E_f dla samogłosek mowy polskiej. Porównanie trzech metod parametryzacji dla mówców z grupy g_4. 71
4.28. Globalny błąd FER wyrażony miarą E_f dla spółgłosek mowy polskiej. Porównanie trzech metod parametryzacji dla mówców z grupy g_4. 71
4.29. Globalny błąd FER wyrażony miarą E_f dla samogłosek mowy polskiej. Porównanie trzech metod parametryzacji dla mówców z grupy g_5. 72
4.30. Globalny błąd FER wyrażony miarą E_f dla spółgłosek mowy polskiej. Porównanie trzech metod parametryzacji dla mówców z grupy g_5. 72
4.31. Globalny błąd FER wyrażony miarą E_f dla samogłosek mowy polskiej. Porównanie trzech metod parametryzacji dla mówców z grupy g_6. 72
4.32. Globalny błąd FER wyrażony miarą E_f dla spółgłosek mowy polskiej. Porównanie trzech metod parametryzacji dla mówców z grupy g_6. 73
A.1. Unormowane histogramy częstotliwości tonu krtaniowego dla głosów damskich (krzywa czerwona) i męskich (krzywa niebieska) wykorzystywanej w pracy bazy. . 88
B.1. Schemat algorytmu filtracji odwrotnej IAIF. 93

Spis tabel

4.1. Zakresy średnich wartości częstotliwości formantowych dla samogłosek mowy polskiej . . . 45
4.2. Częstość występowania fonemów mowy polskiej wyrażona w [%] wraz zapisem transkrypcji fonemów wg. standardu IPA. . . . 47
4.3. Wartości odchyleń standardowych $\sigma_{f,p}$ poszczególnych samogłosek obliczone oddzielnie dla wszystkich P współrzędnych wektorów parametrów cepstralnych. Pierwsza wartość dla każdego współczynnika została obliczona z parametrów HFCC, druga z parametrów PS-HFCC, a trzecia z IF-HFCC. . . . 59
4.4. Wartości odchyleń standardowych $\sigma_{f,p}$ poszczególnych spółgłosek obliczone oddzielnie dla wszystkich P współrzędnych wektorów parametrów cepstralnych. Pierwsza wartość dla każdego współczynnika została obliczona z parametrów HFCC, druga z parametrów PS-HFCC, a trzecia z IF-HFCC. . . . 60
4.5. Względne różnice Δ_{KLD} odległości rozkładów GMM samogłosek pomiędzy fonemami uzyskane dla metod HFCC i PS-HFCC. Wartości wyrażone są w procentach, kolor zielony oznacza wzrost odległości po korekcji, a kolor czerwony jej zmniejszenie. . . . 62
4.6. Względne różnice Δ_{KLD} odległości rozkładów GMM spółgłosek pomiędzy fonemami uzyskane dla metod HFCC i PS-HFCC. Wartości wyrażone są w procentach, kolor zielony oznacza wzrost odległości po korekcji, a kolor czerwony jej zmniejszenie. . . . 63
4.7. Względne różnice Δ_{KLD} odległości rozkładów GMM samogłosek pomiędzy fonemami uzyskane dla metod HFCC i IF-HFCC. Wartości wyrażone są w procentach, kolor zielony oznacza wzrost odległości po korekcji, a kolor czerwony jej zmniejszenie. . . . 64
4.8. Względne różnice Δ_{KLD} odległości rozkładów GMM spółgłosek pomiędzy fonemami uzyskane dla metod HFCC i IF-HFCC. Wartości wyrażone są w procentach, kolor zielony oznacza wzrost odległości po korekcji, a kolor czerwony jej zmniejszenie. . . . 65
C.1. Wartości błędu $E_{f,g}$, określającego częstość pomyłek pomiędzy parami fonemów. Rozpoznawanie samogłosek w metodzie PS-HFCC. . . . 95

C.2. Wartości błędu $E_{f,g}$, określającego częstość pomyłek pomiędzy parami fonemów. Rozpoznawanie spółgłosek w metodzie PS-HFCC. 96
C.3. Wartości błędu $E_{f,g}$, określającego częstość pomyłek pomiędzy parami fonemów. Rozpoznawanie samogłosek w metodzie IF-HFCC. 97
C.4. Wartości błędu $E_{f,g}$, określającego częstość pomyłek pomiędzy parami fonemów. Rozpoznawanie spółgłosek w metodzie IF-HFCC. 98

Wykaz ważniejszych skrótów stosowanych w pracy

AM	(ang. *Acoustic Model*)	model akustyczny
ARM	Automatyczne Rozpoznawanie Mowy	
CMVN	(ang. *Cepstrum Mean Variance Normalization*)	algorytm normalizacji polegający na usunięciu wartości średniej cepstrum i jego standaryzacji
CNN	(ang. *Convolutional Neural Network*)	konwolucyjne sieci neuronowe
DCT	(ang. *Discrete Cosine Transform*)	dyskretna transformata cosinusowa
DFT	(ang. *Discrete Fourier Transform*)	dyskretna transformata Fouriera
DNN	(ang. *Deep Neural Network*)	głębokie sieci neuronowe
DTFT	(ang. *Discrete Time Fourier Transform*)	transformata Fouriera sygnału o czasie dyskretnym
EM	(ang. *Expectation Maximization*)	algorytm iteracyjny służący do wyznaczania parametrów rozkładów GMM
ERB	(ang. *Equivalent Rectangular Bandwidth*)	ekwiwalent szerokości pasma prostokątnego
FC	(ang. *Frame Classification*)	rozpoznawanie na poziomie pojedynczych ramek sygnału

FE	(ang. *Feature Extraction*)	ekstrakcja cech
FER	(ang. *Frame Error Rate*)	wskaźnik błędnie rozpoznanych ramek sygnału
FFT	(ang. *Fast Fourier Transform*)	szybka transformata Fouriera
GMM	(ang. *Gaussian Mixture Model*)	model mieszaniny rozkładów gaussowskich
HFCC	(ang. *Human Factor Cepstral Coefficient*)	współczynniki cepstralne uwzględniające właściwości percepcji dźwięku przez człowieka
HMM	(ang. *Hidden Markov Model*)	ukryte modele Markova
IAIF	(ang. *Iterative Adaptive Inverse Filtering*)	metoda filtracji odwrotnej stosowana do estymacji impulsów pobudzenia
IDFT	(ang. *Inverse Discrete Fourier Transform*)	odwrotna transformata DFT
IF-HFCC	(ang. *Inverse Filtering - HFCC*)	współczynniki HFCC wyznaczone z sygnału po zastosowaniu filtracji odwrotnej
IPA	(ang. *International Phonetic Alphabet*)	międzynarodowy alfabet fonetyczny
KLD	(ang. *Kullback Leibler Distance*)	odległość Kullbacka-Leiblera
LM	(ang. *Language Model*)	model językowy
LP	(ang. *Linear Prediction*)	liniowa predykcja
LPC	(ang. *Linear Predictive Coding*)	liniowe kodowanie predykcyjne
LPCC	(ang. *Linear Prediction Cepstral Coefficients*)	współczynniki cepstralne obliczone z wykorzystaniem metody LPC
MFCC	(ang. *Mel Frequency Cepstral Coefficients*)	melowe współczynniki cepstralne

MVDR	(ang. *Minimum Variance Distortionless Response*)	sposób estymacji widma sygnału, parametryczny estymator minimalnej wariancji
PLP	(ang. *Perceptual Linear Prediction*)	modyfikacja metody LPCC polegająca na uwzględnieniu niektórych właściwości systemu słuchowego człowieka
PM	(ang. *Pronunciation Model*)	model wymowy (słownik)
PS-HFCC	(ang. *Pitch Synchronized HFCC*)	modyfikacja parametryzacji HFCC wykorzystująca synchronizację długości ramki z okresem podstawowym
RNN	(ang. *Recurrent Neural Network*)	rekurencyjne sieci neuronowe
SA	(ang. *speaker adaptive*)	(system) adaptujący się do mówcy
SD	(ang. *speaker dependent*)	(system) dedykowany do pracy z jednym mówcą
SI	(ang. *speaker independent*)	(system) dedykowany do pracy z wieloma mówcami
SM	(ang. *Sequential Model*)	model sekwencyjny
STFT	(ang. *Short Time Fourier Transform*)	krótkookresowa transformata Fouriera
VTLN	(ang. *Vocal Tract Length Normalization*)	normalizacja długości kanału głosowego
WER	(ang. *Word Error Rate*)	wskaźnik błędnie rozpoznanych słów

Rozdział 1

Wstęp

Pod pojęciem Automatycznego Rozpoznawania Mowy (ARM) należy rozumieć przetwarzanie sygnału mowy za pomocą urządzeń cyfrowych i oprogramowania w celu ekstrakcji użytecznej dla człowieka informacji, którą jest zawartość semantyczna (znaczeniowa) mowy, czyli jej treść. Jeśli przetwarzanie sygnału ma na celu identyfikację mówcy, rozważamy systemy Automatycznego Rozpoznawania Głosu (ARG), a jeśli użyteczną dla człowieka informacją jest stan emocjonalny mówcy (np. złość, smutek, zdenerwowanie) i projektowane algorytmy mają na celu automatyczne jego rozpoznawanie, rozważamy systemy Automatycznego Rozpoznawania Emocji (ARE). Pierwsze wzmianki w literaturze o automatycznym rozpoznawaniu mowy sięgają lat 50. XX wieku i dotyczyły rozpoznawania pojedynczych słów wypowiadanych przez jednego mówcę. Mimo upływu dziesięcioleci, przed osobami projektującymi takie systemy nadal stoi wiele wyzwań, gdyż nie udało się w zadowalającym stopniu opracować w pełni skutecznego systemu realizującego chociażby konwersję na tekst swobodnej dyskusji wielu osób. Trudnością w zaprojektowaniu uniwersalnego systemu ARM o odpowiednio dużej skuteczności jest bogactwo językowe i złożona natura mowy, która charakteryzuje się dużą zmiennością nawet przy przekazywaniu tej samej informacji przez tego samego mówcę. Wpływ na mowę ma bardzo wiele czynników, m.in. technicznych, środowiskowych oraz wewnątrz- i międzyosobniczych, o czym szczegółowo napisano w rozdz. 2.3.1. Duże zespoły badawcze na całym świecie cały czas pracują nad udoskonaleniem istniejących systemów w zakresie złożoności obliczeniowej algorytmów i szybkości ich działania (szczególnie dla systemów działających w czasie rzeczywistym i tych, które wykorzystują głębokie modele sieci neuronowych), poprawy jakości dostarczanej transkrypcji, dokładności klasyfikacji i rozpoznawania, czy dostarczeniu nowych metod odpornej parametryzacji sygnału.

1.1. Klasyfikacja systemów ARM

Systemy ARM można sklasyfikować, biorąc pod uwagę różne kryteria [6] [69] [57]. Są to m.in.:

- **zależność od mówcy** - wyróżnić można systemy SD (ang. *speaker-dependent*), czyli dedykowane do pracy z jednym mówcą, systemy SI (ang. *speaker-independent*), czyli zaprojektowane do pracy z wieloma mówcami oraz systemy SA (ang. *speaker-adaptive*), mające możliwość adaptacji swoich parametrów i dostrajania się do aktualnego mówcy.
- **rodzaj rozpoznawanej mowy** - przy projektowaniu systemu ARM, aby nadać odpowiednie znaczenie jednostkom językowym, należy zdefiniować zbiór rozpoznawanych słów, czyli słownik. Wyróżnić można: (i) systemy zaprojektowanie do rozpoznawania komend, izolowanych wyrazów, charakteryzujące się stosunkowo niewielkim słownikiem, (ii) systemy zaprojektowane do rozpoznawania sekwencji wyrazów, (iii) systemy do rozpoznawania mowy ciągłej, dla których dodatkowym utrudnieniem jest niewystępowanie przerw pomiędzy słowami, co wymaga od osób projektujących takie systemy, dostarczenia wysokiej skuteczności algorytmów segmentacji,
- **warunki środowiskowe i stany emocjonalne** - bardzo duży wpływ na działanie systemu mają zniekształcenia i zakłócenia, ponieważ powodują zmianę widm fragmentów sygnału. Wyróżnić można: (i) zniekształcenia typu inercyjnego związane np. z rejestracją i transmisją sygnału, (ii) zniekształcenia odbiciowe będące efektem odbić fali akustycznej w pomieszczeniu, (iii) zniekształcenia nieliniowe, powodujące dużą degradację sygnału mowy, (iv) zakłócenia szumowe spłaszczające widmo sygnału, (v) zakłócenia impulsowe zmieniające bardzo znacznie te fragmenty mowy, w których występują. Znaczny poziom zakłóceń (SNR poniżej 6 dB) powoduje problemy ze zrozumiałością mowy również przez człowieka. Ważnym czynnikiem jest także zmienność stanu emocjonalnego mówcy, tj. zmienność sygnału mowy spowodowana artykulacją słów w warunkach stresu, złości i innych silnych stanów emocjonalnych, ale również mowa szeptana czy krzyk,
- **uwarunkowania sprzętowe** - możliwości systemów ARM, ich skuteczność, złożoność proponowanych algorytmów przetwarzania zależą znacząco od uwarunkowań sprzętowych. Można je projektować na specjalizowanych serwerach składających się z bardzo wielu procesorów, ale także na komputerze PC z pojedynczym procesorem, czy tanim procesorze sygnałowym o niewielkiej mocy obliczeniowej. Istotnym czynnikiem jest też czas realizacji rozpoznawania. Jeśli nie odgrywa on kluczowej roli, możliwe jest stosowanie bardziej złożonych algorytmów, ale w aplikacjach czasu rzeczywistego priorytetem staje minimalizacja złożoności obliczeniowej.

1.2. Cele i teza pracy

Teza pracy

Możliwe jest zaprojektowanie metody odpornej parametryzacji sygnału mowy dla systemów ARM, kompensującej wpływ okresowości pobudzenia i jej zmienności, o skuteczności nie mniejszej niż skuteczność metod znanych z literatury.

Przyjęte założenia

W celu uszczegółowienia tezy pracy przyjęto następujące założenia:

- baza nagrań wykorzystywana w badaniach składa się z krótkich wypowiedzi (pojedynczych wyrazów) różnych mówców płci męskiej,
- w badaniach skoncentrowano się na wyselekcjonowanych fonemach dźwięcznych mowy polskiej, a to ograniczenie wynika z badanych metod kompensacji,
- badania wykonano, korzystając z bazy nagrań zawierającej wypowiedzi 40 przypadkowych mówców zarejestrowanych w różnych regionach Polski,
- skuteczność poszczególnych metod kompensacji weryfikowano w oparciu o rozpoznawanie fonemów w pojedynczych ramkach sygnału (pojęcie ramki sygnału zdefiniowano w rozdz. 2.3.2), a miary skuteczności opisano w rozdz. 4.2.

Zadania niezbędne do osiągnięcia celu naukowego zawartego w tezie

Do osiągnięcia celu naukowego zawartego w tezie pracy wymagane jest wykonanie następujących zadań:

- przegląd literatury przedmiotu,
- analiza teoretyczna i eksperymentalna znanych algorytmów parametryzacji sygnału mowy,
- zaprojektowanie i implementacja własnych algorytmów odpornej parametryzacji,
- badania i porównanie skuteczności własnych rozwiązań z algorytmami dostępnymi w literaturze, analiza wyników.

1.2.1. Układ pracy

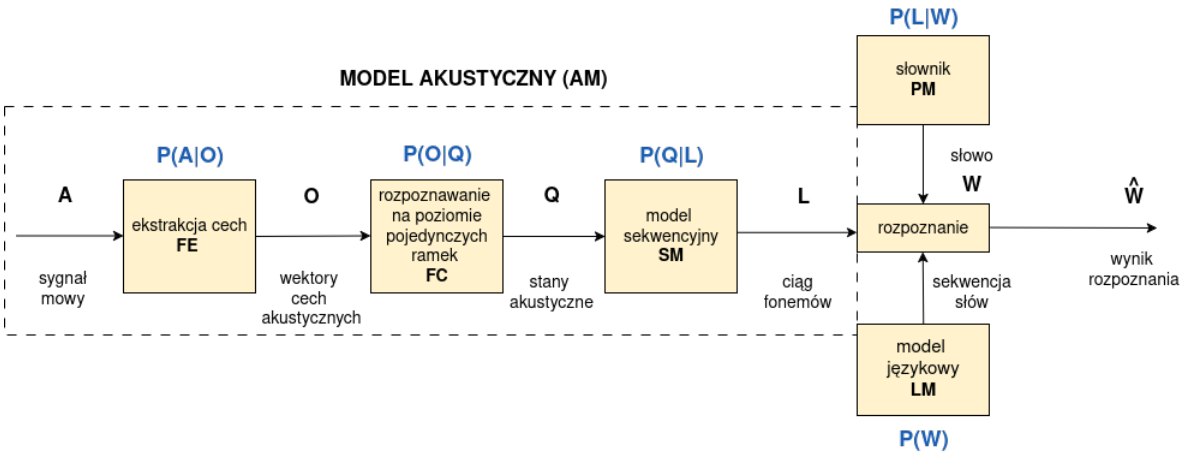
Rozdział drugi stanowi teoretyczne wprowadzenie do niniejszej pracy doktorskiej. Zaprezentowano w nim schemat blokowy konwencjonalnych systemów ARM, omówiono proces powstawania mowy, opierając się na teoretycznym modelu źródło-filtr Fanta. Opisano w nim także cel i kolejne etapy procesu parametryzacji sygnału, dokonano przeglądu literaturowego istniejących algorytmów parametryzacji, wymieniono czynniki, które należy brać pod uwagę przy projektowaniu systemu ARM oraz metody kompensacji ich negatywnego wpływu. Rozdział ten kończy krótki przegląd najnowocześniejszych architektur systemów opartych o głębokie modele sieci neuronowych. Rozdział trzeci koncentruje się na szczegółowej analizie problemu badawczego pracy, tj. wpływu zmienności częstotliwości tonu krtaniowego na wartości współczynników wektora obserwacji. Przedstawiono w nim także autorskie algorytmy z zakresu odpornej parametryzacji wraz z szczegółowymi opisami proponowanych rozwiązań. Rozdział czwarty to opis metodologii badawczej wraz z analizą skuteczności proponowanych algorytmów w odniesieniu do sygnałów modelowych i rzeczywistych, a także przykładowe wyniki oraz analiza błędów rozpoznawania. Pracę zamyka rozdział piąty, stanowiący podsumowanie oraz wnioski.

Rozdział 2

Podstawy teoretyczne i przegląd literaturowy algorytmów parametryzacji sygnałów systemach ARM

2.1. Schemat konwencjonalnego systemu ARM

Schemat blokowy klasycznych systemów ARM przedstawiono na rys. 2.1 [22].



Rys. 2.1: Szczegółowy schemat architektury konwencjonalnych systemów ARM wykorzystujących model akustyczny AM, model językowy LM i słownik PM.

Celem automatycznego rozpoznawania mowy jest znalezienie dla danej wypowiedzi najbardziej prawdopodobnego ciągu słów $\hat{\mathbf{W}}$, czyli maksymalizacja prawdopodobieństwa $P(\mathbf{W}|\mathbf{A})$,

że dana sekwencja słów W jest poprawną transkrypcją wypowiedzi $\mathbf{A}$:

$$\hat{\mathbf{W}} = \arg\max_{W} P(\mathbf{W}|\mathbf{A}). \tag{2.1}$$

Bezpośrednie obliczenie prawdopodobieństwa $P(\mathbf{W}|\mathbf{A})$ jest trudne, natomiast stosując twierdzenie Bayesa, można je przekształcić do postaci:

$$\hat{\mathbf{W}} = \arg\max_{W} P(\mathbf{W}|\mathbf{A}) = \arg\max_{W} \frac{P(\mathbf{A}|\mathbf{W}) \cdot P(\mathbf{W})}{P(\mathbf{A})}, \tag{2.2}$$

gdzie $P(\mathbf{A}|\mathbf{W})$ jest **modelem akustycznym - AM** (ang. *Acoustic Model*), określającym dopasowanie wypowiedzi $\mathbf{A}$ do konkretnego ciągu słów $\mathbf{W}$, $P(\mathbf{W})$ jest **modelem językowym - LM** (ang. *Language Model*), natomiast $P(\mathbf{A})$ jest całkowitym prawdopodobieństwem zaobserwowania wypowiedzi $\mathbf{A}$. Jest ono stałe i stanowi jedynie człon normalizacyjny, zatem można je pominąć przy analizie. Dalsze rozważania ograniczają się do maksymalizacji iloczynu:

$$\hat{\mathbf{W}} = \arg\max_{W} P(\mathbf{A}|\mathbf{W}) \cdot P(\mathbf{W}). \tag{2.3}$$

System ARM nie rozpoznaje słów bezpośrednio z dźwięku, lecz przekształca sygnał przez kolejne etapy przetwarzania zaprezentowane na schemacie z rys. 2.1, zatem zagregowane prawdopodobieństwo $P(\mathbf{A}|\mathbf{W})$ można zapisać przy pomocy iloczynu prawdopodobieństw związanych z kolejnymi blokami systemu:

$$P(\mathbf{A}|\mathbf{W}) \cdot P(\mathbf{W}) = P(\mathbf{A}|\mathbf{O}) \cdot P(\mathbf{O}|\mathbf{Q}) \cdot P(\mathbf{Q}|\mathbf{L}) \cdot P(\mathbf{L}|\mathbf{W}) \cdot P(\mathbf{W}), \tag{2.4}$$

gdzie:

- $P(\mathbf{A}|\mathbf{O})$ to prawdopodobieństwo zaobserwowania wypowiedzi $\mathbf{A}$ przy danych wektorach cech $\mathbf{O}$ reprezentowane przez blok **ekstrakcji cech - FE** (ang. *Feature Extraction*). W praktyce, dla danej wypowiedzi prawdopodobieństwo to uznaje się za stałe i pomijalne w obliczeniach, ze względu na deterministyczny charakter parametryzacji sygnału opisanej szczegółowo w rozdz. 2.3
- $P(\mathbf{O}|\mathbf{Q})$ stanowi prawdopodobieństwo obserwacji wektora cech $\mathbf{O}$ pod warunkiem zajścia konkretnego stanu akustycznego $\mathbf{Q}$ (np. fonemu). Jest to kluczowy etap **modelu akustycznego AM**, w którym odbywa się **rozpoznawanie pojedynczych ramek - FC** (ang. *Frame Classification*), czyli mapowanie cech akustycznych do konkretnych stanów fonetycznych. W tym celu wyznacza się statystyczne modele poszczególnych fonemów.
- $P(\mathbf{Q}|\mathbf{L})$ to **model sekwencyjny SM** (ang. *Sequential Model*), przy pomocy którego modeluje się prawdopodobieństwo przejść między stanami $\mathbf{Q}$ dla zadanych fonemów $\mathbf{L}$. Uwzględniane są w nim warianty artykulacyjne, kontekstowość i sąsiedztwo innych fonemów.
- $P(\mathbf{L}|\mathbf{W})$ to **słownik - PM** (ang. *Pronunciation Model*) będący zestawem wyrazów i odpowiadających im zapisów transkrypcji fonetycznej. Na tym etapie przetwarzania, obliczane jest prawdopodobieństwo, że dana sekwencja fonemów $\mathbf{L}$ odpowiada konkretnemu słowu $\mathbf{W}$.

Blok ten stanowi pomost pomiędzy modelem akustycznym i językowym, dokonując powiązania ciągu jednostek fonetycznych rozpoznanych przez model akustyczny z konkretnymi słowami występującymi w danym języku.

- $P(\mathbf{W})$ to prawdopodobieństwo wystąpienia w danym języku ciągu słów $\mathbf{W}$, stanowiące **model językowy LM**, którego głównym zadaniem jest określenie, jak bardzo analizowana fraza jest prawdopodobna w danym języku. Pomaga prognozować kolejne ciągi jednostek fonetycznych w oparciu o historię (np. poprzednie słowa), jednocześnie eliminując występowanie nielogicznych i nienaturalnych ciągów. W klasycznych podejściach modele te reprezentowane są przez n-gramy, opierające się na statystyce występowania słów danego języka, (np. dla n=3, tzw. trigram określa prawdopodobieństwo wystąpienia danego słowa na podstawie historii dwóch poprzednich). We współczesnych rozwiązaniach model ten realizowany jest za pomocą rekurencyjnych sieci neuronowych RNN (ang. *Recurrent Neural Network*), które w warstwach ukrytych przechowują informacje dotyczące kontekstu i całej historii słów. Lepiej radzą sobie one z zależnościami długoterminowymi, jednak muszą być trenowane na ogromnych bazach danych.

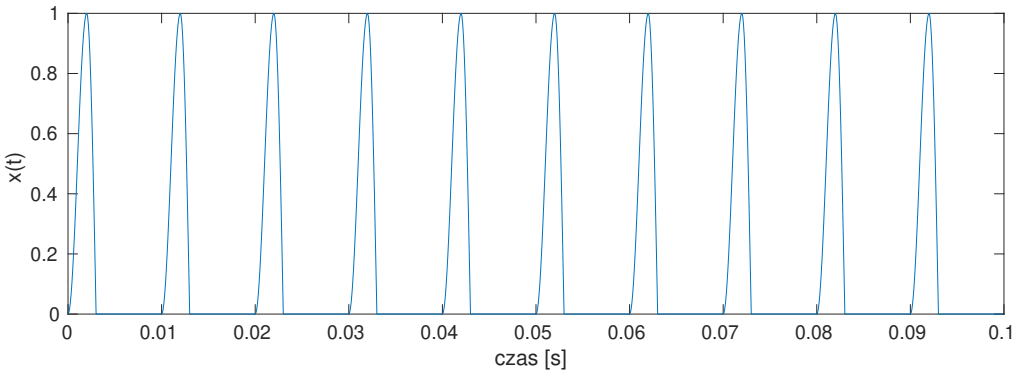
2.2. Model sygnału mowy

Mowa jest sygnałem akustycznym, a więc zmianą ciśnienia akustycznego w przestrzeni. Powszechnie akceptowalnym modelem sygnału mowy $s(n)$ jest model źródło-filtr Fanta [20] postaci:

$$s(n) = x(n) \star v(n) \star r(n), \tag{2.5}$$

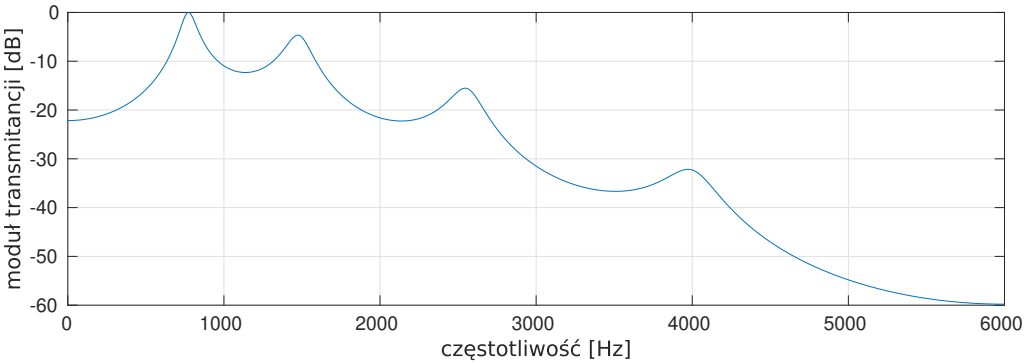
gdzie $x(n)$ jest pobudzeniem, $v(n)$ - odpowiedzią impulsową filtru modelującego kanał głosowego, a $r(n)$ - odpowiedzią impulsową opisującą emisję sygnału mowy przez usta mówcy, n - czasem dyskretnym, a $\star$ - operatorem splotu.

W procesie generowania mowy udział biorą płuca, pobudzające wydychanym powietrzem organ mowy człowieka, w tym struny głosowe. Gdy powietrze wychodzące z płuc wprawia je w drgania, to pobudzenie $x(n)$ ma postać ciągu impulsów tonu krtaniowego, sygnału w przybliżeniu okresowego, co jest związane z cyklicznością pracy strun głosowych. Generowana wówczas mowa ma charakter dźwięczny. Przykładowy przebieg sygnału pobudzenia $x(n)$ w postaci ciągu impulsów według modelu Rosenberga zaprezentowano na rys. 2.2. Gdy struny głosowe nie są naprężone, pobudzenie ma charakter szumowy, a generowana mowa jest bezdźwięczna. Kolejnym rodzajem pobudzenia jest pobudzenie mieszane, charakterystyczne dla fonemów dźwięcznych szczelinowych (np. ż lub w). Ma ono cechy pobudzenia dźwięcznego, ale jednocześnie powstaje szum związany ze zwężeniami jamy ustnej. Ostatnim rodzajem pobudzenia jest pojedynczy impuls, czyli gwałtowna zmiana ciśnienia akustycznego w związku z otwarciem kanału głosowego, charakterystyczna dla fonemów plozyjnych. W niniejszej pracy skoncentrujemy się na analizie wyłącznie dźwięcznych fragmentów mowy.



Rys. 2.2: Sygnał pobudzenia w postaci ciągu impulsów tonu krtaniowego według modelu Rosenberga

Strumień powietrza charakteryzujący się zmiennym ciśnieniem akustycznym przechodzi dalej przez kanał głosowy, składający się z gardła, jamy ustnej, języka i jamy nosowej, tworzących komory rezonansowe. Z punktu widzenia teorii systemów, kanał głosowy zamodelować można układem liniowym o zmiennych w czasie parametrach i odpowiedzi impulsowej $v(n,t)$. Kanał głosowy składa się z 4-5 komór rezonansowych, stąd w module jego funkcji transmitancji widoczne są obszary koncentracji energii zwane formantami reprezentowane m.in. przez częstotliwości rezonansowe. Ich położenie na osi częstotliwości jest zależne od ukształtowania w danej chwili kanału głosowego, sterowanego przez mózg w celu wygenerowania określonej jednostki fonetycznej. Kanał głosowy odpowiada zatem za kształtowanie zawartości semantycznej mowy. Przykładowy moduł transmitancji kanału głosowego w przypadku samogłoski a zaprezentowano na rys. 2.3.



Rys. 2.3: Moduł transmitancji kanału głosowego z formantami odpowiadającymi samogłosce a.

Ostatni komponent modelu Fanta opisujący generację sygnału mowy przez usta, można zamodelować przy użyciu układu różniczkującego, a więc górnoprzepustowego filtru 1. rzędu o transmitancji $H_r(z)$ danej zależnością

$$H_r(z) = 1 - \alpha z^{-1}. \tag{2.6}$$

Zwykle przyjmuje się, że $\alpha = 0.95$ [85].

2.3. Parametryzacja sygnału

Ze względu na wiele czynników mających wpływ na mowę, m.in. jej losowy charakter, dużą zmienność i redundancję, przebieg czasowy nie jest dobrą reprezentacją zawartości fonetycznej sygnału. W tym celu stosuje się parametryzację, nazywaną inaczej ekstrakcją cech (ang. *feature extraction*). Ma ona na celu transformację sygnału w możliwie mało liczny zbiór parametrów, wyodrębnienie informacji istotnych z punktu widzenia rozpoznawania. Spośród wielu dostępnych w literaturze metod parametryzacji sygnału wyróżnić można:

1. **Podejścia bazujące na filtracji sygnału**
   Do tej grupy należą: metoda MRA (ang. *Multiresolution Analysis*), w której sygnał poddawany jest wienerowskiej filtracji odszumiającej, a następnie przy pomocy transformacji falkowej rozfiltrowywany na podpasma [9], metoda EIH (ang. *Ensemble Interval Histogram*), w której organ słuchu człowieka modelowany jest za pomocą 165 filtrów i analizowane są poziomy symulujące aktywność neuronów [25].

2. **Metody bazujące na prognozie liniowej LPC (ang. *Linear Predictive Coding*)**
   W tej grupie wymienić należy metodę LPCC (ang. *Linear Prediction Cepstral Coefficients*), w której zakłada się, że analizowany sygnał jest procesem AR (ang. *autoregressive*) p-tego rzędu, a rozwiązanie problemu prognozy polega na wyznaczeniu zbioru współczynników filtru prognozującego optymalnego w sensie średniokwadratowym. Metody te sprowadzają się do rozwiązania układu równań liniowych Wienera-Hopfa [84] [110]. Modyfikacją metody LPCC, uwzględniającą niektóre właściwości organu słuchu człowieka jest parametryzacja PLP (ang. *Perceptual Linear Prediction*), w której wyznacza się współczynniki prognozy liniowej na podstawie widma sygnału wyrażonego w barkowej skali częstotliwości i poddanego operacjom maskowania w pasmach krytycznych oraz nałożenia krzywej jednakowej głośności [37].

3. **Metody wykorzystujące reprezentacje cepstralne**
   Metody te bazują na cepstrum, czyli odwrotnej transformacji Fouriera z logarytmu widma sygnału. W efekcie analizuje się sygnał w dziedzinie tzw. pseudoczasu (ang. *quefrency*). Motywacją do opracowania tych algorytmów było jak najwierniejsze odwzorowanie ludzkiej percepcji dźwięku. Główna różnica pomiędzy tymi metodami parametryzacji leży w sposobie zaprojektowania odpowiedniego banku filtrów słuchowych, uwzględniającego nieliniową zależność pomiędzy zakresami częstotliwości sygnału mowy, a subiektywnym wrażeniem wysokości dźwięku. Do tej grupy metod zaliczyć można: (i) LFCC (ang. *Linear-Frequency Cepstral Coefficients*), w której współczynniki cepstralne obliczane są na podstawie widma amplitudowego w liniowej skali częstotliwości, (ii) MFCC (ang. *Mel-Frequency Cepstral Coefficients*) [11] oraz HFCC (ang. *Human Factor Cepstral Coefficients*) [94], wykorzystujące melową skalę częstotliwości, (iii) BFCC (ang. *Bark Frequency Cepstral Coefficients*), która powstała na bazie parametryzacji MFCC i PLP z wykorzystaniem skali barkowej [87]

[65], (iv) wiele innych wyrafinowanych metod uwzględniających percepcję dźwięku przez układ słuchowy i mózg człowieka, np. metodę naśladującą działanie błony podstawnej w uchu wewnętrznym, która w literaturze również opatrzona jest akronimem BFCC (ang. *Basiliar-membrane Frequency-band Cepstral Coefficient*) [51], wykorzystujących bank filtrów gammatone - GTCC (ang. *Gammatone Cepstral Coefficient*) [106] lub jedno z najnowszych rozwiązań, PWCC (ang. *Power Wavelet Cepstral Coefficients*) z użyciem banku filtrów NWFB (ang. *Normalized Wavelet FilterBank*), opartego na falkach Morleta [114].

Metody bazujące na filtracji sygnału nie są powszechnie wykorzystywane, ze względu na dużą złożoność obliczeniową. Dodatkowo, porównywalną jakość parametryzacji można uzyskać innymi, nowszymi metodami [41]. Metody LPC źle sprawdzają się w warunkach szumu i dużych wartości częstotliwości tonu krtaniowego oraz w małym stopniu uwzględniają właściwości słuchowe człowieka. W literaturze najbardziej popularne i powszechnie stosowane są algorytmy bazujące na reprezentacjach widmowych i cepstralnych. Badania dowodzą, że są one bardziej odporne na szum, efektywne i przydatne w praktycznych zastosowaniach, i to właśnie na tej grupie metod bazuje niniejsza praca.

2.3.1. Odporna parametryzacja

Używanie systemu ARM w innych warunkach niż te, w których został zaprojektowany, skutkuje znacznym obniżeniem jego skuteczności. Sygnał mowy charakteryzuje się bardzo dużą zmiennością i losowością, zatem w tych systemach zachodzi konieczność uwzględnienia wielu czynników, takich jak:

- techniczne warunki rejestracji, w tym liniowe i nieliniowe zniekształcenia transmisyjne,
- zmienność wewnątrzosobnicza spowodowana m.in. nastrojem, chwilowymi stanami emocjonalnymi, stanem zdrowia,
- zmienność międzyosobnicza spowodowana różnicami płci, wieku, osobowości, wykształcenia czy budowy organu mowy,
- różnice regionalne, kulturowe i środowiskowe,

które negatywnie wpływają na jakość rozpoznawania. Straty w skuteczności działania systemu mogą być kompensowane poprzez [57]:

1. **odporną parametryzację** (ang. *robust feature extraction*) polegającą na takiej ekstrakcji cech, aby wpływ ww. czynników na mowę był jak najmniejszy lub zredukowany całkowicie [69] [70].

2. **grupowanie mówców** (ang. *speaker clustering*), polegające na budowaniu osobnych modeli statystycznych dla grup mówców charakteryzujących się podobnymi cechami osobniczymi [58],

3. **normalizację** (ang. *normalization*), polegającą na modyfikacji wartości współczynników parametryzacji aktualnego mówcy tak, aby upodobnić je do parametrów uśrednionego mówcy zbioru uczącego [81],

4. **adaptację** (ang. *adaptation*), czyli zmiany parametrów modelu GMM bez modyfikacji wartości współczynników parametryzacji [109].

Odporna parametryzacja to taka, dla której wynikowe zbiory wektorów obserwacji tego samego fonemu charakteryzują się małą wariancją pod wpływem wyżej wymienionych czynników. Poglądowy schemat podsystemu akustycznego z odporną parametryzacją w systemie ARM zaprezentowano na rys. 2.4. Klasyczne metody parametryzacji cepstralnych, poprzez zastoso-

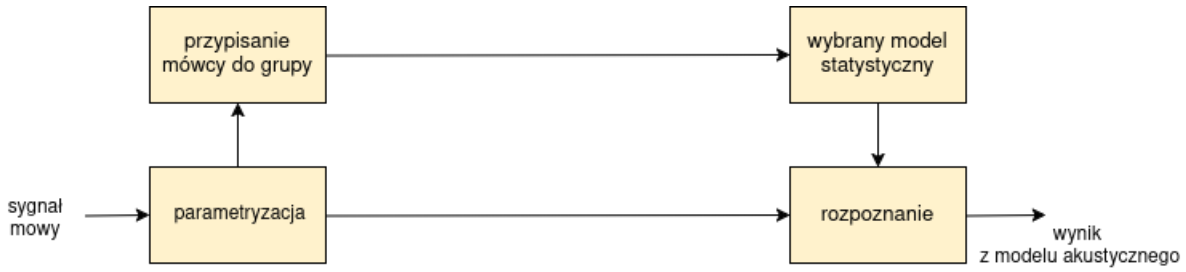
Rys. 2.4: Fragment podsystemu akustycznego z odporną parametryzacją w systemie ARM.

wanie filtrów melowych i uśredniania w pasmach częstotliwości, posiadają w sposób naturalny pewne mechanizmy odporności na niewielkie zakłócenia szumowe i zmienność warunków, jednak nie można ich nazwać metodami odpornymi. Można je natomiast uzupełnić o algorytm RASTA (ang. *Relative Spectra*), który teoretycznie usuwa te składowe, które nie są związane z artykulacją mowy. W oparciu o takie rozwiązanie powstał algorytm hybrydowy RASTA-PLP [50].

W kategoriach odpornej parametryzacji rozpatrywać można także algorytm MVDR-MFCC (ang. *Minimum Variance Distortionless Response Mel Frequency Cepstral Coefficients*), który bazuje nie na klasycznym estymatorze DFT zwanym periodogramem, lecz na estymatorze minimalnej wariancji MDVR. Jest to estymator parametryczny, dzięki któremu możliwe jest znaczne zredukowanie wpływu przecieku widma oraz zmniejszenie wariancji współczynników cepstralnych poprzez zastosowanie wygładzania w dziedzinie czasu. Metoda jest kosztowna obliczeniowo, jednak w badaniach literaturowych wykazano przewagę tej metody względem parametryzacji MFCC, gdyż błędy rozpoznawania komend w różnych warunkach zostały zredukowane o 40%. W literaturze istnieje kilka modyfikacji metody MVDR-MFCC mających na celu zmniejszenie złożoności obliczeniowej. Są to m.in. algorytmy PMVDR (ang. *Perceptual MVDR*) oraz zmodyfikowana metoda PLP z zastosowaniem estymatora widma MVDR.

Wśród pozostałych metod kompensacji, wymienionych w tym rozdziale, warto wspomnieć o **grupowaniu mówców**. Poglądowy schemat podsystemu akustycznego, wykorzystującego grupowanie mówców, przedstawiono na rys. 2.5. Spośród wielu parametrów, na podstawie których można podzielić mówców na grupy, warto wyróżnić ich cechy osobnicze, w tym parametry ka-

nału głosowego. Jednym z nich jest częstotliwość tonu krtaniowego F_0 , którą można traktować jako wskaźnik wielkości kanału głosowego, gdyż jest skorelowana z długością strun głosowych mówcy [57]. Jako podstawę do grupowania traktować można długości części ustnej i gardłowej kanału głosowego, ponieważ mają one bezpośredni wpływ na położenie formantów na osi częstotliwości [71], a także współczynniki parametryzacji, skupiając się na maksymalizacji odległości pomiędzy wielowymiarowymi rozkładami prawdopodobieństw wektorów cech w sensie wybranej miary odległości. Podziału można dokonać także w sposób hierarchiczny, wykorzystując więcej niż jeden czynnik odróżniający mówców od siebie. Przykładem takiego podejścia może być chociażby grupowanie na podstawie płci i szybkości mówienia [35]. Jednym z nowszych algorytmów zaproponowanych w literaturze jest podejście bazujące na adaptacji wag modelu UBM (ang. *Universal Background Model*) zaproponowane w pracy [58].

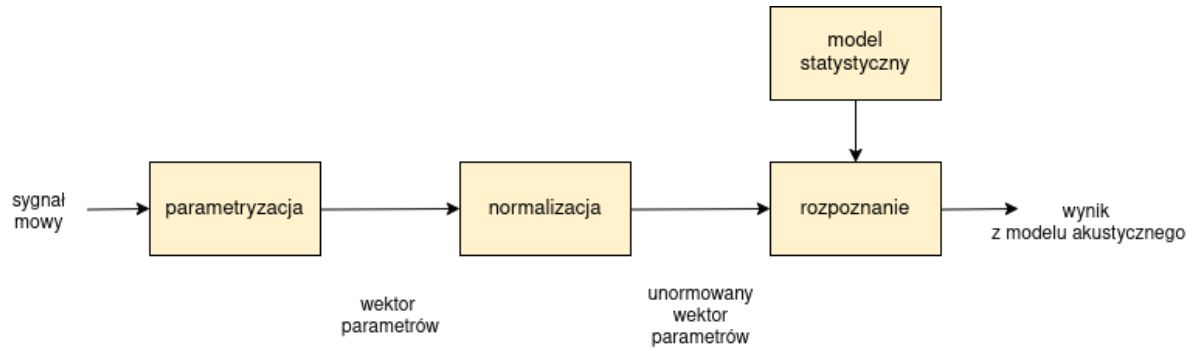


Rys. 2.5: Fragment podsystemu akustycznego z grupowaniem mówców w systemie ARM.

Do algorytmów normalizacji (rys. 2.6) należą rozwiązania oparte na aproksymacji funkcji zniekształceń wprowadzanych przez tor transmisyjny. Są to zniekształcenia inercyjne, a zakłócenia przyjmują postać szumu addytywnego, normalizacja natomiast realizowana jest w dziedzinie logarytmu widma mocy sygnału. Głównym celem takiego podejścia jest wyznaczenie parametrów rozkładów logarytmu widma mocy sygnału zniekształconego i zakłóconego oraz logarytmu widma mocy czystego sygnału mowy. W celu uwzględnienia nieliniowego charakteru tych zależności stosuje się rozwinięcia ww. wielkości w szereg Taylora. Metoda ta nosi nazwę VTS (ang. *Vector Taylor Series*) [66], a jej późniejsza modyfikacja - DVTS (ang. *Delta Vector Taylor Series*) [79]. W badaniach eksperymentalnych przeprowadzonych w zacytowanych wyżej pracach udowodniono efektywność tej metody, uzyskując znacznie lepsze wyniki rozpoznawania pojedynczych słów.

Z racji tego, że cechy osobnicze, takie jak wielkość kanału głosowego, mają wpływ na częstotliwości formantowe, a co za tym idzie, na widmo sygnału mowy, stosuje się metody normalizacji długości kanału głosowego VTLN (ang. *Vocal Tract Length Normalization*). Polegają one na skalowaniu osi częstotliwości przy użyciu rozmaitych funkcji, którymi moga być przekształcenia liniowe, dwuparametryczne, dwuliniowe lub przekształcenia opisane linią łamaną [69, 63, 103]. Podejście to można stosować także w rozpoznawaniu mowy dziecięcej przy użyciu systemów ARM dedykowanych do pracy z głosami dorosłych mówców. Niedopasowanie akustyczne spowodowane odmienną długością kanału głosowego między dorosłymi a dziećmi

stanowi kluczowy czynnik obniżający skuteczność systemu ARM. W takich przypadkach zaproponowano użycie metody VTLN i modyfikację formantów mowy dziecięcej z wykorzystaniem widma LP (ang. *Linear Prediction*) i funkcji korygującej w dziedzinie częstotliwości (ang. *frequency warping*). Dzięki temu możliwa jest synteza mowy o strukturze formantowej zbliżonej do formantów dorosłych mówców [46].



Rys. 2.6: Fragment podsystemu akustycznego z normalizacją wartości wektora parametrów w systemie ARM.

Jednym z najbardziej popularnych i powszechnie stosowanych algorytmów normalizacji jest CMVN (ang. *Cepstrum Mean Variance Normalization*), stosowany m.in. do kompensacji wpływu zniekształceń liniowych wprowadzanych przez tor transmisyjny oraz uodpornienia wektora parametrów na wpływ szumu [100]. Metoda ta polega na standaryzacji zbioru obserwacji wypowiedzi. W wyniku otrzymuje się ustandaryzowany wektor współczynników $c^{(s)}(t,p)$:

$$c^{(s)}(t,p) = \frac{c(t,p) - m_c(p)}{\sigma_c(p),}, \tag{2.7}$$

gdzie $t = 1, ..., T$ jest indeksem ramki sygnału, a T liczbą ramek, zaś $p = 0, ..., P-1$ indeksem współczynnika cepstralnego, a P liczbą tych współczynników. Wartość średnia $m_c(p)$ zbioru obserwacji dana jest zależnością:

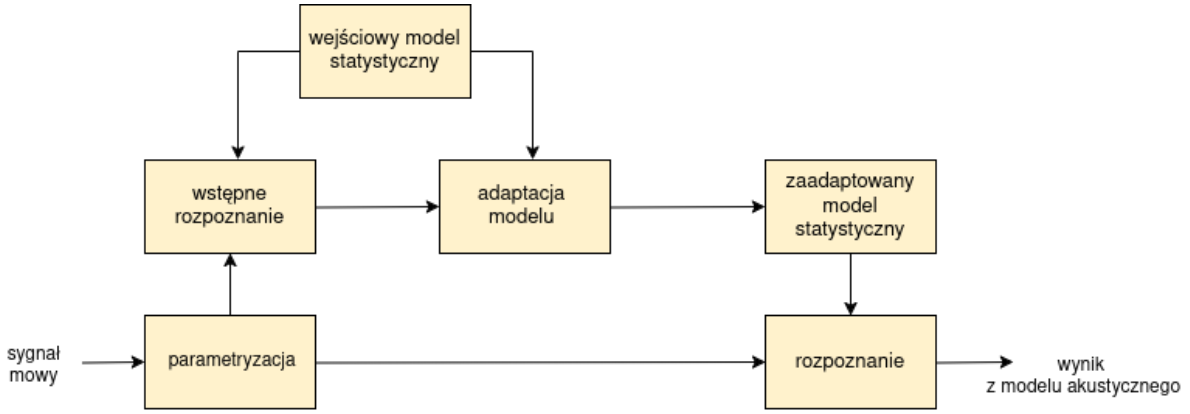
$$m_c(p) = \frac{1}{T}\sum_{t=1}^{T} c(t,p), \tag{2.8}$$

a $\sigma_c(p)$ jest odchyleniem standardowym zbioru obserwacji:

$$\sigma_c^2(p) = \frac{1}{T}\sum_{t=1}^{T}[c(t,p) - m_c(p)]^2. \tag{2.9}$$

Zakładając, że algorytm ma pracować w czasie rzeczywistym, stosuje się modyfikacje wyżej wyznaczonych wielkości w oparciu o uśrednianie statystyk na podstawie sąsiedztwa kolejnych ramek sygnału, a w celu uwzględnienia nieliniowego wpływu szumu na wartości współczynników cepstralnych (w tym nie tylko na wartości średnie i wariancje, ale również na statystyki wyższych rzędów) stosuje się nieliniowe korekcje realizowane przy pomocy algorytmu CTN (ang. *Cepstrum Third-Order Normalization*) [96] lub normalizacje rozkładów prawdopodobieństw [97].

Algorytmy adaptacji polegają na zmianie parametrów modelu statystycznego mowy w celu jego dopasowania do aktualnego mówcy. Poglądowy schemat takiego podejścia przedstawiono na rys. 2.7. W adaptacji wykorzystuje się m.in. metodę regresji liniowej i maksymalizacji prawdopodobieństwa MLRR (ang. *Maximum Likelihood Linear Regression*) [54] wraz z jej wersją z ograniczeniami - cMLLR (ang. *constrained MLLR*) [16], metodę EVSA (ang. *Eigenvoice Speaker Adaptation*), bazującą na adaptacji w przestrzeni wektorów własnych [52], metodę MAP (ang. *Maximum a posteriori*) nazywana także metodą Bayesa [23], a także jej późniejsze modyfikacje - metody RMP (ang. *Regression-Based Model Prediction*) [105], SMAP (ang. *Structural Maximum a posteriori*), które przyspieszają proces adaptacji i pozwalają przewidywać przekształcenia dla stanów, które nie zostały wystarczająco dobrze odwzorowane w dostępnych danych obserwacyjnych [92, 93]. Kolejne modyfikacje klasycznej metody MAP zawierają elementy progowania, adaptację z przesunięciem i zbliżeniem do siebie średnich, adaptacje z dodatkowymi składnikami mikstury modelu statystycznego oraz inne, opisane w pracy [109].



Rys. 2.7: Fragment podsystemu akustycznego z adaptacją modelu statystycznego w systemie ARM.

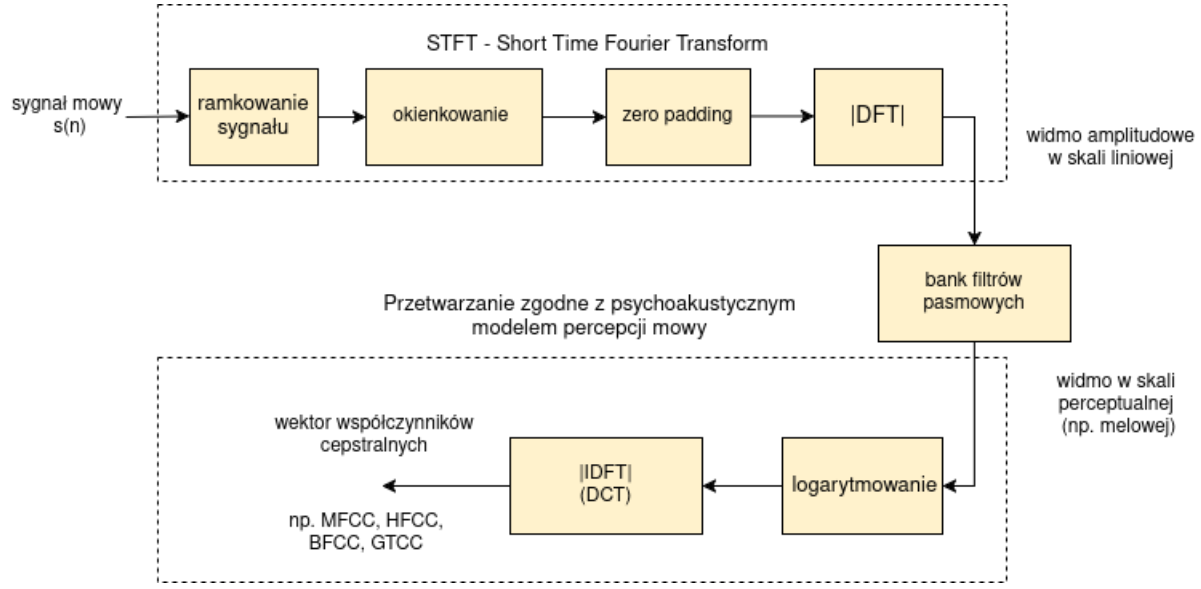
2.3.2. Schemat parametryzacji HFCC

Zgodnie z celem pracy wskazanym w rozdz. 1.2, postawionym zadaniem jest opracowanie takiej metody parametryzacji, aby uczynić ją bardziej odporną na zmienność międzyosobniczą i wewnątrzosobniczą. Schemat wyznaczania współczynników cepstralnych w klasycznej parametryzacji HFCC przedstawiono na rys. 2.8.

W pierwszym kroku sygnał mowy $s(n)$ dzielony jest na ramki, tj. fragmenty o długości N_r i przesunięciu N_s. Zakładając lokalną stacjonarność sygnału, uznaje się, że w obrębie jednej ramki statystyki mowy się nie zmieniają. Operację wydzielenia ramek $z(t,n)$ z sygnału $s(n)$ określa zależność (2.10).

$$z(t,n) = s(tN_s + n)w(n), \qquad t = 0, \ldots, T-1, \tag{2.10}$$

gdzie t oznacza kolejną ramkę sygnału, T jest liczbą ramek, n - indeksem próbki, $w(n)$ - funkcją okna czasowego. Operacja ramkowania jest tożsama z przemnożeniem sygnału przez okno



Rys. 2.8: Schemat obliczania współczynników cepstralnych.

prostokątne, jednak w systemach ARM w celu minimalizacji wpływu przecieku widma stosuje się inne funkcje okien czasowych. W niniejszej pracy przyjęto długość ramki $N_r = 30$ ms, przesunięcie (skok) $N_s = 10$ ms oraz okno Hamminga, wyrażone zależnością (2.11).

$$w(n) = \begin{cases} 0,54 - 0,46 \cdot cos(\frac{2\pi n}{N_r}); & n = 0, ..., N_r - 1 \\ 0; & \text{poza tym.} \end{cases} \tag{2.11}$$

Ponieważ faza sygnału mowy nie ma wpływu na rozpoznawanie, kolejnym krokiem parametryzacji jest obliczenie widma ramki sygnału. z wykorzystaniem Dyskretnej Transformaty Fouriera (DFT), zgodnie z zależnością:

$$|Z(t,l)| = \sqrt{|\mathcal{F}\{z(t,n)\}|^2}, \qquad l = 0, \ldots, L, \tag{2.12}$$

gdzie $\mathcal{F}\{\cdot\}$ oznacza dyskretną transformatę Fouriera, l jest indeksem prążka w dziedzinie częstotliwości, L liczbą prążków widma.

Uwaga.

Ze względu na możliwe dwie interpretacje sygnału mowy, konieczne jest precyzyjne zdefiniowanie pojęcia widma, które będzie wykorzystywane w dalszej części pracy:

- w przypadku analizy pojedynczej, konkretnej realizacji sygnału o skończonym czasie trwania – np. ramki czasowej sygnału mowy o długości 30 ms – sygnał $z(t,n)$ można traktować jako przebieg deterministyczny. Wówczas jego reprezentacją wykorzystywaną w procesie obliczania współczynników cepstralnych jest widmo amplitudowe $|Z(t,l)|$, wyznaczane bezpośrednio za pomocą dyskretnej transformaty Fouriera (DFT),
- w przypadku transformacji zbioru ramek sygnału mowy, pochodzących z różnych nagrań wielu mówców, dla danego fonemu sygnał należy traktować jako proces stochastyczny. Wówczas pojedyncza ramka sygnału stanowi fragment realizacji procesu i zasadne staje się odwołanie do pojęcia widmowej gęstości mocy.

W praktyce obliczeniowej, dla procesów stacjonarnych i ergodycznych widmowa gęstość mocy P_{xx} może być wyznaczona z zależności [113]:

$$P_{xx}(f) = \lim_{N\to\infty} E\left[\frac{1}{2N+1}|Z(t,l)|^2\right], \tag{2.13}$$

gdzie E jest operatorem wartości oczekiwanej. Przy skończonej liczbie próbek N, wyrażenie $\frac{1}{2N+1}|Z(t,l)|^2$, nazywane periodogramem, stanowi estymator widmowej gęstości mocy. W związku z tym, wyrażenie $|Z(t,l)|$ jest pierwiastkiem widmowej gęstości mocy, zgodnie z zależnością 2.12.

Pomimo formalnego rozróżnienia, w dalszej części pracy pojęcie widma ramki sygnału (lub krócej: widma) będzie używane zamiennie.

Dalej, aby uwzględnić nieliniowość postrzegania częstotliwości dźwięku przez człowieka, widmo w skali liniowej przeliczane jest na melową skalę częstotliwości poprzez przemnożenie go przez bank filtrów komplementarnych o trójkątnych charakterystykach amplitudowych. W klasycznej parametryzacji MFCC widmo przelicza się na skalę melową, w której częstotliwości środkowe filtrów do 1 kHz rozłożone są liniowo na osi częstotliwości, a dalej tworzą szereg geometryczny, zwykle o wykładniku $\kappa = 1,149$. Warto mieć na uwadze, że modyfikacji podlegać może wiele różnych parametrów banku filtrów, m.in.: (i) rodzaj aproksymacji skali melowej i rozłożenie częstotliwości środkowych, (ii) kształt, szerokości i wzmocnienie poszczególnych filtrów, (iii) zakres pasma częstotliwości przyjęty do analizy np. rezygnacja z pierwszych pasm melowych w celu wyeliminowania wpływu zakłóceń dolnopasmowych.

W parametryzacji HFCC wykorzystuje się bank filtrów trójkątnych z częstotliwościami środkowymi równomiernie rozłożonymi zgodnie z aproksymacją Fanta melowej skali częstotliwości:

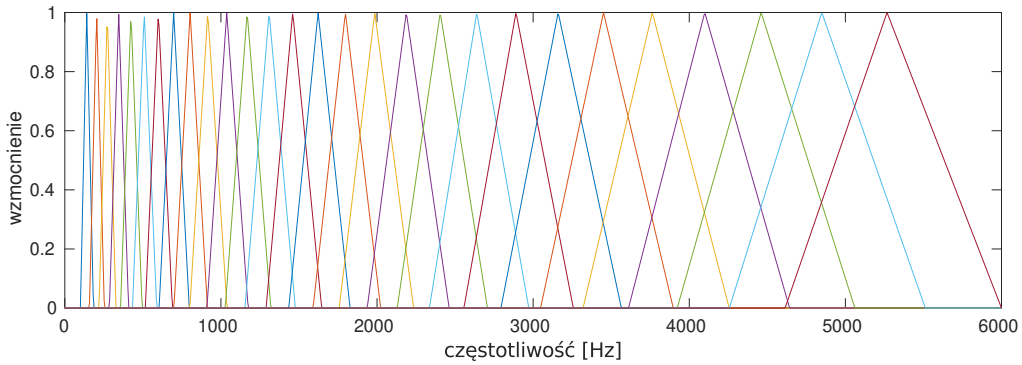
$$\bar{f} = 2595 \log_{10} 1 + \frac{f}{700}, \tag{2.14}$$

gdzie f jest częstotliwością wyrażoną w Hz, a $\bar{f}$ częstotliwością w skali melowej. Główną różnicą względem parametryzacji MFCC jest fakt, że szerokości poszczególnych okien trójkątnych e stanowią niezależny parametr projektowania algorytmu, a nie są zdeterminowane przyjętym zakresem częstotliwości i liczbą filtrów. Wyrażone są one w skali ERB (ang. *Equivalent Rectangular Bandwidth*), będącej funkcją częstotliwości środkowych filtrów [67]. Aproksymację Moore'a i Glasberga tej skali wyraża zależność (2.15).

$$e = 6.23 \cdot 10^{-6} f_c^2 + 93.39 \cdot 10^{-3} f_c + 28.52, \tag{2.15}$$

gdzie f_c oznacza częstotliwości środkowe filtrów wyrażone w Hz. Bank filtrów wykorzystywany w parametryzacji HFCC zaprezentowano na rys. 2.9.

Badania pokazują [94], że zastosowana aproksymacja w parametryzacji HFCC wykazuje większą odporność na szum względem innych przyjętych metod parametryzacji, w tym szczególnie względem MFCC.



Rys. 2.9: Bank filtrów wykorzystywany w stosowanej parametryzacji HFCC

Przeliczenie widma na skalę melową odbywa się według zależności:

$$Y^l(t,j) = log(\sum_{l=0}^{L} |Z(t,l)|\, G(l,j)), \qquad j = 1, ..., J \tag{2.16}$$

gdzie $G(l,j)$ jest zbiorem filtrów komplementarnych, j jest indeksem okna częstotliwościowego, J - liczbą tych okien, a zarazem podpasm częstotliwościowych w skali melowej i prążków widma otrzymanego w wyniku tej operacji, a log - logarytmem dziesiętnym. Logarytmowanie służy uwzględnieniu nieliniowości postrzegania intensywności dźwięku przez ucho ludzkie. W tym miejscu można zastosować inne funkcje nieliniowe, np. pierwiastki określonego stopnia. Ostatnim elementem parametryzacji jest odwrotna transformata Fouriera IDFT (ang. *Inverse Discrete Fourier Transform*), w wyniku której otrzymuje się wektor współczynników cepstralnych $c(t,p)$:

$$c(t,p) = \sum_{j=1}^{J} Y_l(t,j) cos\left(p\left(j - \frac{1}{2}\right)\frac{\pi}{J}\right), \qquad p = 1, ..., P, \tag{2.17}$$

gdzie p jest numerem wyznaczanego współczynnika, a P liczbą tych współczynników. Ostatnią operację można także interpretować jako dyskretną transformatę cosinusową DCT (ang. *Discrete Cosine Transform*). Autorzy parametryzacji HFCC w pracy [94] wskazują, że operacja ta powoduje częściową dekorelację współczynników parametryzacji. Nie ma ona, z biologicznego punktu widzenia, odniesienia w działaniu słuchu człowieka. Jest więc zabiegiem sztucznym, jednak świetnie sprawdza się, gdy na dalszych etapach przetwarzania stosowane są algorytmy i klasyfikatory wymagające przynajmniej częściowej dekorelacji współczynników lub wykorzystujące diagonalne macierze kowariancji. W tym miejscu warto także wspomnieć o kompresji informacji. Z uwagi na fakt, że transformacja DCT ma zdolność do kumulowania energii w pasmach niskich częstotliwości, zwykle do dalszej analizy wykorzystuje się pierwsze kilkanaście współczynników; ich liczba $P < J$ jest mniejsza od liczby melowych pasm częstotliwości.

2.3.3. Wyznaczenie wektora obserwacji

Poza współczynnikami cepstralnymi, wyznaczonymi wg. zależności (2.17), wektor cech stanowiący reprezentację ramki sygnału mowy w systemie ARM można uzupełnić o kilka para-

metrów, dzięki którym opis i analiza sygnału będą pełniejsze. Zerowy współczynnik cepstralny nazywany jest wskaźnikiem głośności. Z racji, że postrzeganie głośności nie jest liniową funkcją intensywności dźwięku, w ujęciu sygnałowym zaproponowano wiele różnych definicji wskaźników głośności, m.in. obliczając współczynnik $c(t,0)$ z zależności (2.17), otrzymujemy logarytm energii ramki sygnału mowy. W celu zmniejszenia złożoności obliczeniowej, sumę kwadratów można zastąpić sumą wartości bezwzględnych i zrezygnować z operacji logarytmowania:

$$e(t) = \sum_{n=0}^{Nr-1} |z(t,n)|. \tag{2.18}$$

Jest to istotny parametr w analizie sygnału mowy, gdyż poszczególne jednostki fonetyczne charakteryzują się znacznie zróżnicowanym poziomem względnej intensywności.

Ludzki słuch jest analizatorem widma sygnału i powszechnie wiadomo, że jest bardziej wyczulony na zmiany odbieranego sygnału, niż jego bezwzględny poziom, tak więc parametry uwzględniające dynamikę zmian mowy powinny dotyczyć rownież widma sygnału. W tym celu wektor obserwacji rozszerza się o zbiór parametrów dynamicznych - pochodne I i II rzędu (ang. *delta cepstrum*). W literaturze można spotkać bardzo wiele definicji tych parametrów [6][13], wśród nich wyróżnić warto współczynniki regresji liniowej [21]:

$$c^{(\Delta)}(t,p) = \eta \sum_{i=-I}^{I} ic(t+i,p), \tag{2.19}$$

gdzie η jest współczynnikiem skalującym opisanym zależnością:

$$\eta = \sum_{i=-I}^{I} i^2, \tag{2.20}$$

a parametr I definiuje zakres sąsiadujących ze sobą ramek poddawanych analizie [84]. Pierwsze pochodne opisują szybkość zmian współczynników parametryzacji, co pozwala na uwzględnienie stanów przejściowych pomiędzy jednostkami fonetycznymi. W analogiczny sposób definiuje się drugie pochodne współczynników parametryzacji $c^{(\Delta^2)}(t,p)$, które mogą być przydatne w analizie wypowiedzi zróżnicowanej emocjonalnie lub prozodycznie. Parametry dynamiczne są ważne także przy analizie fonemów plozyjnych, charakteryzujących się dużymi zmianami ciśnienia akustycznego i turbulentnymi przepływami powietrza.
W wyniku parametryzacji każda z analizowanych ramek sygnału mowy otrzymuje swoją reprezentację nazywaną wektorem obserwacji lub wektorem cech o_t

$$o_t = \left[o_{t,1} \ldots o_{t,D}\right]^T, \tag{2.21}$$

gdzie indeks t jest numerem ramki, indeks d numerem współczynnika wektora obserwacji, a D liczbą wszystkich współczynników.

W niniejszej pracy $\mathbf{O}$ oznacza zbiór wektorów obserwacji dla wszystkich nagrań. Wyróżnić w nim można dwa podzbiory:

$$\mathbf{O}_T = \{o_t : t = 1, ..., T\}, \tag{2.22}$$

który oznacza sekwencję wszystkich wektorów obserwacji wyznaczonych dla danej wypowiedzi lub jej fragmentu, a także:

$$\mathbf{O}_f = \{o_t \in \mathbf{O}_T | f(t) = f\}, \quad f = 1, ..., F, \tag{2.23}$$

będący zbiorem wektorów obserwacji o_t przypisanych fonemowi f, gdzie F jest liczbą fonemów.

2.4. Architektury nowoczesnych systemów ARM

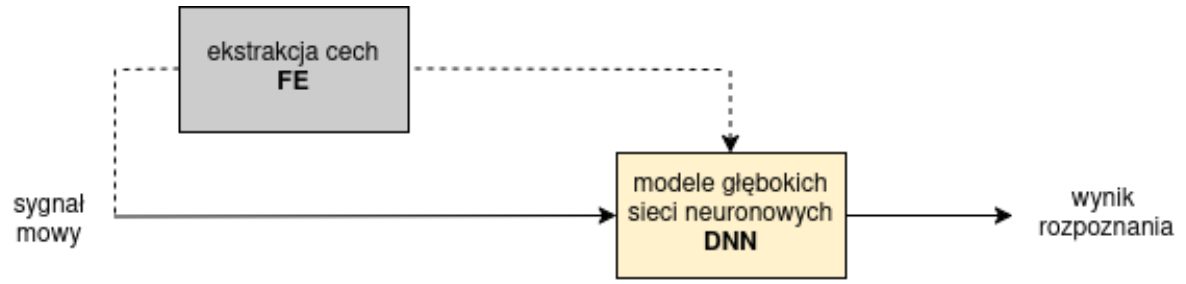
W ciągu ostatnich dziesięcioleci obserwuje się ewolucję systemów automatycznego rozpoznawania mowy - od klasycznych architektur konwencjonalnych, poprzez rozwiązania hybrydowe, łączące modele statystyczne z sieciami neuronowymi, aż po współczesne systemy oparte na modelach głębokich i architekturach typu End-to-End, związane z dynamicznym rozwojem sztucznej inteligencji. Wyróżnia się zatem następujące rozwiązania w projektowaniu systemów ARM:

- **podejścia klasyczne**, w których modele AM, LM i PM wymieniają między sobą informacje, ale są trenowane i optymalizowane oddzielnie. Podejścia te, do obliczania prawdopodobieństw $P(\mathbf{O}|\mathbf{Q})$ (blok FC) oraz $P(\mathbf{Q}|\mathbf{L})$ (model SM) wykorzystują modele statystyczne HMM-GMM, czyli ukryte modele Markova (ang. *Hidden Markov Models*) i probabilistyczny model GMM (szczegółowo opisany w rozdz 4.1.3) [15][44][107],

- **podejścia hybrydowe**, w których statystyczny model GMM został zastąpiony sztucznymi sieciami neuronowymi ANN (ang. *Artifical Neural Network*) i modelami głębokimi DNN-HNN (ang. *Deep Neural Network*) przy jednoczesnym zachowaniu wszystkich ww. komponentów systemu ARM [107][10][39][1] [98]. Taka koncepcja została zapisana w patentach Google LLC US8442821B1 [77] i US8442821B1 [78], w których jako wektory cech w modelach uczenia głębokiego wykorzystywane są m.in. współczynniki PLP lub MFCC.

- **głębokie modele E2E** (ang. *End-To-End*), w których odchodzi się od konwencjonalnych rozwiązań opartych o analizę akustyczną, struktury językowe i leksykalne, a operuje się na zintegrowanym modelu głębokich sieci neuronowych działającym na surowych danych wejściowych z przynajmniej częściowym pominięciem etapu ekstrakcji cech. Takie rozwiązanie przedstawiono na rys. 2.10. Dzięki temu możliwe jest wdrażanie systemów rozpoznawania niezależnych od konkretnego języka. Pomimo tego, że systemy E2E osiągają świetne wyniki dokładności i są stosowane w najnowocześniejszych rozwiązaniach, trzeba mieć na względzie, że taka architektura wymaga ogromnych mocy obliczeniowych i trenowania na bardzo dużych zbiorach danych. W związku z tym odpowiednio zoptymalizowane klasyczne systemy ARM, nad którymi pracowano od dziesięcioleci nadal są powszechnie wykorzystywane i rozwijane.

W architekturach typu E2E pierwotnie stosowano rekurencyjne sieci neuronowe RNN [91] przeznaczone do analizy danych sekwencyjnych (takich jak mowa), ponieważ zachowują informacje kontekstowe i przydatne są przy rozpoznawaniu długich i skomplikowanych sekwencji



Rys. 2.10: Schemat architektury End-To-End systemów ARM. Blok ekstrakcji cech jest w niektórych przypadkach zintegrowany wewnątrz modelu DNN, jednak najczęściej nawet w wewnętrznych reprezentacjach modeli głębokich dokonywana jest parametryzacja sygnału.

dźwiękowych. Do mapowania surowego sygnału audio na transkrypcję tekstową wykorzystywane są także różne ich warianty, np. LSTM (Long Short-Term Memory) [90] czyli sieci długiej pamięci krótkotrwałej, które są zdolne do pamiętania długich sekwencji i informacji o poprzednich stanach oraz wykorzystywania ich do bardziej precyzyjnego rozpoznawania kolejnych fragmentów wypowiedzi.

W architekturach E2E wykorzystywane są także sieci konwolucyjne CNN (ang. *Convolutional Neural Network*) przeznaczone do analizy obrazów [1] [112]. Stosuje się je do rozpoznawania mowy z wykorzystaniem graficznej reprezentacji sygnału postaci spektrogramu lub mel-spektrogramu [108]. Taką reprezentację można uzyskać, wykorzystując transformacje czasowo-częstotliwościowe STFT (ang. *Short-Time Fourier Transform*) wraz z analizą cepstralną sygnału.

W modelach opartych o mechanizm CTC (ang. *Connectionist Temporal Classification*) [34] sieci neuronowe są w stanie bezpośrednio przetwarzać mowę ciągłą na tekst bez konieczności wcześniejszej segmentacji danych, co znajduje szczególne zastosowanie w analizie i rozpoznawaniu nagrań mowy o zmiennej długości i skomplikowanej strukturze. Technologia ta została opisana w patencie Google LLC US10229672B1 [75], dotyczącym metody trenowania modeli akustycznych (z wykorzystaniem wektorów obserwacji postaci współczynników MFCC). Modele CTC bardzo często działają w połączeniu z konwolucyjnymi sieciami neuronowymi. Rozwinięciem tego podejścia są sieci typu Transducer, tj. RNN-T (ang. *Recurrent Neural Network Transducer*) [33], które dzięki wysokiej dokładności i niskim opóźnieniom przetwarzania używane są do rozpoznawania mowy w czasie rzeczywistym.

Najnowsze architektury typu Transformer [17] oparte o modele AED (ang. *Attetnion-Based Encoder-Decoder*) [111] [3] wykorzystują:

- **enkodery**, które przetwarzają wejściowy sygnał dźwiękowy na wysokopoziomową reprezentację wewnętrzną. Najpierw jednak wyznaczają one wektory cech akustycznych na podstawie surowych danych wejściowych, stosując powszechnie znane metody parametryzacji opisane w rozdz. 2.3,

- **dekodery** generujące wyjściową sekwencję etykiet (fonemów, słów lub innych jednostek językowych) na podstawie wysokopoziomowej reprezentacji enkoderów. Działają w sposób autoregresyjny, wykorzystując informację z poprzednich prognoz,
- **mechanizmy uwagi** (ang. *multi-head attention*) lub **samouważności** (ang. *self-attention*) [99], które dzięki równoległemu przetwarzaniu informacji pozwalają na bardziej precyzyjne, wydajne i szybsze rozpoznawanie. Mechanizmy te umożliwiają modelowi skupienie się na istotniejszych fragmentach sygnału dźwiękowego (niezależnie od zależności czasowych pomiędzy nimi), zamiast jednolitej analizy całego sygnału. Takie przypisanie odpowiednich wag do poszczególnych fragmentów mowy pozwala na analizę sygnału bez konieczności jego wcześniejszej segmentacji. Mechanizm ten poprawia znacznie zdolność modelu do rozumienia kontekstu poprzez analizę sąsiedztwa danej jednostki fonetycznej, co pozwala na lepszą interpretację poszczególnych fraz i zwiększenie odporności systemu na negatywny wpływ zmienności sygnału spowodowanej różnicami kontekstowymi. Takie podejście zostało opisane w patencie Google LLC US10452978B2 [76].

Podsumowując, blok ekstrakcji cech FE stanowi fundament zarówno tradycyjnych, jak i nowoczesnych architektur systemów ARM. W podejściach klasycznych (HMM-GMM) i hybrydowych (HMM-DNN) parametryzacja sygnału realizowana jest w oparciu o klasyczne rozwiązania szczegółowo opisane w rozdz. 2.3. W architekturach typu E2E, choć proces ekstrakcji cech jest zintegrowany (całkowicie lub częściowo) wewnątrz głębokich sieci neuronowych, nadal stosowane są wewnętrzne reprezentacje w postaci wektorów cech akustycznych (w enkoderach) lub spektrogramy i mel-spektrogramy (w sieciach CNN).

Parametryzacja sygnału pozostaje więc kluczowym etapem mającym wpływ na dokładność i efektywność systemów ARM. Uzasadnionym pozostaje więc fakt konieczności dalszego poszukiwania metod odpornej parametryzacji, w celu niwelowania negatywnego wpływu czynników wymienionych w rozdz. 2.3.1. Warto podkreślić, że tak korzystne modyfikacje reprezentacji sygnału mowy przyczyniają się także do zmniejszenia wymagań dotyczących architektur sieci RNN, CNN w głębokich modelach i systemach E2E poprzez zmniejszenie ich złożoności obliczeniowej, stanowiącej kluczowy aspekt projektowania tych architektur.

Rozdział 3

Problem badawczy i autorskie rozwiązania w zakresie odpornej parametryzacji

W rozdziale skoncentrowano się na analizie problemu badawczego tej pracy, czyli wpływu zmienności częstotliwości tonu krtaniowego na wartości wyznaczanych współczynników parametryzacji. W drugiej części opisano szczegółowo własne propozycje kompensacji tego negatywnego wpływu. Niniejsze rozważania stanowią zasadniczy wkład własny autora i są podstawą badań eksperymentalnych zaprezentowanych w kolejnych rozdziałach.

3.1. Wpływ częstotliwości tonu krtaniowego na współczynniki parametryzacji

Przedmiotem badań niniejszej pracy są dźwięczne fragmenty mowy w postaci sygnału dyskretnego, dla których model pobudzenia $x(n)$ przybiera formę ciągu impulsów:

$$x(n) = g(n) \star p(n) \approx \sum_{k=0}^{+\infty} g(nT_s - kT_0), \tag{3.1}$$

a sygnał mowy modelowany jest zależnością:

$$s(n) = x(n) \star h(n) \approx \sum_{k=0}^{+\infty} s_p(nT_s - kT_0), \tag{3.2}$$

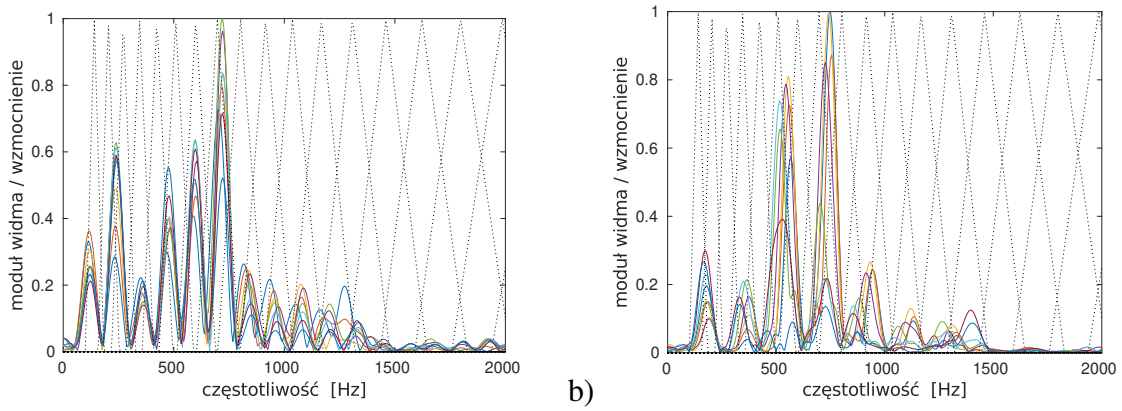
gdzie $g(n)$ to kształt pojedynczego impulsu pobudzającego, $p(n) = \sum_{k=0}^{+\infty} \delta(nT_s - kT_0)$ jest w przybliżeniu funkcją grzebieniową - ciągiem impulsów pobudzających (ang. *pulse train*) o czasie repetycji równym okresowi podstawowemu T_0, $s_p(n)$ jest odpowiedzią pełnego systemu na pojedynczy impuls pobudzenia $\delta(n)$, natomiast T_s jest odstępem próbkowania.

W zastosowaniach praktycznych korzystamy ze skończonej w czasie reprezentacji sygnału mowy $s(n)$, czyli $z(t,n)$, która jest wynikiem ramkowania sygnału według zależności (2.10). W efekcie widmo ramki sygnału przybiera postać:

$$Z(t,\omega) = DTFT\left\{z(t,n)\right\} = DTFT\left\{s(n)\right\} \star DTFT\left\{w(n)\right\} = S(\omega) \star W(\omega), \qquad (3.3)$$

gdzie DTFT (ang. *Discrete Time Fourier Transform*) jest operatorem transformacji Fouriera sygnału dyskretnego. W ogólności pewne aspekty wpływu operacji okienkowania w postaci członu $W(\omega)$ można kompensować poprzez zastosowanie okna czasowego (np. Hamminga, Kaisera) i eliminację efektu obcięcia (ang. *cut-off*), co jest celem tej pracy.

Obserwacja widm poszczególnych dźwięcznych ramek sygnału mowy wskazuje na istotny wpływ częstotliwości F_0 na te widma. Na rys. 3.1 a-b przedstawiono widma amplitudowe kolejnych ramek sygnału mowy zawierających fonem a wybranych z dłuższych nagrań wypowiedzi tego samego mówcy, nagranych w identycznych warunkach, różniących się tylko wartością częstotliwości F_0. Dla rys. 3.1a) $F_0 = 130$ Hz, a dla rys. 3.1b) $F_0 = 195$ Hz. Linią kropkowaną w tle, wykreślono bank filtrów używany podczas parametryzacji HFCC.

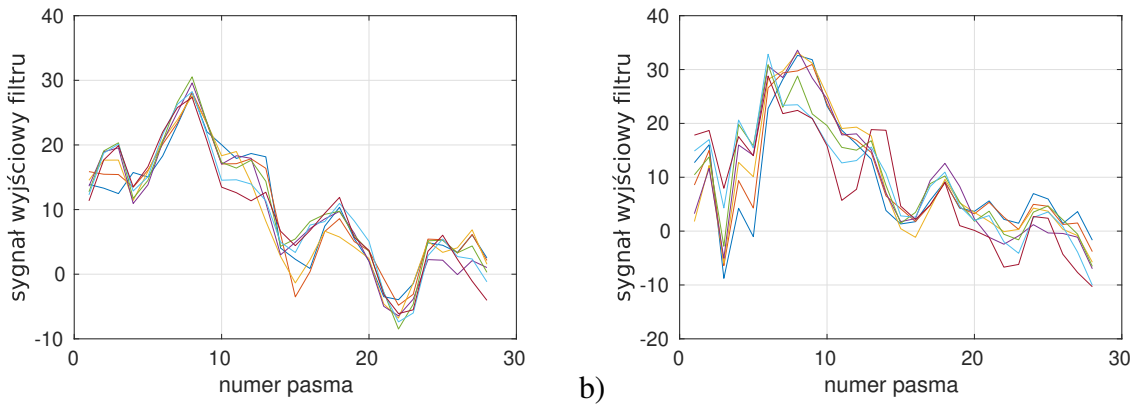


a) b)

Rys. 3.1: Widma amplitudowe ramek sygnału mowy zawierających fonem a z naniesionymi w tle bankami filtrów stosowanymi w parametryzacji HFCC. Częstotliwość tonu krtaniowego wynosi: a) $F_0 = 130$ Hz, b) $F_0 = 195$ Hz.

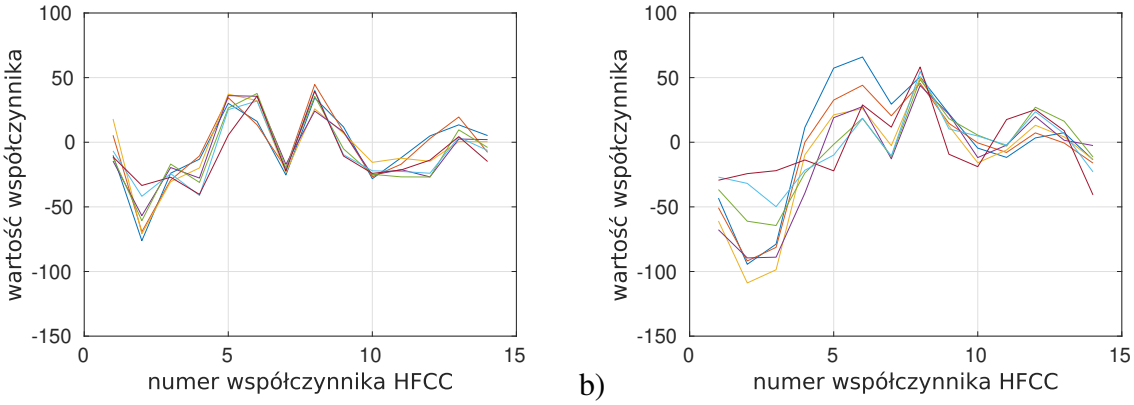
Główne różnice pomiędzy tymi widmami leżą w różnych położeniach lokalnych maksimów, które są wielokrotnościami częstotliwości podstawowej. Zmiana wartości częstotliwości F_0 determinuje odstęp pomiędzy kolejnymi harmonicznymi. Wskutek obecności zafalowań (ripples) w widmach formanty nie są wyraźnie widoczne i trudno je jednoznacznie zlokalizować, ale można oszacować, że częstotliwości środkowe dwóch pierwszych formantów wynoszą ok. 800 Hz i 1.3 kHz. Częstotliwości trzeciego i czwartego formantu wynoszą ok. 2.4 kHz, 4 kHz, jednak ze względu na niskie poziomy mocy w tym zakresie częstotliwości, dla lepszej czytelności rysunków i lepszej ilustracji problemu występowania ripples, zakres częstotliwości ograniczono do 2 kHz.

Konsekwencją odmiennego położenia maksimów lokalnych widma są różne poziomy mocy przypadające na kolejne filtry trójkątne z rys. 3.1. Prowadzi to do innych wartości widm melo-



Rys. 3.2: Widma melowe poszczególnych ramek sygnału mowy zawierających fonem a. Częstotliwość tonu krtaniowego wynosi: a) $F_0 = 130$ Hz, b) $F_0 = 195$ Hz.

wych przy różnych F_0, co można zaobserwować na rys. 3.2. Wykresy te potwierdzają, że różnice te są istotne, mimo że w obu przypadkach wypowiedziany został ten sam fonem przez tego samego mówcę z zachowaniem identycznych warunków rejestracji. Szczególnie duże różnice występują dla pasma nr 4. Wpływ tych różnic na współczynniki cepstralne można zaobserwować na rys. 3.3.



Rys. 3.3: Cepstra poszczególnych ramek sygnału mowy zawierających fonem a. Częstotliwość tonu krtaniowego wynosi: a) $F_0 = 130$ Hz, b) $F_0 = 195$ Hz.

Warto zauważyć duże różnice pomiędzy współczynnikami HFCC dla obu analizowanych przypadków. Co więcej, zmiana F_0 powoduje znaczny rozrzut współczynników, szczególnie tych o mniejszych indeksach, które są bardziej istotne z punktu widzenia rozpoznawania w systemach ARM. Może to prowadzić do znacznych błędów w klasyfikacji poszczególnych ramek sygnału, np. poprzez przypisanie im różnych etykiet, mimo że reprezentują dokładnie ten sam fonem. Z kolei błędy klasyfikacji mogą mieć degradujący wpływ na skuteczność algorytmów w dalszych etapach przetwarzania.

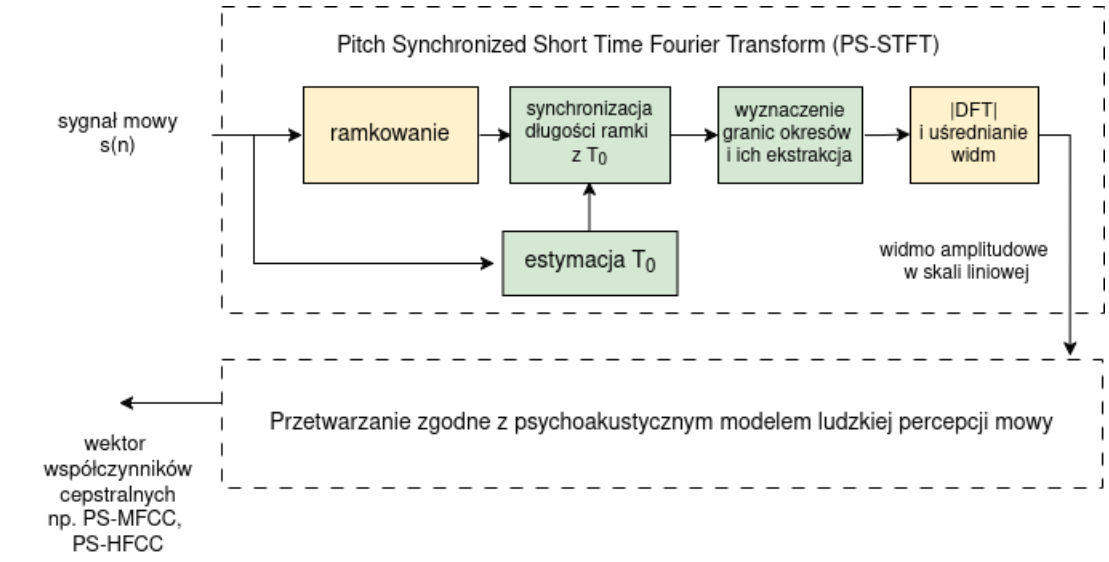
Podsumowując, analiza rysunków 3.1 - 3.3 wskazuje na problem niepożądanej zmienności współczynników cepstralnych, związanej nie z zawartością semantyczną sygnału, lecz ze zmiennością częstotliwości F_0. Klasyczne metody parametryzacji nie są na nią odporne, gdyż nie

kompensują wpływu okresowości pobudzenia. Uzasadniona jest zatem konieczność poszukiwania metod odpornych, tak aby współczynniki obliczane na etapie ekstrakcji cech były mniej zależne od tej częstotliwości. Powyższe rozważania dostarczają intuicyjnego i empirycznego uzasadnienia dla modyfikacji istniejących metod parametryzacji sygnału mowy.

3.2. Własne rozwiązania problemu korekcji widma ramki sygnału

3.2.1. Zastosowanie zmiennej długości ramki sygnału

Pierwsza z zaproponowanych w tej pracy metod nosi nazwę PS-HFCC (ang. *Pitch Synchronized Human Factor Cepstral Coefficient*) i polega na modyfikacji klasycznej parametryzacji HFCC poprzez zastosowanie zmiennej długości ramek, których długość zsynchronizowana jest z okresem podstawowym T_0. Schemat blokowy metody przedstawiono na rys. 3.4.



Rys. 3.4: Schemat obliczania współczynników parametryzacji PS-HFCC

Aby uniknąć efektu obcięcia (ang. *cut-off*), przyjmijmy, że długość ramki analizy spełnia warunek $T_w = N \cdot T_0, \quad N \in \mathcal{Z}$, czyli jest całkowitą wielokrotnością bieżącej wartości okresu tonu podstawowego. W takiej sytuacji, przy założeniu niezmienności T_0 w obszarze ramki, sygnał mowy $z(t,n)$ przyjmuje postać:

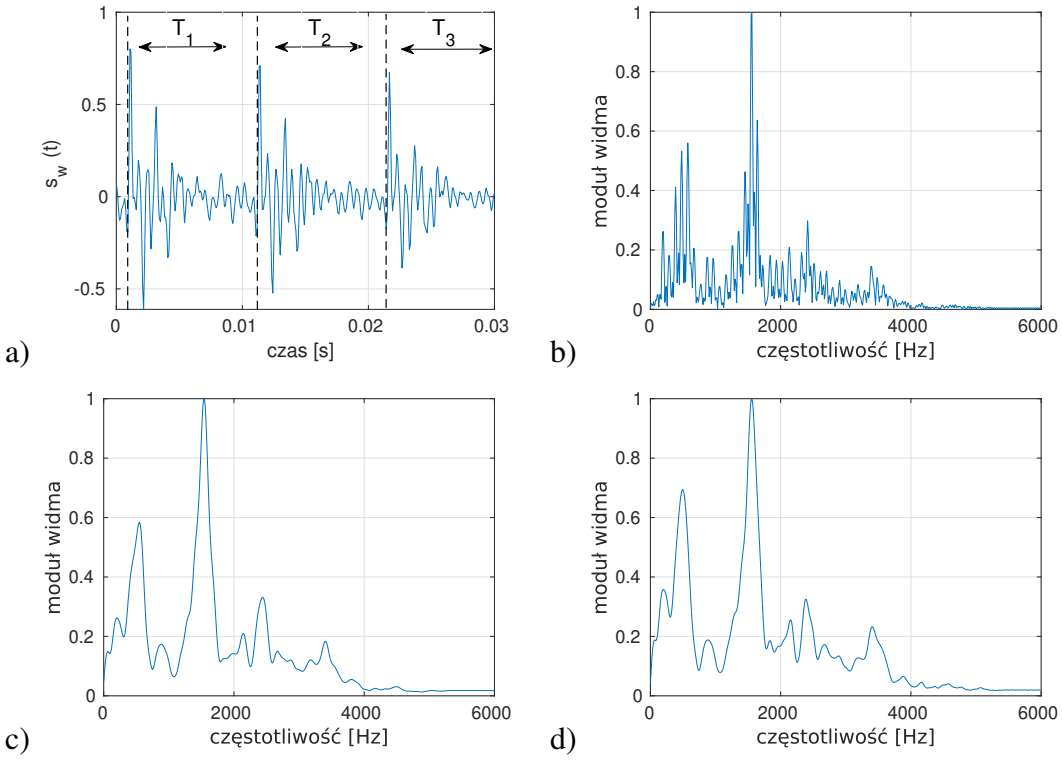
$$z(t,n) = \sum_{k=0}^{N-1} s_p(nT_s - kT_0), \tag{3.4}$$

a jego reprezentacja widmowa:

$$Z(t,\omega) = \left\{ S_p(\omega) \cdot \sum_{k=0}^{N-1} e^{-j\omega k T_0} \right\} \star W(\omega) = \left\{ S_p(\omega) \frac{sin(\omega T_0 \cdot \frac{N}{2})}{sin(\omega T_0 \cdot \frac{1}{2})} \cdot e^{-j\omega \frac{N-1}{2} T_0} \right\} \star W(\omega), \tag{3.5}$$

ilustruje skalę problemu występowania znaczących zafalowań w widmie amplitudowym ramki analizowanego sygnału $z(t,n)$ nawet w tak specyficznym przypadku. Jednak, dla $N=1$ wpływ okna na widmo $Z(t,\omega)$ znika, ale implikuje konieczność estymacji każdej kolejnej wartości T_0.

Biorąc pod uwagę powyższe rozważania, zaproponowany algorytm polega na: (i) ustaleniu liczby N pełnych okresów T_0 mieszczących się w ramce sygnału o długości N_r, czyli $N=\left\lfloor \frac{N_r}{T_0} \right\rfloor$, gdzie $\lfloor\cdot\rfloor$ oznacza część całkowitą, (ii) oznaczeniu początków okresów w ramce poprzez lokalizację maksimów wartości bezwzględnej sygnału i znalezienie punktów przejścia przez zero, (iii) ekstrakcji N pojedynczych okresów sygnału, (iv) uśrednieniu ich widm. W związku z tym przetwarzanie odbywa się w ramkach o zmiennej długości, zależnej od okresu podstawowego T_0. Na rys. 3.5 przedstawiono ilustrację działania tej metody.



Rys. 3.5: Ilustracja wpływu synchronizacji długości ramki z okresem podstawowym na widma sygnałów rzeczywistych: a) przebieg czasowy sygnału w ramce z zaznaczonymi okresami T_1, T_2, T_3, b) widmo ramki sygnału z widocznymi zafalowaniami, c) widmo okresu T_2, d) widmo okresu T_3.

Rysunek 3.5 $a)$ przedstawia przebieg sygnału w ramce o długości 30 ms, wraz z zaznaczonymi kolejnymi okresami T_0 sygnału (oznaczonymi jako T_1, T_2, T_3). Widmo takiego sygnału, o długości $3T_0$, wraz z widocznymi zafalowaniami zaprezentowano na rys. 3.5 $b)$, a na rys. 3.5 $c)-d)$ - przykładowe widma okresów T_2 i T_3. Dla celów porównawczych i lepszej czytelności widma zostały unormowane. Z powyższych wykresów wywnioskować można, że widma pojedynczych okresów sygnału, pozbawione degradującego wpływu okresowości pobudzenia, są bardziej gładkie, a różnice pomiędzy nimi są niewielkie, stąd ich uśrednianie jest zasadne. Dodatkowo przy zmiennej długości okresu T_0 dla zachowania identycznej rozdzielczości widm kolejnych ramek stosowano uzupełnianie zerami. Tak przygotowane widmo ramki sygnału mowy

podlega dalszemu przetwarzaniu, czyli parametryzacji HFCC zgodnie ze schematem na rys. 2.8. W efekcie otrzymuje się wektory współczynników cepstralnych PS-HFCC, na podstawie których opracowuje się model statystyczny GMM, pozwalający na przeprowadzenie klasyfikacji. Podejście to zostało także opisane przez autora w pracach [28] [27]. Zaproponowane rozwiązanie może stanowić uogólnienie także innych algorytmów parametryzacji cepstralnych, np. MFCC lub BFCC.

3.2.2. Metoda wykorzystująca filtrację odwrotną

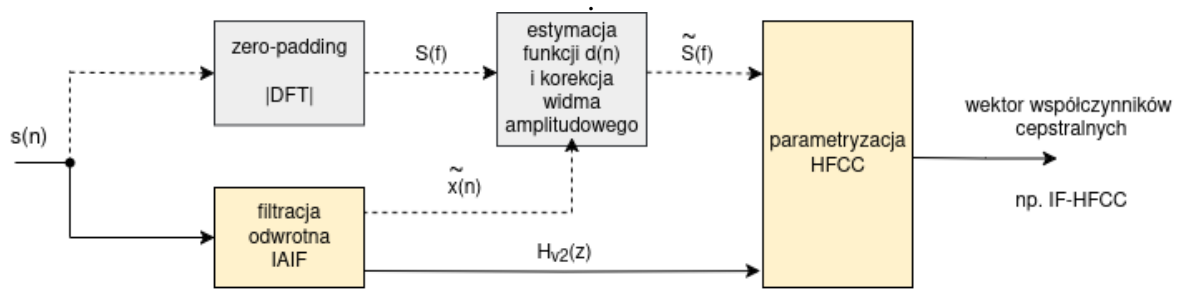
W drugiej metodzie zaproponowano ograniczenie wpływu pobudzenia tonem krtaniowym poprzez operację jego odfiltrowania. Wyznaczane w ten sposób współczynniki parametryzacji nazwano IF-HFCC (ang. *Inverse Filtering Human Factor Cepstral Coefficient*). Impulsy pobudzenia można zarejestrować wykonując specjalistyczne badania przy użyciu elektroglotografu [36], jednak w literaturze zaproponowano wiele algorytmów ich estymacji bezpośrednio z sygnału mowy [19]. Jest to jedno z ważniejszych zagadnień w przetwarzaniu mowy, w praktycznych zastosowaniach wykorzystywane m.in. do rozpoznawania mówcy [80], analizy jego stanu emocjonalnego [101], czy też do syntezy mowy [86]. W literaturze najczęściej stosowane są algorytmy filtracji odwrotnej polegające na odfiltrowaniu z sygnału mowy wpływu komponentów $v(n)$ i $r(n)$ w oparciu o ich modele parametryczne wyznaczone przy pomocy analizy LPC. W podejściu tym ważne jest wyznaczenie wiarygodnego modelu kanału głosowego, co jest możliwe na kilka sposobów [102]. Wśród nich warto wyróżnić: (i) algorytm CPIF (ang. *Closed Phase Inverse Filtering*) [104], w którym analizowana jest tylko faza zamknięcia cyklu drgań strun głosowych oraz (ii) algorytmy, w których wykorzystywane jest podejście iteracyjne i mechanizmy synchronizacji np. IAIF (ang. *Iterative Adaptive Inverse Filtering*) [86] i PS-IAIF (ang. *Pitch Synchronized Iterative Adaptive Inverse Filtering*) [2].

Oprócz filtracji odwrotnej istnieją także algorytmy bazujące na mieszano-fazowym modelu sygnału mowy. Zakłada się w nich, że odpowiedź impulsowa kanału głosowego oraz część pobudzenia odpowiedzialna za fazę powrotu są traktowane jako komponenty przyczynowe, a część pobudzenia stanowiąca fazę otwarcia w cyklu drgań strun głosowych jako komponent nieprzyczynowy. Separacji tych komponentów można dokonać za pomocą algorytmu ZZT (ang. *Zeros of the Z-Transform*) [8] lub CCD (ang. *Complex Cepstrum Decomposition*) [18].

Schemat blokowy wyznaczania współczynników IF-HFCC zaprezentowano na rys. 3.6.

Punktem wyjścia w tej metodzie jest algorytm IAIF, za pomocą którego można dokonać automatycznej dekompozycji sygnału mowy na składowe pobudzenia oraz odpowiedzi impulsowej traktu głosowego. Teoretycznie, sygnał pobudzenia, dla każdej dźwięcznej ramki, można wyznaczyć z zależności (3.6) [86], tj.

$$x(n) = s(n) \star \left(v(n) \star r(n)\right)^{-1}, \tag{3.6}$$



Rys. 3.6: Schemat wyznaczania współczynników parametryzacji IF-HFCC.

gdzie $(.)^{-1}$ oznacza odwrotność w sensie splotowym. Wprowadzając $d(n) = x(n) \star r(n)$, tj. jako splot sygnału pobudzenia i funkcji opisującej emisyjność ust, wielkość $d(n)$ można wyznaczyć z zależności

$$\tilde{d}(n) = s(n) \star \tilde{v}(n)^{-1}, \tag{3.7}$$

Zależność (3.7) opisuje przypadek ślepego rozplatania. Wymaga on estymacji wielkości $v(n)$, a następnie wyznaczenia odwrotności w sensie splotowym tej wielkości. W takiej sytuacji pojawia się problem stabilności. Stabilność ta jest zagwarantowana, jeśli wielkość $v(n)$ jest minimalnofazowa lub stosowany jest algorytm, który minimalnofazowość wymusza. Takim algorytmem jest filtracja średniokwadratowa [82] i jest ona wykorzystywana w analizowanej metodzie filtracji odwrotnej. Szczegółowy opis wraz z schematem blokowym algorytmu filtracji odwrotnej przedstawiono w Dodatku B.

Współczynniki parametryzacji IF-HFCC można obliczyć na dwa sposoby: i) wykorzystując estymator transmitancji kanału głosowego $H_{v2}(z)$ i poddając go parametryzacji HFCC, ii) wykorzystując uzyskany w procesie filtracji odwrotnej estymator sygnału pobudzenia $\tilde{x}(n)$, który może być dalej wykorzystany do estymacji funkcji $d(n)$. Ponieważ współczynniki HFCC liczone są w oparciu o moduł widma sygnału i ponieważ $|S(f)| = |H(f)||D(f)|$, to znając estymator $\tilde{d}(n)$ możemy dokonać korekcji widma amplitudowego, zgodnie z zależnością:

$$|\tilde{S}(f)| = \frac{|S(f)|}{|\tilde{D}(f)|} \tag{3.8}$$

Należy się spodziewać, że wyniki klasyfikacji uzyskane w oparciu o współczynniki parametryzacji wyznaczone z modułu transmitancji kanału głosowego będą bardzo podobne do tych obliczonych na podstawie skorygowanego widma sygnału. Potwierdzają to eksperymenty, dlatego w niniejszej pracy zdecydowano się na podejście pierwsze.

Warto pamiętać o dwóch faktach: (i) wyniki uzyskiwane są przy założeniu minimalnofazowości wszystkich elementów zależności (2.5) i (ii) ponieważ w parametryzacji HFCC faza sygnału nie jest uwzględniana, można mieć nadzieję, że modelowanie elementów zależności (2.5) z wykorzystaniem modelu LPC, przyniesie oczekiwane efekty. W pracach [30] [29] autor opisał i szczegółowo przeanalizował wyniki otrzymane przy użyciu metody IF-HFCC, a także jej połączenie z metodą zmiennej długości ramki [31].

3.2.3. Podsumowanie

Podsumowując: (i) analiza widm z efektem obcięcia wykazuje zasadność kompensacji i poszukiwania metod odpornej parametryzacji sygnału, (ii) zaproponowano dwie własne metody parametryzacji, mające na celu minimalizację wpływu zmienności częstotliwości F_0, (iii) w następnym rozdziale zaproponowane metody poddano weryfikacji przy użyciu sygnałów modelowych i rzeczywistych.

Rozdział 4

Badania skuteczności algorytmów i jakości rozpoznawania

Proces weryfikacji skuteczności zaproponowanych w poprzednim rozdziale metod odpornej parametryzacji stanowi kluczową część oceny przydatności tych rozwiązań w praktycznych zastosowaniach. W pierwszej części rozdziału przedstawiono charakterystykę danych - zarówno modelowych fragmentów mowy, jak i rzeczywistych, pozyskanych od wielu mówców. Opisano także sposób wyznaczania statystycznego modelu GMM fonemów, za pomocą którego możliwa jest ilościowa ocena skuteczności rozpoznawania oraz zdefiniowano odpowiednie miary skuteczności do obiektywnej oceny działania rozważanych metod. W kolejnym kroku przeprowadzono badania na sygnałach modelowych i rzeczywistych oraz dokonano globalnej analizy błędów.

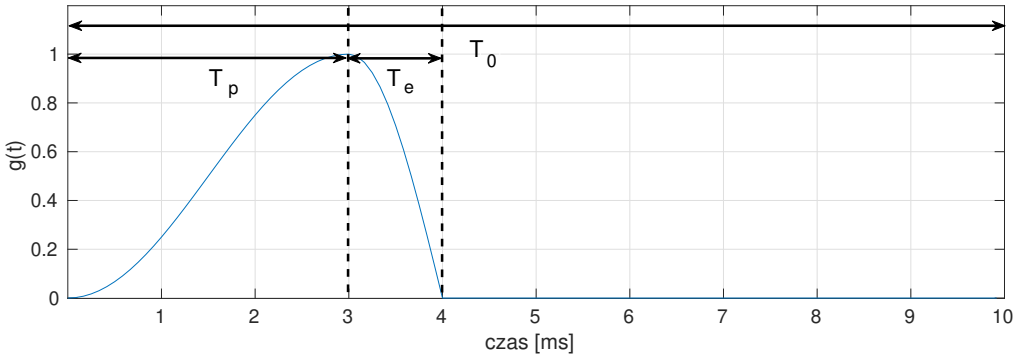
4.1. Metodologia badań

4.1.1. Sygnały modelowe i ich charakterystyka

Pojedynczy impuls tonu krtaniowego $g(t)$ można zamodelować na wiele różnych sposobów. W modelu generowania mowy LP ma on postać delty Kroneckera, a cały ciąg pobudzenia $p(n) = \sum_{k=0}^{+\infty} \delta(nT_s - kT_0)$ jest funkcją grzebieniową z czasem repetycji równym okresowi podstawowemu T_0. W rzeczywistości jednak nieskończenie mała szerokość impulsu jest zbyt daleko idącym uproszczeniem, dlatego w przyjętym w niniejszej pracy modelu Fanta (zob. rozdz. 2.2) zakłada się określony kształt i szerokość impulsu $g(n)$ pobudzającego kanał głosowy, co ma swoje odzwierciedlenie w działaniu aparatu mowy i jest związane z rozwieraniem i zamykaniem się strun głosowych. Spośród wielu modeli opisanych w literaturze, jednym z najbardziej popularnych jest model Rosenberga dany zależnością:

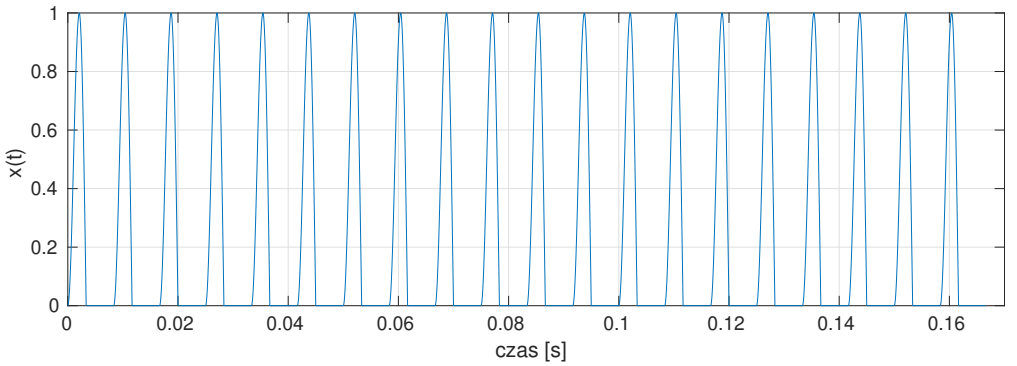
$$g(t) = \begin{cases} 0,5A[1-\cos(\frac{\pi t}{T_p})]; & 0 \leq t \leq T_p \\ A\cos(\frac{\pi(t-T_p)}{2T_c}); & T_p \leq t \leq T_e \\ 0; & T_e \leq t \leq T_0, \end{cases} \tag{4.1}$$

gdzie $T_p = \alpha_1 T_0$ jest czasem otwierania strun głosowych, aż do ich maksymalnego rozwarcia, $T_e = \alpha_2 T_0$ - czasem zwierania strun głosowych, T_0 - okresem podstawowym, a parametry α_1 oraz α_2 są stałymi wyrażającymi stosunek trwania faz otwarcia i zamknięcia względem okresu T_0. Przykładowy impuls Rosenberga dla parametrów $\alpha_1 = 0.3$, $\alpha_2 = 0.1$ wraz z zaznaczonymi czasami trwania poszczególnych jego faz przedstawiono na rys. 4.1.



Rys. 4.1: Impuls pobudzenia według modelu Rosenberga.

Do wygenerowania modelowego sygnału, zgodnie z zależnością (3.1), wykorzystano ciąg takich impulsów. Przykładowy przebieg czasowy sygnału pobudzenia $x(n)$ o częstotliwości $F_0 = 120$ Hz przedstawiono na rys. 4.2.



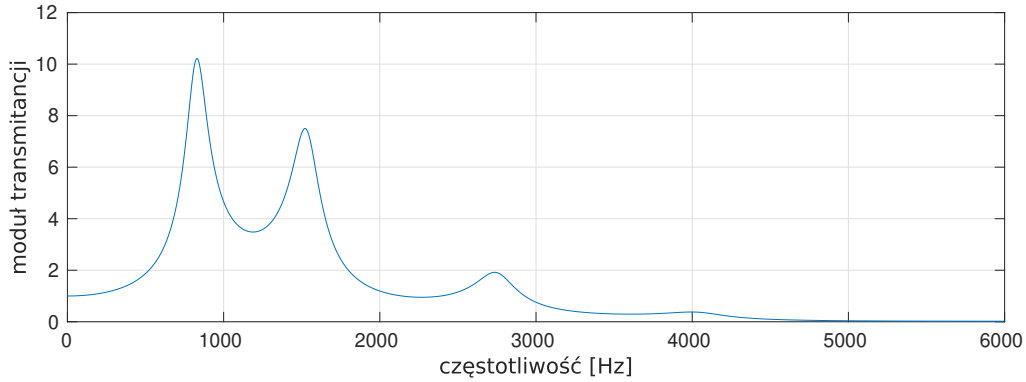
Rys. 4.2: Przykładowy przebieg czasowy modelowego sygnału pobudzenia $x(n)$ o częstotliwości $F_0 = 120$ Hz.

Tabela 4.1 zawiera zaczerpnięte z literatury zakresy średnich wartości częstotliwości formantowych F_1, F_2, F_3, F_4 dla samogłosek mowy polskiej [42] [5] i stanowi podstawę do wygenerowania modelu kanału głosowego.

Tab. 4.1: Zakresy średnich wartości częstotliwości formantowych dla samogłosek mowy polskiej

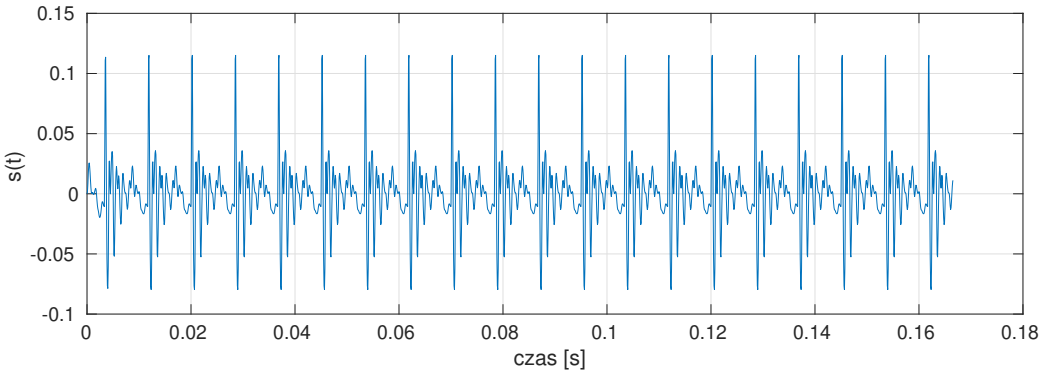
samogłoska	$\mathbf{F_1}$ [Hz]	$\mathbf{F_2}$ [Hz]	$\mathbf{F_3}$ [Hz]	$\mathbf{F_4}$ [Hz]
'i'	188-275	2078-2836	2670-3432	3316-4144
'y'	262-391	1689-2362	2424-3146	3124-4226
'e'	524-630	1580-2228	2468-3146	3064-4034
'a'	683-1021	1132-1566	2328-2860	3098-4088
'o'	493-679	788-1100	2410-3026	3194-3954
'u'	242-338	558-789	2266-3188	2942-4058

Na rys. 4.3 przedstawiono przykładowy moduł funkcji transmitancji modelowego kanału głosowego dla samogłoski a z wartościami częstotliwości formantowych: $F_1 = 825$ Hz, $F_2 = 1530$ Hz, $F_3 = 2750$ Hz, $F_4 = 4048$ Hz,



Rys. 4.3: Przykładowy moduł funkcji transmitancji modelowego kanału głosowego dla samogłoski a.

Przykładowy przebieg czasowy 20 okresów modelowej samogłoski a przedstawiono na rys. 4.4. Wykorzystując technikę opisaną w tym rozdziale, wygenerowano sygnały syntetyczne dla każdej z samogłosek. W rezultacie otrzymano sygnały pozbawione cech mowy naturalnej i zakłóceń.



Rys. 4.4: Przykładowy przebieg czasowy sygnału dla samogłoski a, wygenerowanego na podstawie modelu źródło-filtr Fanta.

4.1.2. Baza sygnałów rzeczywistych

Kluczowa weryfikacja skuteczności proponowanych metod bazuje na nagraniach mowy. Istotnym wyzwaniem w kontekście przetwarzania i klasyfikacji sygnału mowy jest uwzględnienie złożoności fonetycznej, artykulacyjnej i kontekstowej. W związku z tym nagrania zostały zrealizowane przez zwykłych ludzi, niebędących profesjonalnymi lektorami, aby uniknąć nienaturalnych wzorców artykulacyjnych w przetwarzanych danych. Zbiór nagrań będący bazą eksperymentów stanowią nagrania 42 dorosłych głosów męskich zarejestrowane w różnych regionach Polski. Dla każdego z mówców nagrano 150 słów języka polskiego. Do rejestracji użyto magnetofonu reporterskiego TASCAM z miniaturowym mikrofonem. Częstotliwość próbkowania wyniosła 48 kHz. W eksperymentach sygnały zostały poddane decymacji rzędu $dr = 4$, czyli wynikowa częstotliwość próbkowania wyniosła 12 kHz. Rozdzielczość przetwornika A/C wyniosła 16 bitów, a stosunek sygnał/szum był na poziomie SNR=35 dB. Niski poziom SNR wynika z realizacji nagrań w samochodzie na postoju w miastach.

Baza została skonstruowana w taki sposób, aby uwzględnić różnorodność regionalną mówców oraz zapewnić różnorodność fonetyczną naturalnej mowy. Słowa zostały dobrane tak, aby odzwierciedlić szeroką gamę kontekstów fonetycznych. Pozwala to na przeprowadzenie analizy i badań zmienności sygnału w zależności od sąsiedztwa poszczególnych fonemów (sąsiedztwo fonemów z lewej i prawej strony, jak i bardziej złożone konteksty obustronne). Słowa nagrane w bazie obejmują zarówno krótkie, jednosylabowe formy takie jak np. *„dwa”, „tak”*, jak i bardziej złożone struktury fonetyczne, np. *„czepek”, „poprzedni”, „zapamiętaj”*. Dzięki tej różnorodności możliwe jest uchwycenie zmian wektorów współczynników cepstralnych, które są determinowane zarówno długością wyrazu, jak i jego strukturą fonetyczną.

Sygnały zostały poddane transkrypcji fonetycznej zgodnej z IPA (ang. *International Phonetic Alphabet*) i listą fonemów mowy polskiej wg. [42]. Przy przygotowywaniu bazy uwzględnione zostały statystyczne właściwości języka polskiego, czyli częstość występowania poszczególnych jego fonemów. Informacje te przedstawiono w Tab. 4.2. Częstość występowania samogłosek wynosi ok. 40%, a spółgłosek ok. 55%. Pozostałe 5% stanowią przerwy i stany przejściowe.

Tab. 4.2: Częstość występowania fonemów mowy polskiej wyrażona w [%] wraz zapisem transkrypcji fonemów wg. standardu IPA.

fonem	symbol IPA	przykład	częstość wyst. [%]	fonem	symbol IPA	przykład	częstość wyst. [%]
i	/i/	k<u>i</u>no	3,4	ż	/ʒ/	<u>ż</u>aba	1,3
y	/ɨ/	r<u>y</u>ba	3,8	si	/ɕ/	<u>ś</u>roda	0,1
e	/e/	rz<u>e</u>ka	10,6	zi	/ʑ/	<u>z</u>ima	0,2
a	/a/	k<u>a</u>wa	9,7	ch	/x/	<u>ch</u>udy	1,0
o	/o/	b<u>o</u>k	8,0	c	/ts/	<u>c</u>ena	1,2
u	/u/	b<u>u</u>k	2,8	dz	/dz/	<u>dz</u>ban	0,2
j	/j/	<u>j</u>eden	4,4	cz	/tʃ/	<u>cz</u>apka	1,2
ł	/w/	<u>ł</u>ysy	1,8	dż	/dʒ/	<u>dż</u>em	0,1
r	/r/	<u>r</u>ak	3,2	ci	/tɕ/	<u>ci</u>sza	1,2
l	/l/	<u>l</u>ewy	1,9	dzi	/dʑ/	<u>dzi</u>wak	0,7
m	/m/	<u>m</u>ak	3,2	p	/p/	<u>p</u>iłka	3,0
n	/n/	<u>n</u>os	4,0	b	/b/	<u>b</u>eczka	1,5
n’	/ɲ/	<u>ni</u>ski, ko<u>ń</u>	2,4	t	/t/	<u>t</u>on	4,8
n,	/ŋ/	pi<u>ę</u>ć	1,6	d	/d/	<u>d</u>usza	2,1
f	/f/	<u>f</u>ajka	1,3	ki	/c/	<u>ki</u>no	0,7
w	/v/	<u>w</u>arta	2,9	gi	/ɟ/	<u>gi</u>l	0,1
s	/s/	<u>s</u>ok	2,8	k	/k/	<u>k</u>ot	2,5
z	/z/	<u>z</u>ero	1,5	g	/g/	<u>g</u>óra	1,3
sz	/ʃ/	<u>sz</u>afa	1,9	#	stany przejściowe		4,7

Przedmiotem badań niniejszej pracy były dźwięczne fonemy mowy polskiej, a to ograniczenie wynika z przyjętych metod kompensacji. Wśród wyselekcjonowanych fonemów znalazły się samogłoski (i, y, e, a, o, u), spółgłoski półsamogłoskowe (j, ł), a także spółgłoski nosowe (m, n, n', $n_{\prime}$), boczna (l) oraz drżąca (r). Nie zostały poddane analizie spółgłoski szczelinowe (np. z lub ż) i zwarto-szczelinowe (np. $dż$) ze względu ich mieszany model pobudzenia, tzn. oprócz składowej dźwięcznej istotną rolę odgrywa także komponent szumowy. Z analizy wyłączono także fonemy plozyjne (np. b lub g), charakteryzujące się bardzo krótkim czasem trwania, nieprzekraczającym 20 ms. Ze względu na powyższe uwarunkowania, biorąc pod uwagę konieczność estymacji bieżącej częstotliwości tonu krtaniowego F_0, stanowiącej podstawę metody PS-HFCC, analiza fonemów bezdźwięcznych oraz tych, w których występuje szumowy komponent pobudzenia, pozostaje bezprzedmiotowa w kontekście problemu badawczego niniejszej pracy.

Wysoka jakość opisu nagrań została zagwarantowana dzięki ręcznej segmentacji i etykietyzacji. Za jednostkę fonetyczną przyjęto fonem, nazywany również w dalszej części pracy stanem. Pierwotna długość ramki wyniosła $N_r = 30$ ms, a skok ramkowania $N_s = 10$ ms. Ze względu na przyjętą w jednej z metod odpornej parametryzacji zmienną długości ramki, podczas obliczania widma sygnału stosuje się klasyczną technikę uzupełniania zerami (ang. *zero-padding*) do długości $N_f = 1024$ próbki. Liczba współczynników cepstralnych wyniosła $P = 14$ (z pominięciem wskaźnika głośności).

4.1.3. Statystyczny model fonemów

Rozpoznawanie mowy przez człowieka jest bardzo skomplikowanym i wieloetapowym procesem neurobiologicznym. Klasyfikacja mowy i dźwięków docierających do nas z otoczenia nie odbywa się w sposób deterministyczny i jednoznaczny, ale wskazuje na rozwiązanie o najwyższym prawdopodobieństwie. Dzięki temu możliwe jest statystyczne odwzorowanie dużej zmienności fonemów, mogących mieć różne realizacje akustyczne w zależności od mówcy, kontekstu, akcentu, stylu i tempa wypowiedzi.

Powszechnie stosowanym modelem fonemów w systemach ARM jest model GMM, będący ważoną sumą K wielowymiarowych rozkładów normalnych. Każda z K składowych reprezentuje określoną klasę akustyczną [89]. Zwiększanie liczby klas powoduje lepsze dopasowanie modelu i polepsza zdolności klasyfikacyjne, jednak wymaga to dysponowania wystarczająco obszerną bazą sygnałów w celu zapewnienia reprezentatywności dla każdej z klas. W niniejszej pracy przyjęto liczbę $K = 7$.

Tak więc, prawdopodobieństwo, że wektor obserwacji $\mathbf{o}$ odpowiada fonemowi f można wyznaczyć z zależności:

$$p_f(\mathbf{o}) = \sum_{k=1}^{K} w_{fk} \mathcal{N}(\mathbf{o}|\mathbf{m}_{f,k}, \mathbf{\Sigma}_{f,k}), \tag{4.2}$$

przy czym w_{fk} oznacza współczynnik wagowy k-tej składowej, f indeksem stanu (fonemu), zaś $\mathcal{N}(\mathbf{o}|\mathbf{m}_{f,k}, \mathbf{\Sigma}_{f,k})$ jest gęstością prawdopodobieństwa rozkładu normalnego:

$$\mathcal{N}(\mathbf{o}|\mathbf{m}_{f,k}, \mathbf{\Sigma}_{f,k}) = \frac{1}{\sqrt{(2\pi)^P |\mathbf{\Sigma}_{f,k}|}} e^{-\frac{1}{2}(\mathbf{o}-\mathbf{m}_{f,k})^T \mathbf{\Sigma}_{f,k}^{-1} (\mathbf{o}-\mathbf{m}_{f,k})} \tag{4.3}$$

o wartości średniej $\mathbf{m}_{f,k}$ i macierzy kowariancji $\mathbf{\Sigma}_{f,k}$. W modelu GMM używa się pełnych bądź diagonalnych macierzy kowariancji. Stosowanie macierzy pełnych wymaga więcej danych i stwarza problemy obliczeniowe. Uważa się, że stosowanie macierzy diagonalnych nie wpływa znacząco na dokładność modelu [89] [57]. Dlatego w pracy stosowane są macierze diagonalne postaci:

$$\mathbf{\Sigma}_{f,k} = \begin{bmatrix} \sigma^2_{f,k,1} & 0 & \cdots & 0 \\ 0 & \sigma^2_{f,k,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2_{f,k,P} \end{bmatrix} \tag{4.4}$$

W związku z powyższym, zależność (4.3) można zapisać jako:

$$\mathcal{N}(\mathbf{o}|\mathbf{m}_{f,k}, \mathbf{\Sigma}_{f,k}) = \prod_{p=1}^{P} \frac{1}{\sqrt{(2\pi)}\sigma_{f,k,p}} e^{-\frac{1}{2\sigma^2_{f,k,p}}[o(p)-m_{f,k}(p)]^2} \tag{4.5}$$

Z kolei współczynniki wagowe można interpretować jako prawdopodobieństwo przynależności wektora obserwacji $\mathbf{o}$ do konkretnej klasy akustycznej. Wartości tych współczynników

spełniają zależność:

$$\sum_{k=1}^{K} w_{f,k} = 1 \tag{4.6}$$

Model statystyczny Λ_f danej jednostki fonetycznej f jest zatem opisany za pomocą zbioru trzech parametrów: wektorów wartości średnich, diagonalnych macierzy kowariancji i współczynników wagowych, czyli:

$$\Lambda_f = \{\mathbf{m}_{f,k}, \mathbf{\Sigma}_{f,k}, w_{f,k}\}, \qquad k \in \{1, ..., K\} \tag{4.7}$$

Parametry modelu Λ_f estymuje się metodą największej wiarygodności ML (ang. *Maximum-Likelihood*) z zastosowaniem algorytmu EM (ang. *Expectation-Maximization*) [14] [89]. Jest to procedura iteracyjna z kryterium stopu oraz losową inicjalizacją parametrów dla każdej składowej mieszaniny. Algorytm składa się z dwóch naprzemiennie występujących kroków: (i) krok E (ang. *Expectation*), w którym wyznacza się prawdopodobieństwa warunkowe przynależności wektorów obserwacji danego fonemu do wszystkich składowych mieszaniny GMM, (ii) krok M (ang. *Maximization*), w którym wykonywane są operacje uśredniania wektorów danych i wyznaczenia ich macierzy autokowariancji dla pełnego modelu GMM. W każdej iteracji algorytmu spełniona jest zależność:

$$P(\mathbf{O}_f|\Lambda_{f,k}^{(j)}) \geq P(\mathbf{O}_f|\Lambda_{f,k}^{(j-1)}), \tag{4.8}$$

gdzie $\Lambda_{f,k}^{(j)}$ jest nowym zestawem parametrów wyznaczonych w j-tej iteracji algorytmu, gwarantującym maksymalizację funkcji podobieństwa. Algorytm kontynuowany jest do momentu spełnienia warunku stopu, którym może być: (i) zadana liczba iteracji, (ii) kryterium zbieżności funkcji kosztów - algorytm kończy swoje działanie, gdy względna zmiana logarytmu funkcji wiarygodności spadnie poniżej założonego progu tolerancji. Otrzymany wówczas zestaw parametrów uznaje się za optymalny.

Należy mieć na uwadze, że wynik estymacji zależy od wyboru punktu startowego, co nie gwarantuje w pełni jednoznacznej reprezentacji O_f, czyli zbioru obserwacji stanu f. Inicjalizacji parametrów algorytmu można dokonać na wiele sposobów, spośród których wyróżnić można: (i) losowe inicjalizacje, w których początkowe wartości średnie przyjmują postać losowo wybranych wektorów obserwacji ze zbioru O_f, ich wariancje tworzą diagonalną macierz kowariancji, a współczynniki wagowe przyjmują jednakowe wartości, (ii) inicjalizacje wykorzystujące algorytm k-średnich (ang. *k-means*), w których po wstępnym podziale danych na klastry, średnie, macierze kowariancji i współczynniki wagowe wyznacza się na podstawie statystyk poszczególnych klastrów, (iii) inicjalizacje z wykorzystaniem algorytmu k-means++, w której pierwszy centroid wybierany jest losowo, a kolejne dobierane są z prawdopodobieństwem proporcjonalnym do kwadratu odległości od już istniejących centroidów. Dzięki temu wartości początkowe są lepiej odseparowane w przestrzeni cech, co prowadzi do szybszej zbieżności algorytmu i zmniejsza ryzyko utknięcia w lokalnym minimum. Taki sposób inicjalizacji wykorzystano w niniejszej pracy.

4.2. Miary skuteczności rozpoznawania

Do oceny skuteczności opracowanych metod kompensacji stosowano 3 miary: (i) odchylenia standardowe współrzędnych wektora obserwacji, (ii) odległość Kullbacka-Leiblera pomiędzy rozkładami - miara KLD [53], (iii) błąd rozpoznawania pojedynczych ramek - miara FER [57]. Choć na te miary wpływ ma przede wszystkim natura mowy, należy się spodziewać, że efekt obcięcia przy liczeniu widma ramki pogarsza dodatkowo każdą z tych miar.

4.2.1. Odchylenia standardowe współrzędnych wektora obserwacji

Pierwszą z miar skuteczności wybraną do analizy porównawczej proponowanych metod odpornej parametryzacji z klasycznym podejściem HFCC są odchylenia standardowe obliczone oddzielnie dla wszystkich P współrzędnych wektorów obserwacji dla każdego z F fonemów. Niech:

$$\mu_{f,p} = \frac{1}{N_f} \sum_{t=0}^{N_f-1} o_{f,t}(p) \tag{4.9}$$

oznacza wartość średnią p-tej współrzędnej wektora obserwacji fonemu f, N_f oznacza liczebność zbioru $\mathbf{O}_f$. Wówczas

$$\sigma_{f,p} = \frac{1}{N_f - 1} \sqrt{\sum_{t=0}^{N_f-1} (o_{f,t}(p) - \mu_{f,p})^2} \tag{4.10}$$

oznacza odchylenie standardowe p-tej współrzędnej wektorów obserwacji fonemu f. Oczekiwanym i pożądanym rezultatem w tym przypadku jest zmniejszenie wartości $\sigma_{f,p}$ po zastosowaniu metod odpornej parametryzacji, co oznacza redukcję rozrzutu danych w przestrzeni cech.

4.2.2. Odległości pomiędzy rozkładami prawdopodobieństwa GMM

W ogólności naturalną miara do określenia odległości pomiędzy rozkładami gęstości prawdopodobieństwa $p_h(\mathbf{o})$ i $p_g(\mathbf{o})$ dla N-wymiarowego wektora zmiennych losowych $\mathbf{o}$ jest dywergencja Kullbacka-Leiblera zdefiniowana jako [53]:

$$KL(p_h \parallel p_q) = \int_{\mathcal{O}} p_h(\mathbf{o}) \log \left(\frac{p_h(\mathbf{o})}{p_g(\mathbf{o})} \right) d\mathbf{o}. \tag{4.11}$$

Niestety, dla przypadku rozkładów reprezentowanych mieszaniną rozkładów gaussowskich GMM postaci:

$$\begin{aligned} p_h(\mathbf{o}) &= \sum_{i=1}^{K} w_{h,i} \mathcal{N}(\mathbf{o}, \mathbf{m}_{h,i}, \mathbf{\Sigma}_{h,i}) = \sum_{i=1}^{K} w_{h,i} p_{h,i}(\mathbf{o}); \\ p_g(\mathbf{o}) &= \sum_{i=1}^{K} w_{g,i} \mathcal{N}(\mathbf{o}, \mathbf{m}_{g,i}, \mathbf{\Sigma}_{g,i}) = \sum_{i=1}^{K} w_{g,i} p_{g,i}(\mathbf{o}), \end{aligned} \tag{4.12}$$

nie istnieje analityczna formuła wyznaczania takich wartości. W konsekwencji możemy w tym przypadku zastosować metody aproksymacji o naturze stochastycznej, czyli metody symulacyjne Monte-Carlo, wykorzystujące formułę:

$$KL(p_h \parallel p_q) = \frac{1}{D}\sum_{i=1}^{D}\log\frac{p_h(\mathbf{o}_i)}{p_g(\mathbf{o}_i)}, \tag{4.13}$$

które wymagają na ogół ogromnej liczby D danych w postaci obserwacji lub też ich generacji w oparciu o znaną postać rozkładu GMM. Jednakże przy braku wystarczającej liczby danych możemy zastosować aproksymację deterministyczną wyrażenia (4.11) przy wykorzystaniu transformacji UT (ang. *Unscented Transform*) [43].

Przy założeniu, że rozkłady $p_h(\mathbf{o})$ i $p_g(\mathbf{o})$ są postaci GMM (4.12) o K składowych z diagonalnymi macierzami kowariancji, czyli $\mathbf{\Sigma}_{h,i} = diag\{\sigma^2_{h,i,k}\}$ i $\mathbf{\Sigma}_{g,i} = diag\{\sigma^2_{g,i,k}\}$ dla $k = 1, 2, \ldots N$, możemy zapisać, że:

$$\begin{aligned} KL(p_h \parallel p_q) &= \int_{\mathbf{O}} p_h(\mathbf{o})\log\left(\frac{p_h(\mathbf{o})}{p_g(\mathbf{o})}\right)d\mathbf{o} = \underset{p_h}{E}[\log p_h(\mathbf{o})] - \underset{p_h}{E}[\log p_g(\mathbf{o})] = \\ &= \sum_{i=1}^{K} w_{h,i}\underset{p_{h,i}}{E}[\log p_h(\mathbf{o})] - \sum_{i=1}^{K} w_{h,i}\underset{p_{h,i}}{E}[\log p_g(\mathbf{o})]. \end{aligned} \tag{4.14}$$

Dla uproszczenia opisu procedury aproksymacji wyrażenia (4.14) analizujemy jedynie drugi ze składników powyższej sumy, ponieważ pierwszy z nich, przyjmując, że $p_g(\mathbf{o}) = p_h(\mathbf{o})$, jest jego przypadkiem szczególnym. Zgodnie z metodą UT, dla każdej z K składowych rozkładów GMM czyli $p_{h,i}(\mathbf{o}) = \mathcal{N}(\mathbf{o}, \mathbf{m}_{h,i}, \mathbf{\Sigma}_{h,i})$ z diagonalną macierzą kowariancji, czyli:

$$p_{h,i}(\mathbf{o}) = \prod_{k=1}^{N}\frac{1}{\sqrt{2\pi}\sigma_{h,i,k}}e^{-\frac{1}{2\sigma^2_{h,i,k}}[o_k - m_{h,i,k}]^2}, \tag{4.15}$$

generujemy zbiór $2N$ tzw. sigma punktów postaci:

$$\begin{aligned} \mathbf{o}_{i,k} &= \mathbf{m}_{h,i} + \sqrt{N\sigma^2_{h,i,k}}\mathbf{e}_k; \\ \mathbf{o}_{i,k+N} &= \mathbf{m}_{h,i} + \sqrt{N\sigma^2_{h,i,k}}\mathbf{e}_k, \end{aligned} \tag{4.16}$$

gdzie $\mathbf{e}_k$ dla $k = 1, 2, \ldots N$ są wektorami bazowymi w N wymiarowym kartezjańskim układzie współrzędnych, a następnie wyznaczamy aproksymację całki $\underset{p_{h,i}}{E}[\log p_g(\mathbf{o})]$ w oparciu o formułę [32]:

$$\underset{p_{h,i}}{E}[\log p_g(\mathbf{o})] = \int_{\mathcal{O}} p_{h,i}(\mathbf{o})\log p_g(\mathbf{o})d\mathbf{o} \approx \frac{1}{2N}\sum_{k=1}^{2N}\log p_g(\mathbf{o}_{i,k}). \tag{4.17}$$

Wszystkie otrzymane w ten sposób wyniki cząstkowe obliczeń wstawiamy do zależności (4.14) i otrzymujemy aproksymację wartości odległości pomiędzy rozkładami czyli KL. Dla spełnienia własności symetrii stosowanej miary odległości d pomiędzy rozkładami GMM $p_g(\mathbf{o})$ i $p_h(\mathbf{o})$ postaci (4.14) za ostateczną jej postać przyjmujemy zależność:

$$d(p_g, p_h) = \frac{1}{2}(KL(p_h \parallel p_q) + KL(p_g \parallel p_h)). \tag{4.18}$$

4.2.3. Wskaźnik błędnie rozpoznanych ramek

W ogólności, miara FER (ang. *Frame Error Rate*) jest tradycyjnie stosowana do oceny jakości rozpoznawania mowy na poziomie pojedynczych ramek sygnału i definiowana jest jako:

$$E_f = \frac{N_f^{(err)}}{N_f} \cdot 100\% \tag{4.19}$$

gdzie E_f oznacza wskaźnik błędu FER dla f-tego fonemu ($f = 1, ..., F$), F jest liczbą fonemów, N_f jest liczbą wszystkich ramek danego fonemu poddawanych rozpoznaniu, zaś $N_f^{(err)}$ jest liczbą ramek źle rozpoznanych.

W celu zrozumienia przyczyn błędów rozpoznawania warto uwzględnić częstość pomyłek pomiędzy poszczególnymi fonemami. W tym celu przeprowadza się rozpoznawanie jeden-do-jeden pomiędzy parami badanych fonemów i definiuje miarę $E_{f,g}$, która stanowi lokalną miarę błędu FER, określającą, jaki procent ramek fonemu f został błędnie sklasyfikowany jako fonem g:

$$E_{f,g} = \frac{N_{f,g}^{(err)}}{N_f} \cdot 100\%, \qquad f \neq g \tag{4.20}$$

gdzie $N_{f,g}^{(err)}$ stanowi liczbę ramek fonemu f błędnie sklasyfikowanych jako fonem g, N_f oznacza liczbę wszystkich ramek danego fonemu poddawanych rozpoznawaniu, a $f = 1, ..., F$ i $g = 1, ..., F$ to indeksy fonemów. Taka analiza pozwala na utworzenie macierzy błędów klasyfikacji i precyzyjne wskazanie fonemów, stanowiących główne źródło obniżenia skuteczności działania systemu.

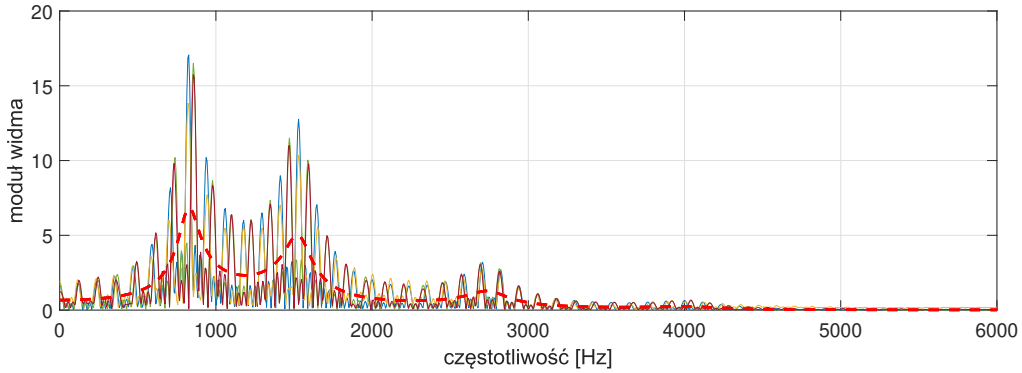
4.3. Wyniki badań i wnioski

4.3.1. Przykładowe wyniki na sygnałach modelowych

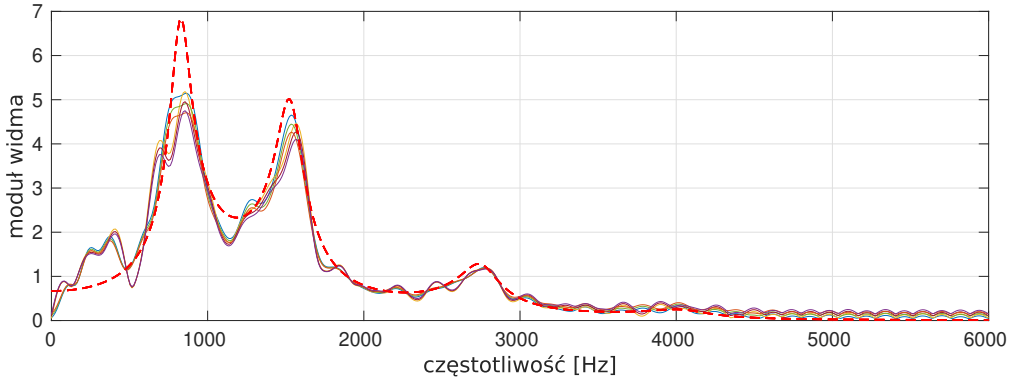
Pierwotna ocena proponowanych metod była realizowana poprzez analizę sygnałów modelowych, reprezentujących samogłoski. Wygenerowano modele każdej z analizowanych samogłosek metodą opisaną w rozdz. 4.1.1. Sygnały te poddano operacji ramkowania zgodnie z techniką opisaną w rozdz. 2.3.2. Na rys. 4.5 przedstawiono unormowane energetycznie widma kilku ramek modelowego sygnału mowy zawierających fonem a.

Na rysunku tym, widoczne są zafalowania, a maksima lokalne występują w wielokrotnościach częstotliwości podstawowej F_0. Linią czerwoną, przerywaną wykreślono moduł funkcji transmitancji modelowego kanału głosowego. Dla lepszej czytelności, widma zostały unormowane. Na wykresie widoczna jest zgodność pomiędzy położeniem maksimów lokalnych widm, a częstotliwościami środkowymi formantów.

Sygnały te poddano parametryzacji PS-HFCC opisanej w rozdz. 3.2.1. Wyniki zastosowania tej metody zaprezentowano na rys. 4.6, na którym wykreślono unormowane widma kilku



Rys. 4.5: Unormowane widma kilku ramek modelowego sygnału mowy zawierających fonem a wraz z naniesionym modułem transmitancji kanału głosowego (czerwona linia przerywana).



Rys. 4.6: Przykładowe wyniki zastosowania odpornej parametryzacji PS-HFCC. Unormowane widma amplitudowe kilku ramek modelowego sygnału mowy wraz z wykreślonym modułem transmitancji kanału głosowego (czerwona linia przerywana).

kolejnych ramek tego sygnału wraz z modułem transmitancji modelowego kanału głosowego (czerwona linia przerywana).

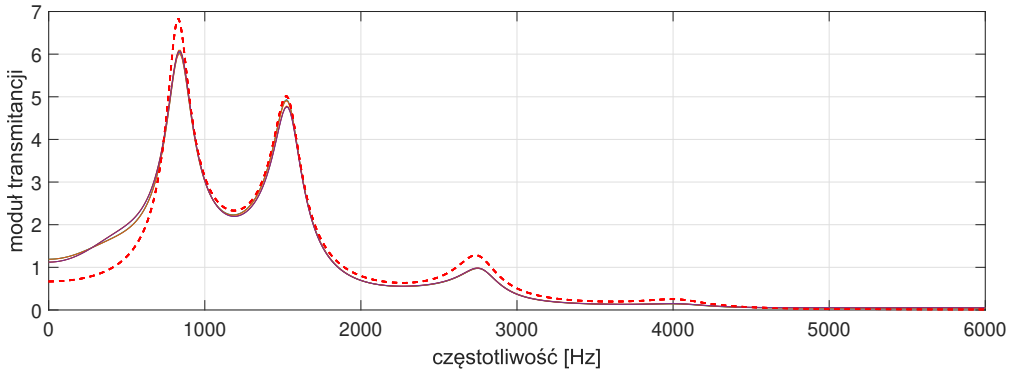
Zastosowanie zmiennej długości ramki sygnału zsynchronizowanej z okresem podstawowym T_0 spowodowało znaczne wygładzenie widm i usunięcie zafalowań. Porównanie tych wyników z widmami sygnałów przed korekcją wskazuje na skuteczność zaproponowanej metody eliminacji zafalowań spowodowanych okresowością tonu krtaniowego, co skutkuje większą zgodnością z modułem transmitancji kanału głosowego. Podobne wyniki uzyskano dla innych sygnałów modelowych.

Badania drugiej z metod, wykorzystującej filtrację odwrotną IAIF opisaną w rozdz. 3.2.2 przeprowadzono, przyjmując następujące wartości parametrów algorytmu:

- $f_c = 70$ Hz - częstotliwość graniczna górnoprzepustowego filtru FIR,
- $p = 20$ - rząd modelów LPC estymatorów transmitancji kanału głosowego $H_{v1}(z)$ i $H_{v2}(z)$,
- $g = 6$ - rząd modelu LPC estymatora pobudzenia $H_{g2}(z)$.

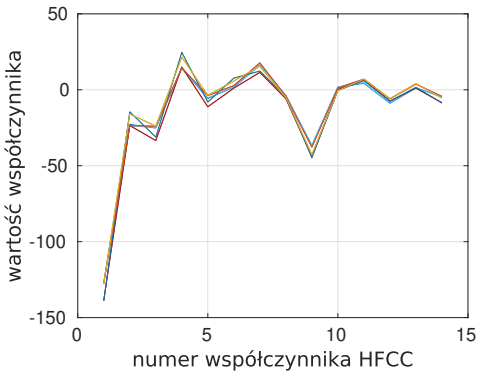
Wyniki przetwarzania tą metodą przedstawiono na rys. 4.7. Widoczne są na nim estymatory $H_{v2}(z)$ modułu funkcji transmitancji kanału głosowego wyznaczone dla tych samych ramek

modelowego sygnału wraz z wykreślonym czerwoną linią modułem transmitancji modelowego kanału głosowego.

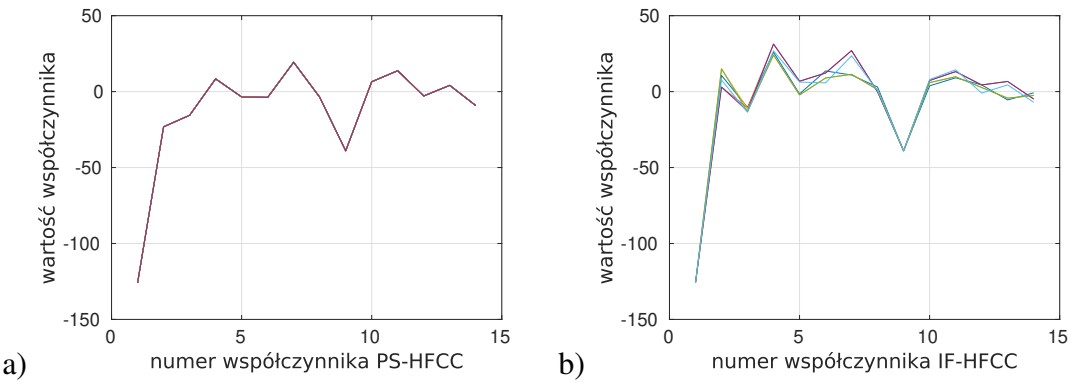


Rys. 4.7: Przykładowe wyniki zastosowania metody filtracji odwrotnej IAIF. Estymatory $H_{v2}(z)$ kilku ramek modelowego sygnału mowy wraz z wykreślonym modułem transmitancji kanału głosowego (czerwona linia przerywana).

Porównanie wyników z rys. 4.7 z widmami z rys. 4.5 także wskazuje na skuteczność eliminacji zafalowań. Bazując na widmach z rys. 4.5 - 4.7 wyznaczono współczynniki cepstralne analizowanych ramek sygnału i wykreślono je na rys. 4.8 - 4.9.



Rys. 4.8: Cepstra kilku ramek modelowego sygnału, wyznaczone z widm z rys. 4.5, uzyskane metodą HFCC



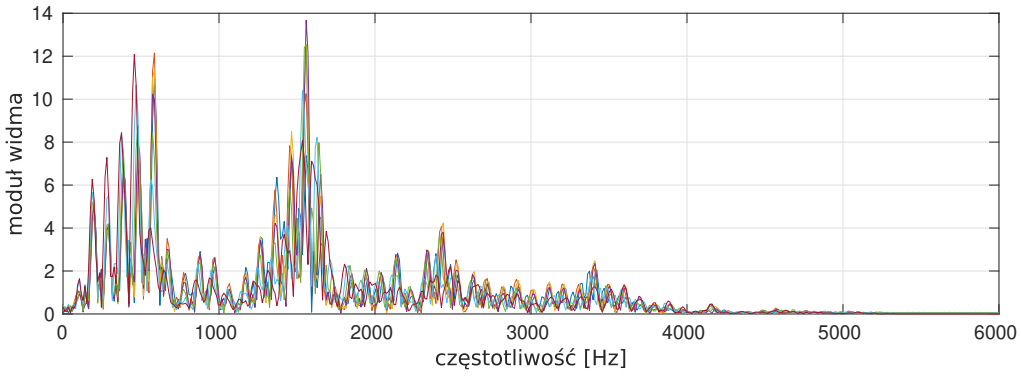
Rys. 4.9: Cepstra kilku ramek modelowego sygnału wyznaczone metodą a) PS-HFCC, b) IF-HFCC.

Na rys. 4.8 przedstawiono współczynniki HFCC obliczone z widm z rys. 4.5 z widocznymi zafalowaniami. Charakteryzują się one znacznym rozrzutem wartości współczynników pomiędzy kolejnymi ramkami sygnału. Stosując proponowane w pracy metody odporne, widoczne jest zmniejszenie wariancji na poszczególnych współrzędnych wektora obserwacji. Stosując metodę PS-HFCC, uzyskuje się zerowy rozrzut wartości współczynników $c(t,p)$, co jest związane z precyzyjną ekstrakcją okresów sygnałów modelowych. Stosując metodę IF-HFCC, uzyskano mniejszy rozrzut współczynników niż na rys. 4.8, jednak nie jest on zerowy. Związane jest to z faktem, że metoda IF-HFCC nie eliminuje efektu obcięcia.

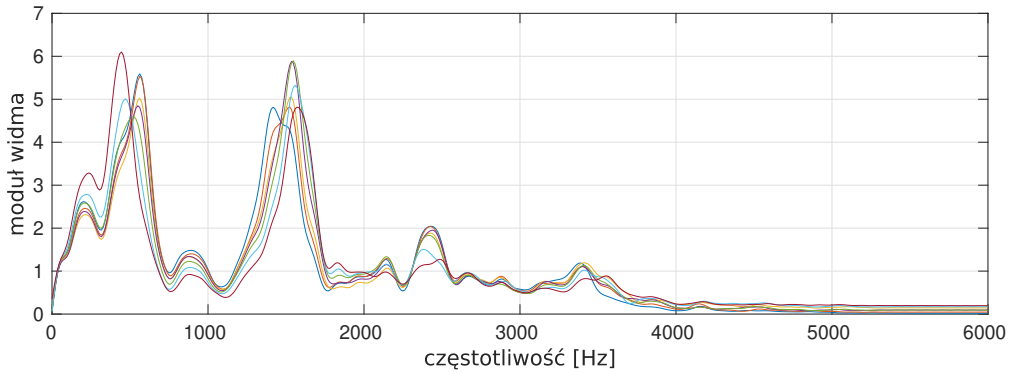
Podsumowując, analiza przeprowadzona z wykorzystaniem sygnałów modelowych potwierdza, że obie zaproponowane metody spowodowały znaczne wygładzenie widm amplitudowych i skuteczną redukcję niepożądanych zafalowań wynikających z okresowej natury dźwięcznych fragmentów mowy. Przeprowadzone eksperymenty skutkują lepszą reprezentatywnością sygnałów, a wygładzone widma stanowią wiarygodne estymatory transmitancji kanału głosowego i dobrze odzwierciedlają formantową strukturę samogłosek. Jest to niezwykle istotne także w kontekście dalszych etapów przetwarzania w systemach ARM, takich jak rozpoznawanie ciągów fonetycznych, gdzie stabilność i odporność cech akustycznych mają kluczowe znaczenie dla skuteczności algorytmów klasyfikacyjnych. Otrzymane wyniki mogą być także podstawą do opracowania algorytmu automatycznego wykrywania formantów.

4.3.2. Przykładowe wyniki dla sygnałów rzeczywistych

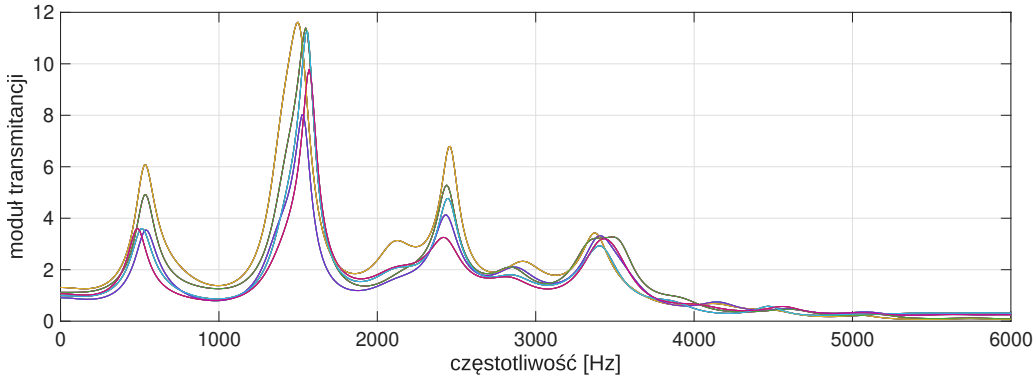
W rozdziale tym zaprezentowano przykładowe wyniki analizy ilustrujące skuteczność metod zaproponowanych w pracy w przypadku sygnałów rzeczywistych, którymi były dźwięczne fonemy mowy polskiej występujące w opisanej w rozdz. 4.1.2 bazie nagrań. Przykładowe widma $|Z(t,l)|$ kilku ramek fonemu e przedstawiono na rys. 4.10, przykłady wygładzonych widm stanowiących podstawę parametryzacji PS-HFCC na rys. 4.11, a IF-HFCC na rys. 4.12. Widma kolejnych ramek wykreślono różnymi kolorami, a dla celów porównawczych zostały one energetycznie unormowane.



Rys. 4.10: Unormowane widma kilku ramek rzeczywistego sygnału mowy zawierających fonem e zaczerpnięte z nagrań głosu tego samego mówcy.



Rys. 4.11: Przykłady wygładzonych widm kilku ramek rzeczywistego sygnału mowy zawierających fonem e, używane do parametryzacji PS-HFCC.

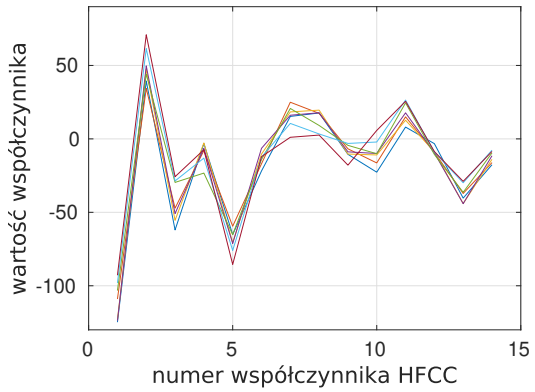


Rys. 4.12: Przykładowe estymatory $H_{v2}(z)$ kilku ramek rzeczywistego sygnału mowy zawierających fonem e, używane do parametryzacji IF-HFCC.

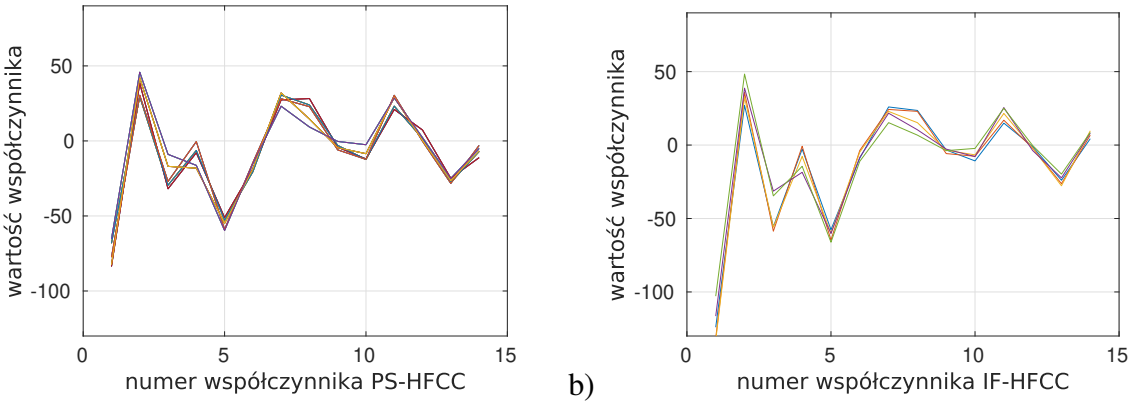
Podobnie jak przy analizie sygnałów modelowych, porównanie wykresów 4.11 - 4.12 z wykresem 4.10 wskazuje na skuteczność eliminacji zafalowań widm $|Z(t,l)|$. Podobieństwo wykresów 4.11 i 4.12 jest duże. Częstotliwości środkowe czterech pierwszych formantów odczytane z wygładzonych widm wynoszą ok. $F_1 \approx 0.50$ kHz, $F_2 \approx 1.60$ kHz, $F_3 \approx 2.45$ kHz, $F_4 \approx 3.40$ kHz i są zgodne z zakresami średnich wartości częstotliwości formantowych tego fonemu. Podobne wygładzenie widm uzyskano dla wszystkich innych analizowanych fonemów.

Bazując na widmach z rys. 4.10 - 4.12, wyznaczono współczynniki cepstralne $c(t,p)$ analizowanych ramek sygnału. Na rys. 4.13 przedstawiono współczynniki obliczone z widm bez korekcji, zawierających okresowe powielenia. Na rys. 4.14a) zaprezentowano współczynniki PS-HFCC, a na rys. 4.14 b) współczynniki wyznaczone z estymatorów $H_{v2}(z)$, a więc metodą IF-HFCC. W obu przypadkach widoczne jest zmniejszenie rozrzutu ich wartości.

Porównanie rys. 4.14 a) - b) z rys. 4.13 pozwala na wyciągnięcie wniosku, że metoda PS-HFCC daje najlepsze wyniki, ponieważ rozrzut współczynników cepstralnych jest najmniejszy spośród trzech badanych wariantów HFCC. Precyzyjne dopasowanie okna analizy i jego synchronizacja z długością okresu podstawowego pozwala na wydobycie cech niemal niezależnych od częstotliwości tonu krtaniowego F_0.



Rys. 4.13: Cepstra poszczególnych ramek sygnału mowy, zawierających fonem e, wyznaczone z widm zawierających okresowe powielenia z rys. 4.10.



Rys. 4.14: Cepstra poszczególnych ramek sygnału mowy, zawierających fonem e, wyznaczone z widm z rys. 4.11 - 4.12

.

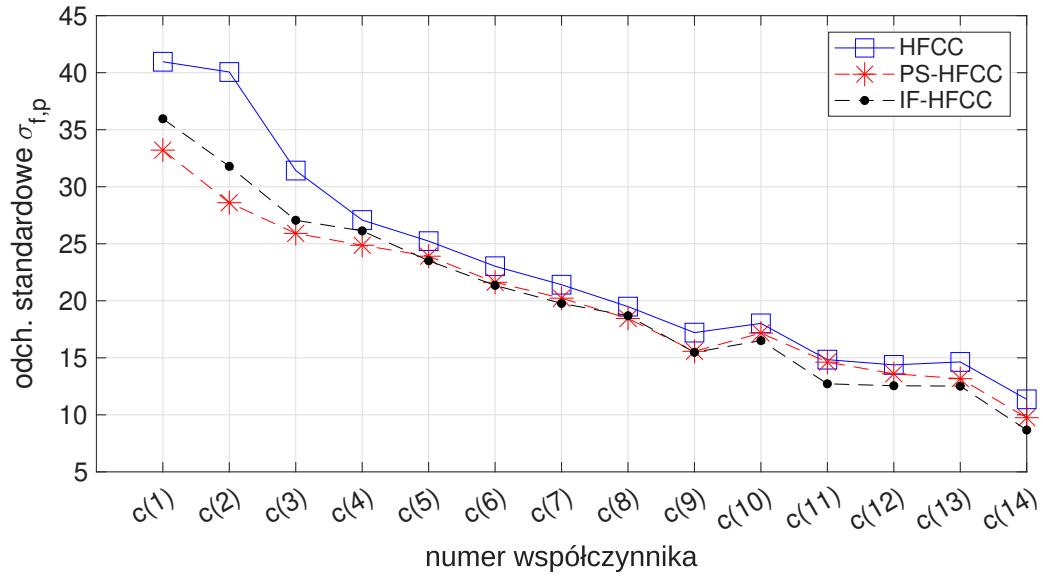
4.3.3. Globalna analiza błędów

W celu ilościowej oceny wpływu zastosowanych metod odpornej parametryzacji dla całej analizowanej bazy sygnałów, dokonano oceny ich skuteczności z wykorzystaniem miar opisanych w rozdz. 4.2. Realizacja takich badań wymagała wcześniejszego opracowania modeli akustycznych fonemów w postaci rozkładów prawdopodobieństw GMM.

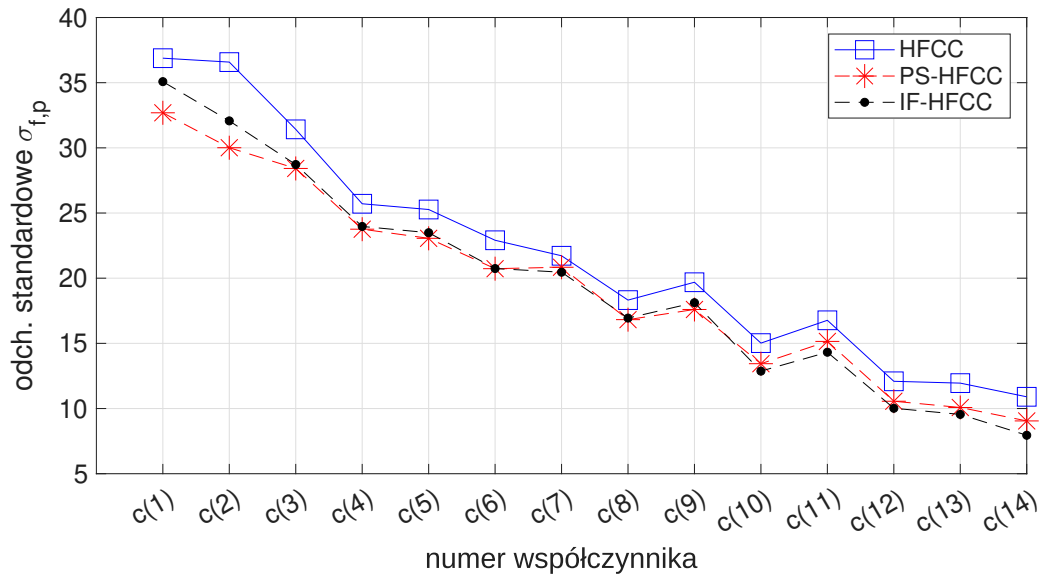
Analiza odchyleń standardowych

W pierwszym kroku porównano ze sobą wartości odchyleń standardowych $\sigma_{f,p}$ wyznaczone osobno dla wszystkich z P składowych wektorów obserwacji. Przykładowe wyniki obliczone dla samogłoski e oraz spółgłoski r zaprezentowano na rys. 4.15 - 4.16. Niebieska krzywa odnosi się do odchyleń parametrów bazowego algorytmu HFCC, czerwona do PS-HFCC, a czarna do IF-HFCC. Analiza wykresów pozwala jedynie na wizualne uchwycenie charakterystycznej tendencji redukcji rozrzutu wartości współczynników cepstralnych. Warto jednak podkreślić, że ma ona charakter systematyczny i powtarzalny, stąd szczegółowe dane liczbowe ilustrujące ten

efekt dla wszystkich analizowanych fonemów przedstawiono w tabelach 4.3 - 4.4. Dla każdego współczynnika cepstralnego zapisano w nich trzy wartości odchyleń standardowych, odnoszące się do metod parametryzacji HFCC, PS-HFCC oraz IF-HFCC. Wyniki te zostały zorganizowane w dwóch tabelach - pierwsza z nich (Tab. 4.3) dotyczy samogłosek, a druga (Tab. 4.4) spółgłosek.



Rys. 4.15: Porównanie wartości odchyleń standardowych $\sigma_{f,p}$ dla samogłoski e i trzech metod parametryzacji.



Rys. 4.16: Porównanie wartości odchyleń standardowych $\sigma_{f,p}$ dla spółgłoski r i trzech metod parametryzacji.

Tab. 4.3: Wartości odchyleń standardowych $\sigma_{f,p}$ poszczególnych samogłosek obliczone oddzielnie dla wszystkich P współrzędnych wektorów parametrów cepstralnych. Pierwsza wartość dla każdego współczynnika została obliczona z parametrów HFCC, druga z parametrów PS-HFCC, a trzecia z IF-HFCC.

wsp. \ fonem	i	y	e	a	o	u
$c(1)$	33.94	38.40	40.96	38.49	38.40	36.13
	30.94	31.16	33.21	32.53	32.91	30.89
	29.67	34.41	35.96	33.38	35.77	35.11
$c(2)$	38.32	36.47	40.06	31.83	30.44	29.67
	22.18	26.69	28.60	28.39	27.40	27.18
	27.19	28.75	31.79	29.50	29.74	29.18
$c(3)$	28.66	30.21	31.42	32.38	37.99	35.91
	22.27	25.27	25.91	29.26	33.31	32.01
	25.21	26.41	27.06	31.24	35.09	32.16
$c(4)$	23.16	25.13	27.08	28.06	27.24	26.61
	19.33	21.94	24.87	26.30	24.64	24.29
	21.44	24.01	26.13	25.51	25.19	26.53
$c(5)$	19.58	23.47	25.23	26.47	24.56	23.90
	15.62	21.72	23.91	25.15	23.80	23.04
	17.17	22.33	23.51	24.46	23.64	24.38
$c(6)$	18.18	23.42	23.03	19.36	23.39	23.07
	15.42	20.61	21.62	18.33	22.55	23.13
	16.68	20.77	21.35	18.49	23.64	24.23
$c(7)$	17.09	19.91	21.41	19.68	18.70	19.23
	15.07	17.96	20.23	18.74	17.45	17.87
	15.08	18.30	19.77	18.85	18.49	18.91
$c(8)$	17.97	17.50	19.49	17.01	18.29	18.48
	14.67	16.00	18.45	15.89	17.34	18.24
	15.41	16.26	18.69	15.85	17.88	18.90
$c(9)$	14.80	16.87	17.21	19.99	17.68	18.13
	11.90	14.99	15.57	19.14	16.92	18.20
	12.28	16.15	15.48	17.97	16.12	18.83
$c(10)$	15.26	17.42	18.02	16.19	14.99	15.79
	12.69	15.75	17.17	15.16	14.61	15.30
	12.63	15.25	16.51	13.81	13.30	15.13
$c(11)$	14.82	15.45	14.84	15.43	15.86	13.62
	12.66	14.65	14.62	14.68	14.92	13.10
	11.62	13.76	12.72	13.60	14.76	13.11
$c(12)$	12.50	14.44	14.38	13.96	13.86	12.25
	10.35	13.21	13.60	12.85	12.78	11.90
	9.50	12.12	12.55	12.32	12.47	10.89
$c(13)$	11.88	15.73	14.64	12.47	13.59	12.33
	9.27	13.53	13.17	11.10	12.66	10.76
	9.18	12.53	12.52	10.70	11.82	9.86
$c(14)$	12.37	11.70	11.36	10.33	10.84	11.43
	9.01	9.58	9.75	9.13	9.95	10.05
	7.69	9.10	8.67	8.57	8.75	8.62

Tab. 4.4: Wartości odchyleń standardowych $\sigma_{f,p}$ poszczególnych spółgłosek obliczone oddzielnie dla wszystkich P współrzędnych wektorów parametrów cepstralnych. Pierwsza wartość dla każdego współczynnika została obliczona z parametrów HFCC, druga z parametrów PS-HFCC, a trzecia z IF-HFCC.

fonem / wsp.	j	l	r	l	m	n	n'	$n,$
	42.99	35.15	36.88	39.64	34.09	35.10	36.56	38.86
$c(1)$	35.29	28.75	32.69	35.42	30.81	31.68	34.59	35.60
	35.58	32.76	35.08	36.83	28.39	30.93	33.36	34.08
	36.09	28.13	36.58	37.24	28.52	29.94	30.55	32.48
$c(2)$	26.99	25.70	30.01	30.21	26.12	27.25	26.43	31.49
	27.52	27.08	32.07	34.21	27.95	28.83	29.42	32.54
	28.03	33.81	31.42	32.87	31.79	29.51	39.24	34.54
$c(3)$	20.85	29.30	28.42	30.56	29.09	27.87	33.73	33.38
	23.83	31.32	28.72	31.29	29.90	29.40	35.46	34.28
	27.91	24.94	25.71	26.85	26.93	26.49	26.05	28.87
$c(4)$	24.33	22.60	23.76	24.43	26.01	25.34	24.43	26.42
	25.34	25.66	23.97	25.46	26.21	27.30	26.04	29.22
	23.77	18.59	25.27	24.92	24.49	22.96	20.98	24.82
$c(5)$	19.26	17.98	23.06	22.47	23.44	22.10	20.14	23.61
	21.24	19.33	23.49	22.08	21.41	22.37	19.52	23.64
	18.76	20.13	22.91	25.98	21.48	21.01	20.39	22.11
$c(6)$	16.58	19.65	20.73	24.34	21.28	20.22	20.27	21.68
	17.78	20.87	20.74	24.39	20.26	19.95	20.34	20.20
	16.79	20.22	21.72	20.06	22.39	20.97	18.76	22.85
$c(7)$	14.59	19.00	20.84	19.19	21.21	20.37	18.36	21.76
	14.52	21.19	20.46	19.23	19.28	19.15	18.14	21.07
	18.36	16.00	18.32	18.78	19.32	18.95	21.22	18.92
$c(8)$	15.65	15.79	16.82	18.07	18.88	18.66	20.87	18.60
	15.79	16.12	16.94	17.14	17.67	17.98	20.80	17.75
	15.09	15.60	19.69	20.70	19.85	19.52	18.72	17.48
$c(9)$	13.17	15.12	17.60	19.46	18.70	18.78	17.76	17.19
	13.56	15.23	18.12	19.14	17.30	17.96	17.65	16.38
	15.81	15.03	15.02	16.28	15.91	16.41	15.94	16.91
$c(10)$	14.18	15.20	13.44	15.37	15.86	15.90	15.74	16.99
	14.08	14.48	12.87	14.68	13.84	13.79	14.09	14.21
	14.08	14.23	16.77	15.60	15.23	15.17	15.78	14.84
$c(11)$	12.82	14.38	15.15	14.31	14.75	15.02	15.58	14.51
	12.17	14.01	14.32	14.01	12.20	13.05	13.93	12.30
	12.99	11.84	12.09	12.70	14.24	14.92	15.36	14.37
$c(12)$	10.80	11.61	10.56	11.75	13.63	14.00	14.47	13.76
	9.77	10.72	10.02	11.19	11.38	10.98	11.89	10.67
	12.56	11.35	11.95	12.60	12.62	13.07	13.29	11.32
$c(13)$	10.40	9.90	10.08	11.18	11.23	11.36	12.21	10.40
	10.06	9.24	9.55	10.23	9.78	8.96	10.71	7.96
	11.34	11.79	10.90	11.06	12.76	11.76	12.77	11.27
$c(14)$	9.25	10.50	9.05	9.55	11.39	10.62	11.02	9.64
	8.98	9.24	7.95	8.34	8.75	8.97	9.14	7.74

Analizując dane liczbowe, zauważyć można, że zmniejszenie wariancji parametrów PS-HFCC zaobserwowano w 97%, a IF-HFCC w 93% przypadków i były to spadki znaczące, sięgające nawet 40% względem metody referencyjnej HFCC. Z kolei zwiększenie wariancji wystąpiło w odpowiednio 3% i 7% przypadków i było mniej znaczące - sięgające 1% w metodzie PS-HFCC i 5% w metodzie IF-HFCC. Redukcja rozrzutu wartości współczynników jest największa dla kilku pierwszych współrzędnych wektorów obserwacji. Są one kluczowe w procesie parametryzacji cepstralnych, gdyż niosą najwięcej informacji fonetycznej, opisują główne właściwości widma sygnału i zawierają kluczowe informacje dotyczące formantów. Większa koncentracja wartości współczynników niższego rzędu ma szczególne znaczenie dla poprawy rozpoznawania, ponieważ mają one największy udział w modelowaniu różnic pomiędzy poszczególnymi fonemami. Poprawa jest obserwowalna także dla współczynników o wyższych indeksach (np. c(9) - c(14)), jednak jest ona mniej znacząca.

Podsumowując, mniejsze wartości odchyleń standardowych $\sigma_{f,p}$ uzyskane w obu proponowanych metodach odpornej parametryzacji świadczą o bardziej skupionych wartościach parametrów, co znajdzie odzwierciedlenie w mniejszej szerokości rozkładów prawdopodobieństwa. W efekcie, powinno to skutkować większymi różnicami pomiędzy rozkładami poszczególnych fonemów i zwiększeniem zdolności detekcji.

Analiza odległości Kullbacka-Leiblera

W kolejnym kroku wyznaczono i poddano analizie wyniki odległości Kullbacka-Leiblera. Obliczono odległości $d_1(p_g, p_h)$ pomiędzy wielowymiarowymi rozkładami GMM dla badanych fonemów przed korekcją oraz $d_2(p_g, p_h)$ po korekcji, czyli stosując metody PS-HFCC i IF-HFCC. Następnie, dla każdej pary fonemów obliczono względną zmianę wartości dywergencji Δ_{KLD}, wyrażoną w procentach, według poniższego wzoru:

$$\Delta_{KLD} = 100\% \cdot \frac{d_2(p_g, p_h) - d_1(p_g, p_h)}{d_1(p_g, p_h)}, \tag{4.21}$$

Sposób prezentacji wyników jest następujący: dla obu metod odpornej parametryzacji przygotowano macierze kwadratowe o wymiarach F x F (odpowiadających liczbie analizowanych stanów), zawierających względne zmiany odległości Δ_{KLD} pomiędzy badanymi fonemami. Ze względów edytorskich każdą macierz podzielono na 2 części: pierwsza obejmuje odległości samogłosek względem pozostałych fonemów, natomiast druga zawiera dane dotyczące spółgłosek. W rezultacie uzyskano 4 tabele (Tab. 4.5 - 4.8). Dodatkowo wprowadzono oznaczenia kolorystyczne – kolor zielony wskazuje pożądane przypadki zwiększenia odległości, natomiast kolor czerwony sygnalizuje ich zmniejszenie.

Porównanie wyników Δ_{KLD} dla metod HFCC i PS-HFCC pozwala stwierdzić, że zwiększenie odległości występuje w 78% przypadków, a zmniejszenie jedynie w 22%. W przypadku metod HFCC i IF-HFCC również dominuje poprawa – sięga ona 60% przypadków, jednak obserwowalna jest większa liczba lokalnych spadków odległości sięgająca 40% przypadków. W obu metodach częściej występuje dodatnia wartość Δ_{KLD}, sugerująca zwiększenie odległości,

Tab. 4.5: Względne różnice Δ_{KLD} odległości rozkładów GMM samogłosek pomiędzy fonemami uzyskane dla metod HFCC i PS-HFCC. Wartości wyrażone są w procentach, kolor zielony oznacza wzrost odległości po korekcji, a kolor czerwony jej zmniejszenie.

fonem / fonem	i	y	e	a	o	u
i		59.60	70.69	14.46	4.78	57.85
y	59.60		118.57	53.51	36.36	53.41
e	70.69	118.57		112.03	49.97	80.10
a	14.46	53.51	112.03		138.61	59.54
o	4.78	36.36	49.97	138.61		64.47
u	57.85	53.41	80.10	59.54	64.47	
j	37.99	−17.27	12.12	16.22	38.09	17.69
ł	−7.73	51.47	81.75	53.51	41.00	43.06
r	−18.97	−12.76	22.79	11.86	312.87	36.98
l	−35.23	11.40	34.86	15.45	85.97	30.61
m	206.03	134.49	164.06	68.32	140.20	158.84
n	30.47	179.24	239.15	177.22	30.26	92.76
n'	−55.04	−1.39	54.85	103.65	9.95	−29.87
$n,$	1.34	5.14	19.40	25.32	16.26	20.82

Tab. 4.6: Względne różnice Δ_{KLD} odległości rozkładów GMM spółgłosek pomiędzy fonemami uzyskane dla metod HFCC i PS-HFCC. Wartości wyrażone są w procentach, kolor zielony oznacza wzrost odległości po korekcji, a kolor czerwony jej zmniejszenie.

fonem / fonem	j	ł	r	l	m	n	n'	$n,$
i	37.99	−7.73	−18.97	−35.23	206.03	30.47	−55.04	1.34
y	−17.27	51.47	−12.76	11.40	134.49	179.24	−1.39	5.14
e	12.12	81.75	22.79	34.86	164.06	239.15	54.85	19.40
a	16.22	53.51	11.86	15.45	68.32	177.22	103.65	25.32
o	38.09	41.00	312.87	85.97	140.20	30.26	9.95	16.26
u	17.69	43.06	36.98	30.61	158.84	92.76	−29.87	20.82
j		135.11	−38.79	−1.07	11.54	11.90	−30.88	149.31
ł	135.11		13.80	15.26	200.62	31.84	−49.67	147.39
r	−38.79	13.80		18.38	−38.07	3.36	−36.31	−42.23
l	−1.07	15.26	18.38		59.40	−18.84	−55.60	−10.33
m	11.54	200.62	−38.07	59.40		317.44	−12.00	−11.82
n	11.90	31.84	3.36	−18.84	317.44		11.57	0.88
n'	−30.88	−49.67	−36.31	−55.60	−12.00	11.57		42.84
$n,$	149.31	147.39	−42.23	−10.33	−11.82	0.88	42.84	

Tab. 4.7: Względne różnice Δ_{KLD} odległości rozkładów GMM samogłosek pomiędzy fonemami uzyskane dla metod HFCC i IF-HFCC. Wartości wyrażone są w procentach, kolor zielony oznacza wzrost odległości po korekcji, a kolor czerwony jej zmniejszenie.

fonem / fonem	i	y	e	a	o	u
i		31.39	101.57	49.18	85.48	56.53
y	31.39		94.86	−20.67	−20.95	−6.08
e	101.57	94.86		172.26	71.59	69.39
a	49.18	−20.67	172.26		−33.33	18.49
o	85.48	−20.95	71.59	−33.33		38.41
u	56.53	−6.08	69.39	18.49	38.41	
j	203.44	−77.63	−55.92	−34.61	−22.11	58.69
ł	35.15	−34.19	24.81	5.26	−0.63	−8.47
r	−15.07	−62.10	18.15	−19.23	130.11	39.25
l	74.79	−37.16	43.14	−1.55	82.19	48.77
m	158.87	51.58	166.54	7.93	73.79	293.05
n	46.60	−18.81	164.33	108.28	45.64	1.14
n'	−28.00	−72.40	5.94	4.04	−24.07	−8.41
$n,$	14.46	−49.13	15.31	33.78	13.94	18.21

Tab. 4.8: Względne różnice Δ_{KLD} odległości rozkładów GMM spółgłosek pomiędzy fonemami uzyskane dla metod HFCC i IF-HFCC. Wartości wyrażone są w procentach, kolor zielony oznacza wzrost odległości po korekcji, a kolor czerwony jej zmniejszenie.

fonem / fonem	j	ł	r	l	m	n	n'	$n,$
i	203.44	35.15	−15.07	74.79	158.87	46.60	−28.00	14.46
y	−77.63	−34.19	−62.10	−37.16	51.58	−18.81	−72.40	−49.13
e	−55.92	24.81	18.15	43.14	166.54	164.33	5.94	15.31
a	−34.61	5.26	−19.23	−1.55	7.93	108.28	4.04	33.78
o	−22.11	−0.63	130.11	82.19	73.79	45.64	−24.07	13.94
u	58.69	−8.47	39.25	48.77	293.05	1.14	−8.41	18.21
j		40.72	−50.12	28.33	−11.14	41.53	−44.28	43.08
ł	40.72		−26.05	8.62	40.26	6.67	−53.02	20.44
r	−50.12	−26.05		−1.37	−14.99	−41.27	−62.96	−42.62
l	28.33	8.62	−1.37		53.83	1.81	−45.48	0.96
m	−11.14	40.26	−14.99	53.83		406.55	−63.87	−64.26
n	41.53	6.67	−41.27	1.81	406.55		−28.16	36.14
n'	−44.28	−53.02	−62.96	−45.48	−63.87	−28.16		−51.45
$n,$	43.08	20.44	−42.62	0.96	−64.26	36.14	−51.45	

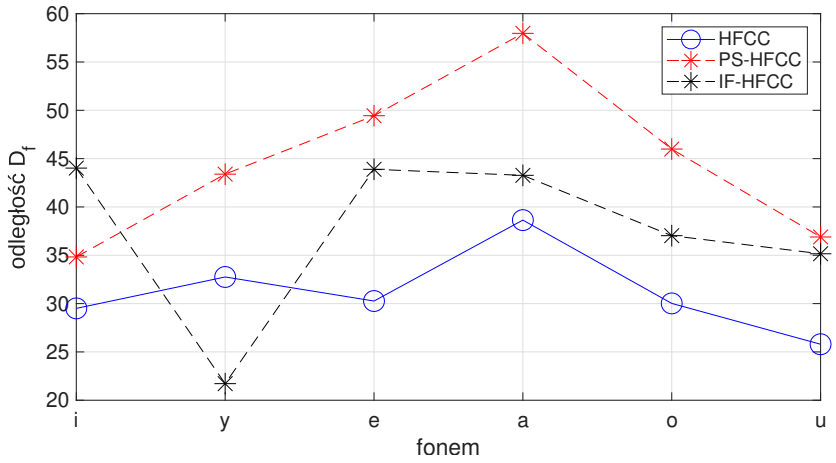
a poprawa jest nie tylko częstsza, ale znacząco silniejsza. Przykładowo, pomiędzy spółgłoskami m i n wartość Δ_{KLD} wynosi ponad 300% dla metody PS-HFCC i ponad 400% dla metody IF-HFCC. Z kolei zmniejszenie odległości nie przekroczyło odpowiednio 60% i 77%.

Analizując szczegółowo wartości Δ_{KLD} pomiędzy poszczególnymi parami badanych fonemów, warto zauważyć, że w metodzie PS-HFCC: (i) nie odnotowano zmniejszenia odległości pomiędzy parami samogłosek, (ii) zaobserwowano zaledwie kilka przypadków zmniejszenia dystansu samogłosek (i oraz y) względem spółgłosek, (iii) najczęstsze przypadki pogorszenia obserwowano pomiędzy parami spółgłosek, najwięcej dla fonemu n'. Z kolei, w metodzie IF-HFCC: (i) najwięcej przypadków zmniejszenia odległości odnotowano dla spółgłosek n', r, j oraz samogłoski y, (ii) przypadki zmniejszenia dystansu wystąpiły pomimo spadku wariancji na wszystkich składowych wektorów obserwacji. Może to sugerować, że za zmniejszenie odległości odpowiada zmiana wartości średniej poszczególnych składowych współczynników po korekcji.

Chcąc zaprezentować wyniki bardziej syntetycznie, obliczono globalny dystans D_f, sumując odległości pomiędzy danym rozkładem GMM a pozostałymi rozkładami, tj. wyznaczając wartości:

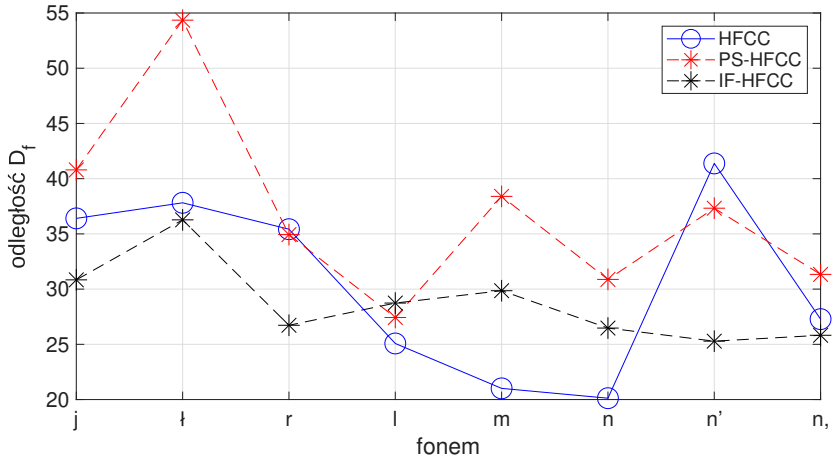
$$D_f = \sum_{i=1}^{F} d(p_f, p_i), \tag{4.22}$$

dla poszczególnych fonemów i metod parametryzacji, gdzie f oraz i są indeksami fonemu, a F ich liczbą. Miary te prezentowane są na rys. 4.17 - 4.18.



Rys. 4.17: Globalna miara odległości D_f dla samogłosek mowy polskiej przed korekcją (krzywa niebieska) i po korekcji (krzywa czarna i czerwona).

Niebieska krzywa odnosi się do modeli GMM wyznaczonych w oparciu o parametry HFCC, zaś krzywa czerwona i czarna koresponduje z metodami PS-HFCC oraz IF-HFCC. Lokalna poprawa rozróżnialności, wyrażona wartościami Δ_{KLD} przekłada się bezpośrednio na globalny efekt - suma odległości D_f dla poszczególnych fonemów wzrasta. Efekt ten jest zauważalny szczególnie w metodzie PS-HFCC, dla której wzrost odległości odnotowano dla każdego stanu za wyjątkiem n'. Dla metody IF-HFCC wygląda to gorzej, ponieważ dla kilku fonemów (samogłoski y i spółgłosek $j, ł, r, n'$ oraz n,) odnotowano zmniejszenie odległości D_f.



Rys. 4.18: Globalna miara odległości D_f dla spółgłosek mowy polskiej przed korekcją (krzywa niebieska) i po korekcji (krzywa czarna i czerwona).

Podsumowując, wzrost odległości D_f sugeruje lepszą separowalność fonemów w przestrzeni cech po zastosowaniu odpornej parametryzacji. Warto jednak pamiętać, że odległość Kullbacka-Leiblera wskazuje jedynie kierunek i wielkość zmian wprowadzanych przez metody odporne. Kluczowe znaczenie ma klasyfikacja, w związku z tym ostateczne wnioski przesądzające o przydatności proponowanych algorytmów powinny być formułowane na podstawie wyników rozpoznawania.

Analiza FER

W kolejnej części pracy przedstawiono wyniki klasyfikacji bazujące na FER, czyli błędach rozpoznawania pojedynczych ramek sygnału. Jest to inne podejście do oceny skuteczności rozpoznawania mowy w praktycznych zastosowaniach, bliższe spodziewanej efektywności całego systemu ARM. W pierwszym kroku przeprowadzono rozpoznawanie jeden-do-jeden dla wszystkich par badanych fonemów. W efekcie otrzymano macierze o rozmiarach $F \times F$, które można interpretować jako macierze pomyłek klasyfikatora z wartościami wyrażonymi w procentach. Pełne wyniki dla całej bazy sygnałów zaprezentowano w Tab.C.1 - C.4 w Dodatku C.

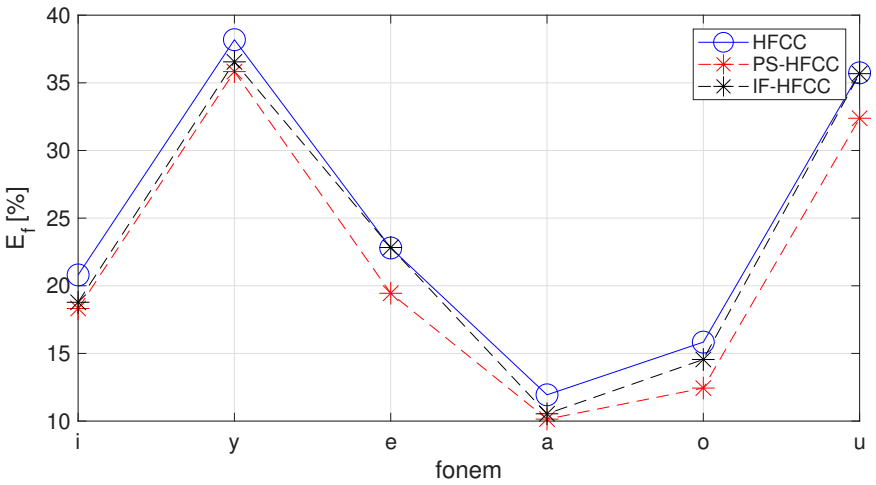
W kolejnym kroku obliczono wartości błędu E_f, zdefiniowanego wg. zależności (4.19), sumując wiersze w macierzach błędów lokalnych $E_{f,g}$, tj.:

$$E_f = \sum_{\substack{g=1 \\ g \neq f}}^{F} E_{f,g}. \tag{4.23}$$

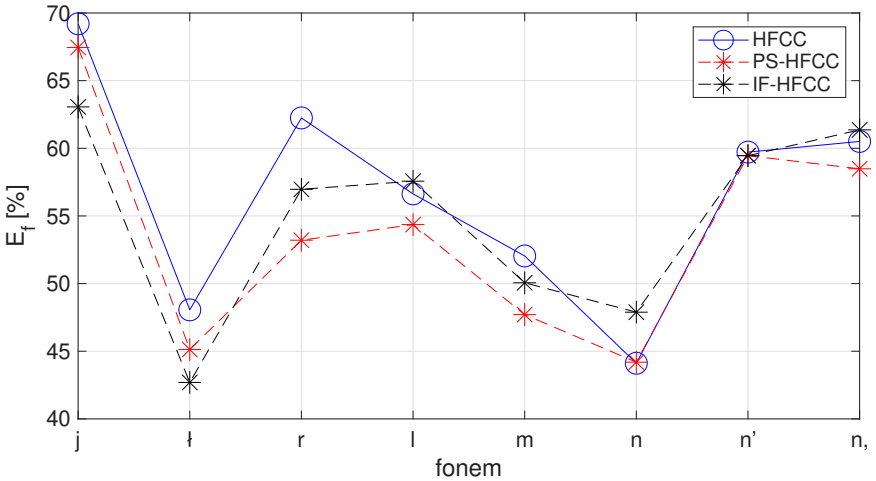
Różnice międzyosobnicze wynikające z wieku, płci, czy budowy kanału głosowego skutkują różnymi realizacjami tych samych fonemów przez poszczególnych mówców. Wprowadza to niepotrzebne rozmycie do modelu GMM, powodując jego spłaszczenie i obniżenie skuteczności klasyfikacji. Jedną z metod kompensacji tego negatywnego wpływu jest grupowanie mówców, polegające na budowaniu osobnych modeli statystycznych dla mówców charakteryzujących się podobnymi cechami osobniczymi (zob. rozdz. 2.3.1).

Z tego względu końcowe rozpoznawanie zostało przeprowadzone z podziałem mówców na 7 grup, oznaczonych jako g_k, gdzie $k = 0, \ldots, 6$ jest indeksem grupy, a grupa g_0 obejmuje

wszystkich mówców w bazie i koresponduje z poprzednimi wynikami przedstawionymi w tym rozdziale, czyli wartościami odchyleń standardowych $\sigma_{f,p}$, względnymi odległościami Δ_{KLD}, odległościami globalnymi D_f, jak i $E_{f,g}$, czyli lokalnymi błędami FER z tab. C.1 - C.4. Podział na grupy $g_1 - g_6$ został dokonany według algorytmu zaproponowanego w pracy [58]. Wykorzystuje on model UBM i zakłada adaptację wag poszczególnych składowych mieszaniny GMM danego mówcy. Wektory wag po adaptacji stanowią podstawę do przypisania poszczególnych mówców do konkretnych klastrów. Wartości miary E_f dla poszczególnych grup, z podziałem na samogłoski i spółgłoski, przedstawiono na rys. 4.19 - 4.32.

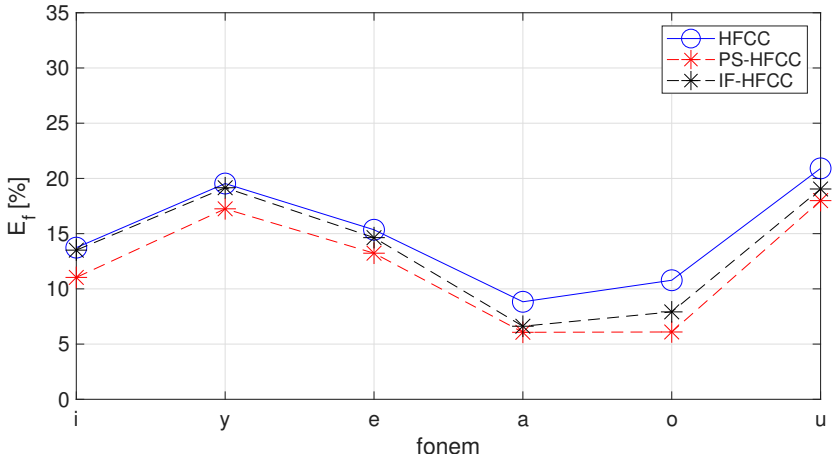


Rys. 4.19: Globalny błąd FER wyrażony miarą E_f dla samogłosek mowy polskiej. Porównanie trzech metod parametryzacji dla grupy g_0, stanowiącej zbiór wszystkich mówców.

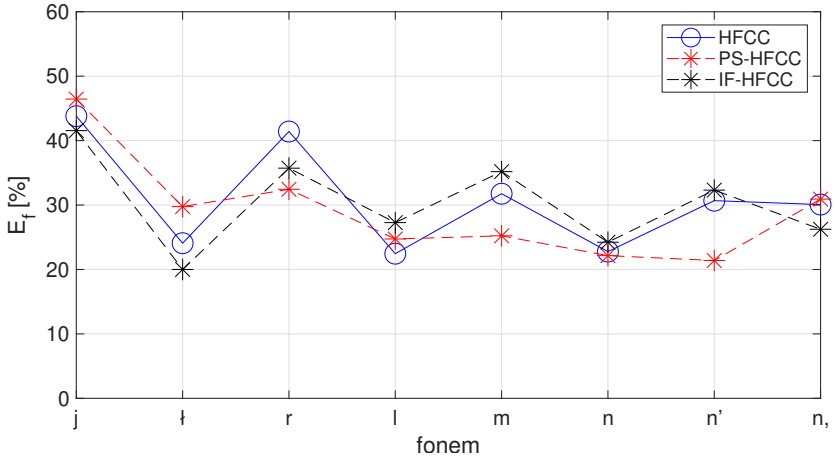


Rys. 4.20: Globalny błąd FER wyrażony miarą E_f dla spółgłosek mowy polskiej. Porównanie trzech metod parametryzacji grupy g_0, stanowiącej bazę wszystkich mówców.

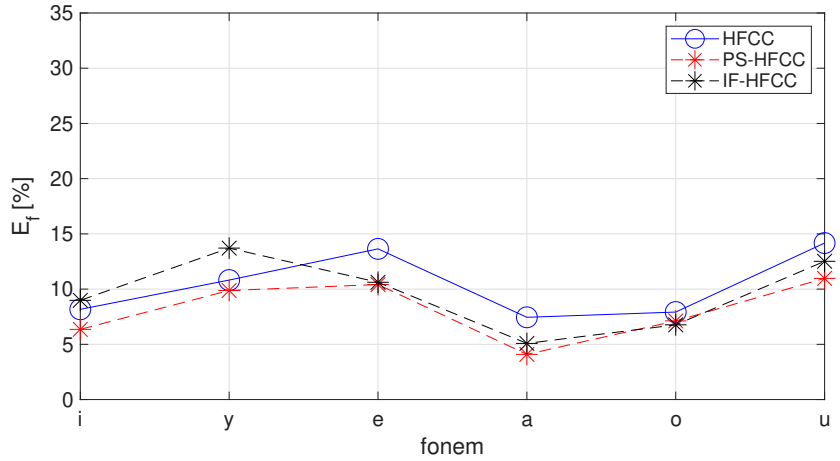
Kluczowe wnioski z analizy wyników rozpoznawania dla całej bazy mówców są następujące: (i) metoda PS-HFCC gwarantuje poprawę dla wszystkich analizowanych stanów, (ii) największe spadki błędów E_f w tej metodzie odnotowano dla fonemów: r - z ok. 62% do 53% , m - z 52% do 44,5%, e - z 22,8% do 19,4% , o - z 15,8% do 12,4%, u - z 35,5% do 32,4%, (iii) w metodzie IF-HFCC poprawę uzyskano dla większości fonemów, najbardziej znacząca jest ona dla spółgłosek – dla j z 69% do 63%., dla ł - z 48% do 42%, a także dla r z 62% do 57%, (iv) w metodzie tej obserwowalny jest wzrost błędu dla fonemu n z 44% do 47,5% i nieznaczne wzrosty dla fonemów l i n,.



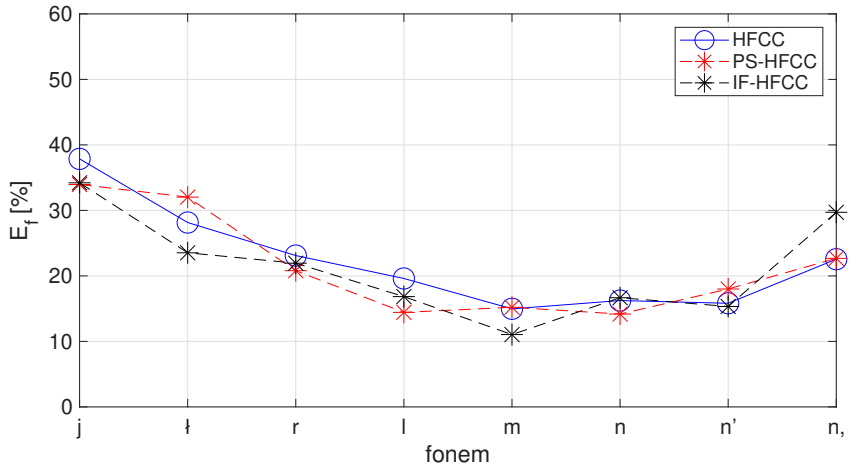
Rys. 4.21: Globalny błąd FER wyrażony miarą E_f dla samogłosek mowy polskiej. Porównanie trzech metod parametryzacji dla mówców z grupy g_1.



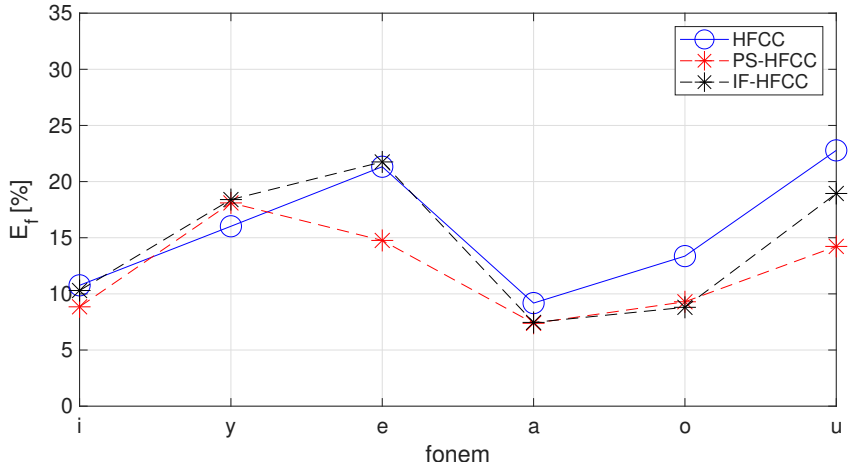
Rys. 4.22: Globalny błąd FER wyrażony miarą E_f dla spółgłosek mowy polskiej. Porównanie trzech metod parametryzacji dla mówców z grupy g_1.



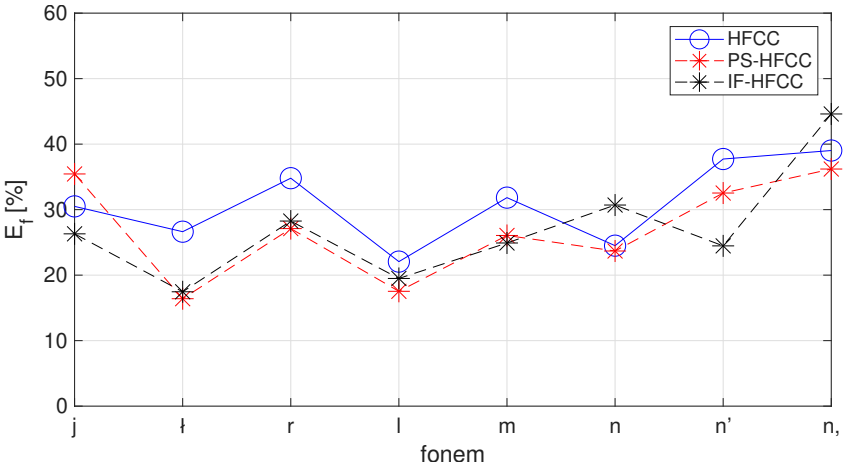
Rys. 4.23: Globalny błąd FER wyrażony miarą E_f dla samogłosek mowy polskiej. Porównanie trzech metod parametryzacji dla mówców z grupy g_2.



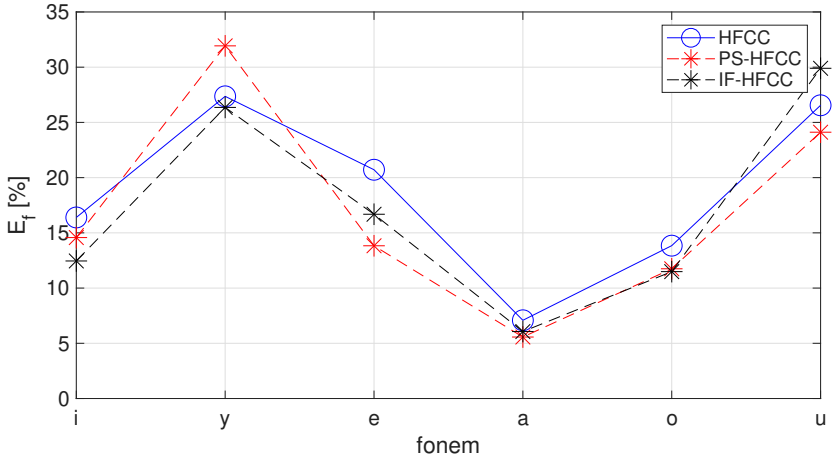
Rys. 4.24: Globalny błąd FER wyrażony miarą E_f dla spółgłosek mowy polskiej. Porównanie trzech metod parametryzacji dla mówców z grupy g_2.



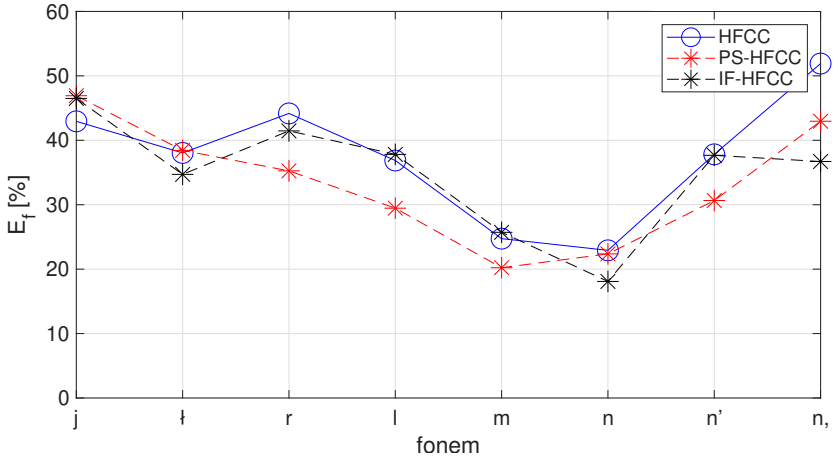
Rys. 4.25: Globalny błąd FER wyrażony miarą E_f dla samogłosek mowy polskiej. Porównanie trzech metod parametryzacji dla mówców z grupy g_3.



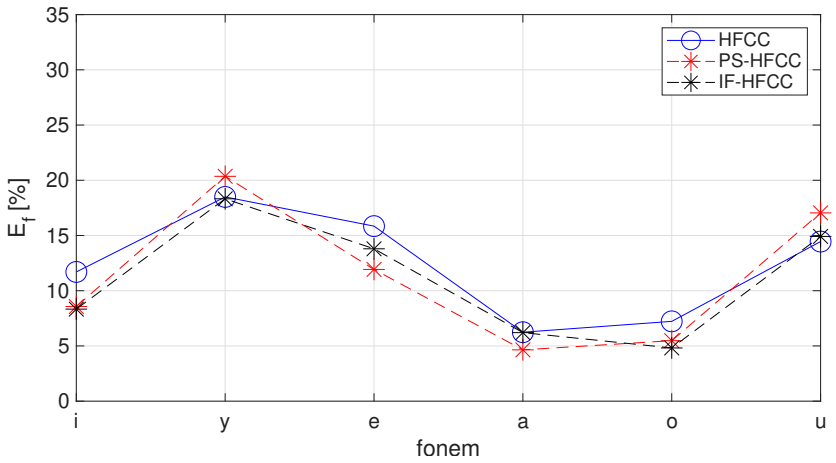
Rys. 4.26: Globalny błąd FER wyrażony miarą E_f dla spółgłosek mowy polskiej. Porównanie trzech metod parametryzacji dla mówców z grupy g_3.



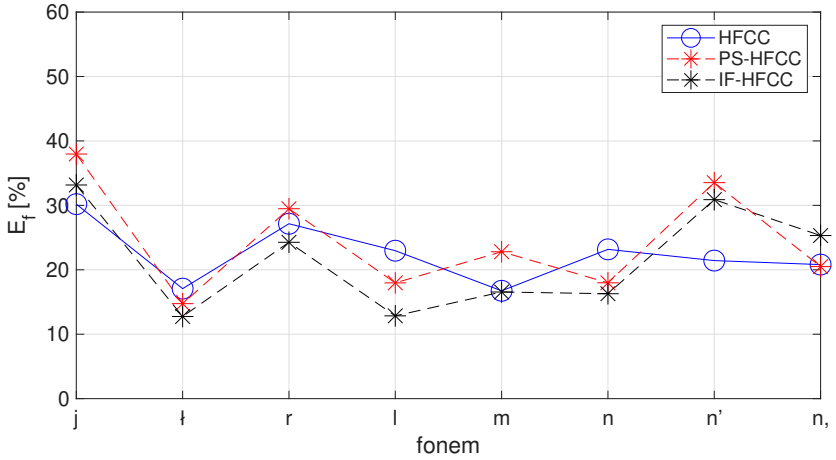
Rys. 4.27: Globalny błąd FER wyrażony miarą E_f dla samogłosek mowy polskiej. Porównanie trzech metod parametryzacji dla mówców z grupy g_4.



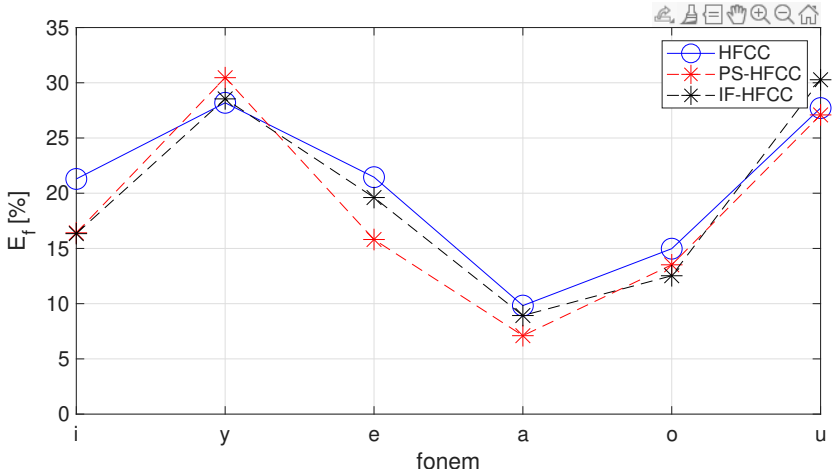
Rys. 4.28: Globalny błąd FER wyrażony miarą E_f dla spółgłosek mowy polskiej. Porównanie trzech metod parametryzacji dla mówców z grupy g_4.



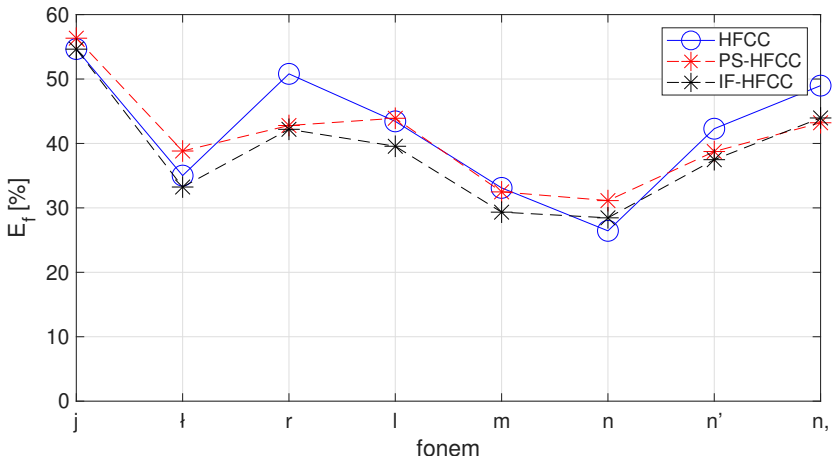
Rys. 4.29: Globalny błąd FER wyrażony miarą E_f dla samogłosek mowy polskiej. Porównanie trzech metod parametryzacji dla mówców z grupy g_5.



Rys. 4.30: Globalny błąd FER wyrażony miarą E_f dla spółgłosek mowy polskiej. Porównanie trzech metod parametryzacji dla mówców z grupy g_5.



Rys. 4.31: Globalny błąd FER wyrażony miarą E_f dla samogłosek mowy polskiej. Porównanie trzech metod parametryzacji dla mówców z grupy g_6.



Rys. 4.32: Globalny błąd FER wyrażony miarą E_f dla spółgłosek mowy polskiej. Porównanie trzech metod parametryzacji dla mówców z grupy g_6.

Na podstawie wyników rozpoznawania dla grup g_1-g_6 (rys. 4.21 - 4.32) wyciągnąć można następujące wnioski:

- wartości błędów E_f były mniejsze dla grup g_1-g_6, niż dla grupy g_0, stanowiącej bazę wszystkich mówców, co potwierdza zasadność stosowania algorytmów grupowania. Jest to szczególnie widoczne w przypadku spółgłosek, dla których wartości E_f w niektórych grupach spadają niemal o połowę,
- najmniejsze błędy rozpoznawania odnotowano dla samogłosek – w grupie g_0, stanowiącej bazę wszystkich mówców, wynoszą one ok. 10% dla samogłoski a oraz ok. 15% dla samogłosek i oraz o, natomiast po zastosowaniu algorytmu grupowania błędy te znacząco spadają - szczególnie w grupach g_1, g_2, g_5 błąd E_f dla samogłoski a spada do poziomu 4-5%, co jest wynikiem bardzo dobrym,
- zgodnie z oczekiwaniami, dla spółgłosek błędy rozpoznawania okazały się większe, przekraczające 40% - jest to szczególnie widoczne dla spółgłoski półsamogłoskowej j, drżącej r oraz nosowej n,. Analiza wyników z tabel C.1 - C.4 wskazuje, że do najczęstszych pomyłek dochodziło: (i) pomiędzy spółgłoską j oraz samogłoskami i (błąd $E_{f,g}$ ok. 30%), a także y oraz e (błąd $E_{f,g}$ ok. 15%, (ii) pomiędzy spółgłoskami nosowymi, dla których błędy $E_{f,g}$ sięgały nawet 20%,
- największym błędem w analizie globalnej obarczona jest spółgłoska j, dla której po zastosowaniu grupowania błąd E_f maleje z blisko 70% do poziomu ok. 40%, a w grupie g_3 nawet do 28%,
- zaobserwowano zróżnicowanie wyników pomiędzy poszczególnymi grupami: w grupach g_2 i g_5 uzyskano znacznie niższe wartości błędów E_f niż w grupach g_4 i g_6. Różnice te wynikają prawdopodobnie z mniejszej zmienności wewnątrzgrupowej i bardziej stabilnej artykulacji mówców w grupach 2. i 5., w przeciwieństwie do dużej zmienności osobniczej w grupach 4. i 6.

4.4. Podsumowanie

Analiza porównawcza metod PS-HFCC oraz IF-HFCC pozwala na wyciągnięcie kluczowych wniosków:

- metoda PS-HFCC gwarantuje poprawę w 80% wszystkich przypadków, a pogorszenie w 20%, a metoda IF-HFCC w odpowiednio 75% i 25% przypadków,
- spadki błędu E_f w obu metodach są nie tylko częstsze, ale bardziej znaczące,
- w metodzie PS-HFCC największą poprawę uzyskano dla fonemu ł w grupie g_3– spadek błędu E_f z poziomu 26,6% do 16,4%, a także znaczącą poprawę dla fonemów r w grupie g_0 i g_1, n w grupie g_4 i m w grupie g_3,
- w metodzie IF-HFCC odnotowano spadek błędu E_f z poziomu 51,9% do 36,7% dla fonemu n, w grupie g_4, z 37,7% do 24,4% dla n' w grupie g_3, z 22,9% do 12,9% dla l w grupie g_5,
- zaobserwowano jeden znaczący wzrost błędu dla fonemu n' w grupie g_5, podczas gdy pozostałe pojedyncze przypadki pogorszenia nie przekroczyły 6 p.p. w metodzie PS-HFCC i 7 p.p. w metodzie IF-HFCC,
- redukcja błędów po zastosowaniu metod odpornych występuje w zdecydowanej większości przypadków, jednak skala wprowadzanych zmian jest także zróżnicowana – za przykład może posłużyć fonem n,, dla którego metoda IF-HFCC gwarantuje rekordową poprawę w grupie g_4, jednak wprowadza pogorszenie w grupie g_2.

Podsumowując, wysoki poziom błędów rozpoznawania w przypadku spółgłosek uzasadnić można ich krótszym czasem trwania i mniejszą energią w porównaniu do samogłosek. Krótkotrwałe zmiany mogą być trudniejsze do uchwycenia w reprezentacji cepstralnej sygnału. Co więcej, płynne przejścia artykulacyjne pomiędzy niektórymi fonemami (jak w przypadku j-e) utrudniają jednoznaczne wyznaczenie granic między nimi, stąd możliwe błędy w segmentacji sygnału mogą odgrywać kluczową rolę w zwiększeniu liczby ramek niepoprawnie rozpoznanych.

Ponadto spółgłoski nosowe charakteryzują się dużym podobieństwem akustycznym, a obecność rezonansu nosowego może prowadzić do powstawania podobnych wektorów współczynników cepstralnych. Warto pamiętać, że wartości Δ_{KLD} pomiędzy spółgłoskami m i n były jednymi z najbardziej znaczących, jednak istotny wpływ kontekstowości sprawia, że ich widma podlegają silnym modyfikacjom, przez co model GMM napotyka trudności w zamodelowaniu tych różnic. Warto jednak zauważyć, że duże wartości błędów E_f spółgłosek nosowych były mocno redukowane przez metody odporne.

Analiza wyników w grupach zwiększa wiarygodność przeprowadzonych badań i pozwala na wyciągnięcie bardziej szczegółowych wniosków. Przy analizie globalnej wyniki są uśrednione, wskutek czego można przeoczyć, że poprawa w jednej z grup jest bardzo znacząca, podczas gdy w innej jest marginalna.

Konkludując, wpływ obu proponowanych w niniejszej pracy metod odpornej parametryzacji jest zauważalny, jednak poprawa uzyskana dla metody PS-HFCC jest częstsza.

Rozdział 5

Podsumowanie

Jednym z kluczowych aspektów związanych z zagadnieniem automatycznego rozpoznawania mowy jest parametryzacja sygnału. Efektywne przeprowadzenie tego procesu jest niezbędne do uzyskania reprezentacji sygnałów akustycznych w takiej formie, która zostanie skutecznie przetworzona przez modele klasyfikacyjne zarówno w konwencjonalnych systemach ARM, jak i tych opartych o modele głębokie oraz architektury typu E2E.

Niniejsza praca doktorska skoncentrowana była wokół tematu odpornej parametryzacji. Polega ona na takiej ekstrakcji cech z sygnału mowy, aby wpływ wielu czynników (opisanych w rozdz. 2.3.1), powodujących obniżenie skuteczności działania systemów ARM był jak najmniejszy lub zredukowany całkowicie.

Celem badawczym pracy było zaprojektowanie algorytmów parametryzacji odpornych na niepożądany wpływ zmian częstotliwości tonu krtaniowego na współczynniki cepstralne. Do osiągnięcia postawionego celu badawczego zrealizowano następujące zadania:

1. Dokonano przeglądu literatury związanej z systemami ARM z zakresu metod parametryzacji oraz ich odporności na negatywny wpływ czynników wymienionych w rozdz. 2.3.1.
2. Określono, czym powinien charakteryzować się algorytm odpornej parametryzacji w kontekście postawionego w rozdz. 3 problemu badawczego.
3. Przeprowadzono analizę teoretyczną istniejących metod parametryzacji wykorzystujących reprezentacje cepstralne sygnału, wybierając jako algorytm referencyjny parametryzację HFCC.
4. Zaproponowano dwie metody odpornej parametryzacji, powodujące wygładzenie widm ramek sygnału mowy poprzez całkowite wyeliminowanie lub kompensację zafalowań związanych z okresowością pobudzenia. Pierwsza z nich polega na analizie ramek sygnału o zmiennej długości, synchronizowanej z bieżącą wartością okresu podstawowego T_0 i uśrednianiu widm policzonych z pojedynczych okresów. Tak działającą metodę nazwano PS-HFCC. Druga, nazwana na potrzeby pracy IF-HFCC, wykorzystuje metodę LPC stosowaną w algorytmie filtracji odwrotnej IAIF i polega na parametryzacji estymatorów modułu transmitancji kanału głosowego.

5. Algorytmy zaimplementowano w środowisku Matlab i poddano je badaniom weryfikującym ich skuteczność. Oceny efektywności zaproponowanych rozwiązań dokonano na dwa sposoby. Pierwszy z nich obejmował analizę sygnałów modelowych, którymi były komputerowo wygenerowane fonemy mowy polskiej i polegał na analizie porównawczej oryginalnych widm (z wyraźnie widocznymi zafalowaniami, związanymi z kolejnymi harmonicznymi częstotliwości F_0) z wygładzonymi widmami obliczonymi po zastosowaniu metod odpornych. Drugi sposób oceny dotyczył sygnałów rzeczywistych z bazy nagrań opisanej w rozdz. 4.1.2 i polegał na analizie porównawczej wyników.
6. Do analizy sygnałów rzeczywistych przyjęto następujące miary skuteczności rozpoznawania: i) odchylenia standardowe współrzędnych wektorów obserwacji $\sigma_{f,p}$, ii) miary Δ_{KLD} i D_f wykorzystujące dywergencję Kullbacka-Leiblera jako miarę odległości między wielowymiarowymi rozkładami GMM fonemów, oraz iii) miarę FER, czyli błędy rozpoznawania pojedynczych ramek sygnału. Przeanalizowano błędy lokalne $E_{f,g}$ w rozpoznawaniu jeden-do-jeden pomiędzy parami fonemów, jak i błędy globalne E_f dla każdego fonemu.
7. Analizę dla sygnałów rzeczywistych przeprowadzono w oparciu o wyselekcjonowane z bazy nagrań fonemy dźwięczne, a to ograniczenie wynika z założeń badanych metod kompensacji.
8. Na koniec oceny jakości rozpoznawania fonemów przeprowadzono, budując osobne modele GMM dla 7 grup mówców.

Obie zaproponowane metody przyniosły w toku przeprowadzonych badań jakościową poprawę, a eliminacja zafalowań widma związanych z okresowością pobudzenia spowodowała wygładzenie widm ramek sygnału. Przełożyło się to globalnie na poprawę skuteczności rozpoznawania w sensie przyjętych miar. Dla wszystkich fonemów zaobserwowano mniejsze wartości odchyleń standardowych $\sigma_{f,p}$. Ponadto, w większości analizowanych przypadków zaobserwowano dodatnie wartości Δ_{KLD}, czyli wzrost zarówno lokalnych odległości pomiędzy rozkładami GMM poszczególnych fonemów, jak i globalnego dystansu D_f. Zwiększenie odległości sugeruje lepszą separowalność fonemów w przestrzeni cech, a to z kolei przekłada się na lepsze wyniki rozpoznawania, co zostało potwierdzone analizą FER. Spadek błędów E_f zaobserwowano dla większości analizowanych fonemów dla obu metod. W ogólności, mniejsze błędy E_f uzyskano po zastosowaniu metod odpornych dla samogłosek. Błąd E_f dla samogłoski a spada do poziomu 4-5%. Należy jednak podkreślić, że skala tej poprawy była różna w zależności od zastosowanej metody i analizowanej grupy mówców.

Spadki błędów (sięgające nawet 15 p.p.) spowodowane zastosowaniem metod odpornych są bardziej znaczące niż pojedyncze przypadki ichwzrostów. Porównanie zaproponowanych przez autora metod wskazuje, że poprawa uzyskana przy użyciu metody zmiennej długości ramki PS-HFCC jest częstsza, gdyż występuje w 80% przypadków. W metodzie IF-HFCC również odnotowano zmniejszenie E_f dla większości analizowanych fonemów (75% przypadków). Są to

spadki rekordowe (m.in. z 51,9% do 36,7% dla fonemu n), jednak częściej obserwuje się przypadki nieznacznego pogorszenia rozpoznawania, szczególnie dla spółgłosek nosowych.

Analiza wyników E_f dla grup mówców zwiększa wiarygodność przeprowadzonych badań i pozwala uchwycić zależności, które w skali globalnej nie są zauważalne, ze względu na to, że wyniki są uśrednione. Zauważalne jest jednak zróżnicowanie wyników pomiędzy poszczególnymi grupami.

Godnym uwagi jest fakt, że dokładność proponowanych metod zależy od kilku czynników. Przede wszystkim płynne przejścia artykulacyjne i krótkie czasy trwania fonemów (niektórych spółgłosek) uniemożliwiają precyzyjne wyznaczenie granic między nimi. Może to skutkować błędami segmentacji sygnału, co bezpośrednio przekłada się na zwiększenie liczby ramek niepoprawnie rozpoznanych. Po drugie, na dokładność metody PS-HFCC ma także wpływ precyzja: (i) algorytmów estymacji częstotliwości F_0, (ii) ekstrakcji pojedynczych okresów sygnału. Z kolei metoda IF-HFCC nie kompensuje efektu obcięcia.

Warto mieć na uwadze, że zarówno w konwencjonalnych, jak i hybrydowych systemach ARM wektory obserwacji wyznaczane są na podstawie klasycznych metod parametryzacji. W systemach E2E etap parametryzacji bywa całkowicie lub częściowo zintegrowany wewnątrz modeli głębokich sieci neuronowych, jednak ich wewnętrzne reprezentacje w ukrytych warstwach nadal odpowiadają klasycznym cechom akustycznym opisanym w literaturze (współczynniki MFCC w przypadku enkoderów, czy spektrogramy w sieciach CNN). W kontekście niniejszej pracy warto pamiętać, że quasiokresowość pobudzenia powoduje degradację widma sygnału, co skutkuje pogorszeniem jakości wszystkich cech wyznaczanych w oparciu o analizę widmową. W związku z powyższym, poprawa skuteczności związana z użyciem metod odpornych, nawet na etapie rozpoznawania pojedynczych ramek sygnału, jest istotna, a sam fakt poszukiwania i opracowywania takich metod jest wciąż aktualny i zasadny.

Przeprowadzone badania i zaprezentowane wyniki mogą stanowić punkt wyjścia do poprawy skuteczności w kolejnych etapach przetwarzania w systemach ARM, tj. na etapie rozpoznawania ciągów jednostek fonetycznych, słów i zdań. Zaproponowane metody spełniają pokładane w nich nadzieje i powinny zaowocować zwiększeniem efektywności kompletnego systemu ARM. Można je także uogólnić na inne metody parametryzacji oparte na reprezentacjach cepstralnych sygnału. Opisane rozwiązania mają także potencjał zmniejszenia wymagań obliczeniowych dotyczących rozmiarów architektur sieci RNN, CNN w głębokich modelach oraz redukcji wymiarowości problemu i skrócenia czasu obliczeń w systemach E2E.

Jednocześnie należy pamiętać, że na współczynniki parametryzacji, oprócz zmienności częstotliwości F_0, wpływa wiele innych czynników, związanych z cechami osobniczymi mówców, kontekstowością, rejestracją sygnału, a także inne aspekty opisane w rozdz. 2.3.1.

Literatura

[1] O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, L. Deng, G. Penn, D. Yu. Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(10):1533–1545, 2014.

[2] P. Alku. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech Communication*, 1991.

[3] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen, J. Chen, J. Chen, Z. Chen, M. Chrzanowski, A. Coates, G. Diamos, K. Ding, N. Du, E. Elsen, Z. Zhu. Deep speech 2: end-to-end speech recognition in english and mandarin. *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, strona 173–182. JMLR.org, 2016.

[4] R. Bachu, S. Kopparthi, B. Adapa, B. Barkana. *Voiced/Unvoiced Decision for Speech Signals Based on Zero-Crossing Rate and Energy*, strony 279–282. Springer Netherlands, 01 2010.

[5] C. Basztura. *Rozmawiać z komputerem*. Wydawnictwo Format, 1992.

[6] C. Becchetti, L. Ricotti. *Speech Recognition: Theory and C++ Implementation*. Wiley, 1999.

[7] B. Bouachache, P. Flandrin. Wigner-ville analysis of time-varying signals. *ICASSP '82. IEEE International Conference on Acoustics, Speech, and Signal Processing*, wolumen 7, strony 1329–1332, 1982.

[8] B. Bozkurt, B. Doval, C. d'Alessandro, T. Dutoit. Zeros of z-transform representation with application to source-filter separation in speech. *Signal Processing Letters, IEEE*, 12:344 – 347, 05 2005.

[9] C. P. Chan, P. C. Ching, T. Lee. Noisy speech recognition using de-noised multiresolution analysis acoustic features. *The Journal of the Acoustical Society of America*, 110(5):2567–2574, 11 2001.

[10] G. E. Dahl, D. Yu, L. Deng, A. Acero. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1):30–42, 2012.

[11] S. Davis, P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366, 1980.

[12] A. de Cheveigné, H. Kawahara. Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111 4:1917–30, 2002.

[13] J. R. Deller, J. G. Proakis, J. H. Hansen. *Discrete Time Processing of Speech Signals*. Prentice Hall PTR, USA, wydanie 1st, 1993.

[14] A. P. Dempster, N. M. Laird, D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[15] L. Deng, P. Kenny, M. Lennig, V. Gupta, F. Seitz, P. Mermelstein. Phonemic hidden markov models with continuous mixture output densities for large vocabulary word recognition. *IEEE Transactions on Signal Processing*, 39(7):1677–1681, 1991.

[16] V. Digalakis, D. Rtischev, L. Neumeyer. Speaker adaptation using constrained estimation of gaussian mixtures. *Speech and Audio Processing, IEEE Transactions on*, 3:357 – 366, 10 1995.

[17] L. Dong, S. Xu, B. Xu. Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, strony 5884–5888, 2018.

[18] T. Drugman, B. Bozkurt, T. Dutoit. Complex cepstrum-based decomposition of speech for glottal source estimation. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, strony 116–119, 09 2009.

[19] T. Drugman, B. Bozkurt, T. Dutoit. A comparative study of glottal source estimation techniques. *Computer Speech & Language*, 26:20–34, 01 2012.

[20] G. Fant. *Acoustic Theory of Speech Production*. De Gruyter Mouton, Berlin, Boston, 1971.

[21] S. Furui. Speaker-independent isolated word recognition based on emphasized spectral dynamics. *ICASSP '86. IEEE International Conference on Acoustics, Speech, and Signal Processing*, wolumen 11, strony 1991–1994, 1986.

[22] M. Gales, S. Young. The application of hidden markov models in speech recognition. *Foundations and Trends in Signal Processing*, 1:195–304, 01 2007.

[23] J.-L. Gauvain, L. Chin-Hui. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Transactions on Speech and Audio Processing*, 2(2):291–298, 1994.

[24] B. Gfeller, C. Frank, D. Roblek, M. Sharifi, M. Tagliasacchi, M. Velimirovic. Spice: Self-supervised pitch estimation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1118–1128, 2020.

[25] O. Ghitza. Auditory nerve representation as a basis for speech processing. *Advances in speech signal processing*, 1991.

[26] S. Gmyrek. Voiced frame detection in automatic speech recognition. *Acoustics, acousto-electronics and electrical engineering*, 2021.

[27] S. Gmyrek, R. Hossa. Improving the vowel classification accuracy using varying signal frame length. *Vibrations in Physical Systems*, 36(1), 2025.

[28] S. Gmyrek, R. Hossa. Robust speech parametrization based on pitch synchronized cepstral solutions. *International Journal of Electronics and Telecommunications*, 71(3):1–7, 2025.

[29] S. Gmyrek, R. Hossa, R. Makowski. Reducing the impact of fundamental frequency on the hfcc parameters of the speech signal. *2023 Signal Processing Symposium (SPSympo)*, strony 49–52, 2023.

[30] S. Gmyrek, R. Hossa, R. Makowski. Amplitude spectrum correction to improve speech signal classification quality. *International Journal of Electronics and Telecommunications*, 70(3):569–574, 2024.

[31] S. Gmyrek, R. Hossa, R. Makowski. The influence of the amplitude spectrum correction in the hfcc parametrization on the quality of speech signal frame classification. *Archives of Acoustics*, 50(1):59–67, 2025.

[32] J. Goldberger, H. Aronowitz. A distance measure between gmms based on the unscented transform and its application to speaker recognition. *Interspeech 2005*, 2005.

[33] A. Graves. Sequence transduction with recurrent neural networks, 2012.

[34] A. Graves, S. Fernández, F. Gomez, J. Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural 'networks. *ICML 2006 - Proceedings of the 23rd International Conference on Machine Learning*, 2006:369–376, 01 2006.

[35] T. J. Hazen. A comparison of novel techniques for rapid speaker adaptation. *Speech Commun.*, 31(1):15–33, Maj 2000.

[36] N. Henrich Bernardoni, C. d'Alessandro, B. Doval, M. Castellengo. On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation. *The Journal of the Acoustical Society of America*, 115:1321–32, 04 2004.

[37] H. Hermansky. Perceptual linear predictive (plp) analysis of speech. *The Journal of the Acoustical Society of America*, 87:1738–1752, 1990.

[38] W. Hess. *Pitch Determination of Speech Signals: Algorithms and Devices*. Springer Series in Information Sciences. Springer Berlin Heidelberg, 2012.

[39] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, B. Kingsbury. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

[40] C. T. Ishi, H. Ishiguro, N. Hagita. Automatic extraction of paralinguistic information using prosodic features related to f0, duration and voice quality. *Speech Communication*, 50(6):531–543, 2008.

[41] C. Jankowski, H.-D. Vo, R. Lippmann. A comparison of signal processing front ends for automatic word recognition. *IEEE Trans. Speech Audio Process.*, 3(4):286–293, 1995.

[42] W. Jassem. *Podstawy fonetyki akustycznej*. Państwowe Wydawn. Naukowe, 1973.

[43] S. Julier, J. Uhlmann. Unscented filtering and nonlinear estimation. *Proceedings of the IEEE*, 92(3):401–422, 2004.

[44] D. Jurafsky, J. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, 02 2008.

[45] S. Kadambe, G. Boudreaux-Bartels. Application of the wavelet transform for pitch detection of speech signals. *IEEE Transactions on Information Theory*, 38(2):917–924, 1992.

[46] H. K. Kathania, S. R. Kadiri, P. Alku, M. Kurimo. A formant modification method for improved asr of children's speech. *Speech Communication*, 136:98–106, 2022.

[47] H. Kawahara, A. de Cheveigné, H. Banno, T. Takahashi, T. Irino. Nearly defect-free f0 trajectory extraction for expressive speech modifications based on straight. *Interspeech*, 2005.

[48] J. W. Kim, J. Salamon, P. Li, J. P. Bello. Crepe: A convolutional representation for pitch estimation, 2018.

[49] H. Kobayashi, T. Shimamura. A weighted autocorrelation method for pitch extraction of noisy speech. *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100)*, wolumen 3, strony 1307–1310 vol.3, 2000.

[50] J. Koehler, N. Morgan, H. Hermansky, H. Hirsch, G. Tong. Integrating rasta-plp into speech recognition. *Proceedings of ICASSP '94. IEEE International Conference on Acoustics, Speech and Signal Processing*, i:I/421–I/424 vol.1, 1994.

[51] T.-W. Kuan, A.-C. Tsai, P.-H. Sung, J.-F. Wang, H.-S. Kuo. A robust bfcc feature extraction for asr system. *Artificial Intelligence Research*, 5:14–23, 2016.

[52] R. Kuhn, J.-C. Junqua, P. Nguyen, N. Niedzielski. Rapid speaker adaptation in eigenvoice space. *IEEE Transactions on Speech and Audio Processing*, 8(6):695–707, 2000.

[53] S. Kullback. *Information Theory and Statistics*. Dover Publications, 1997.

[54] C. Leggetter, P. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Comput. Speech Lang.*, 9:171–185, 1995.

[55] J. Lim, M. Scordilis. New and imprived pitch determination for the imbe vocoder. *Proceedings of the Fourth Australian International Conference on Speech Science and Technology, SST-92, Brisbane, Australia,*, strony 375–380, 1992.

[56] Y. Liu, P. Wu, A. Black, G. Anumanchipalli. A fast and accurate pitch estimation algorithm based on the pseudo wigner-ville distribution, 10 2022.

[57] R. Makowski. *Automatyczne rozpoznawanie mowy: wybrane zagadnienia*. Oficyna Wydawnicza Politechniki Wrocławskiej, 2011.

[58] R. Makowski, R. Hossa. An effective speaker clustering method using ubm and ultra-short training utterances. *Archives of Acoustics*, 41(1):107–118, 2016.

[59] R. Makowski, R. Hossa. Voice activity detection with quasi-quadrature filters and gmm decomposition for speech and noise. *Applied Acoustics*, 166:107344, 09 2020.

[60] S. Mallat. *A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way*. Academic Press, Inc., USA, wydanie 3rd, 2008.

[61] J. Markel. The sift algorithm for fundamental frequency estimation. *IEEE Transactions on Audio and Electroacoustics*, 20:367–377, 1972.

[62] M. Mauch, S. Dixon. Pyin: A fundamental frequency estimator using probabilistic threshold distributions. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, strony 659–663, 2014.

[63] J. Mcdonough, T. Schaaf, A. Waibel. Speaker adaptation with all-pass transforms. *Speech Communication*, 42:75–91, 01 2004.

[64] Y. Medan, E. Yair, D. Chazan. Super resolution pitch determination of speech signals. *IEEE Trans. Signal Process.*, 39:40–48, 1991.

[65] A. N. Mishra, N. Awasthy. Hybrid features for speaker independent hindi speech recognition. *International Journal of Scientific & Engineering Research*, 4, 12 2013.

[66] P. Moreno, B. Raj, S. R.M. A vector taylor series approach for environment-independent speech recognition. *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, 2:733–736 vol. 2, 1996.

[67] B. Morre, B. R. Glasberg. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *The Journal of the Acoustical Society of America*, 74:750–753, 1983.

[68] E. Moulines, F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5):453–467, 1990. Neuropeech '89.

[69] P. Mrówka. Algorytmy kompensacji warunków transmisyjnych i cech osobniczych mówcy w systemach automatycznego rozpoznawania mowy, 2007.

[70] P. Mrówka, R. Makowski. Normalization of speaker individual characteristics and compensation of linear transmission distortions in command recognition systems. *Archives of Acoustics*, 33(2):221–242, 2008.

[71] M. Naito, L. Deng, Y. Sagisaka. Speaker clustering for speech recognition using vocal tract parameters. *Speech Communication*, 36:305–315, 03 2002.

[72] A. M. Noll. Cepstrum pitch determination. *The Journal of the Acoustical Society of America*, 41(2):293–309, 02 1967.

[73] T. T. Nway, T. Shimamura. Weighted autocorrelation with convolutional neural network for noisy speech pitch estimation. *2024 IEEE 13th Global Conference on Consumer Electronics (GCCE)*, strony 510–511, 2024.

[74] A. Oppenheim, R. Schafer. *Digital Signal Processing*. Prentice Hall international editions. Prentice-Hall, 1975.

[75] Training acoustic models using connectionist temporal classification. Patent US10229672B1 granted on March 12, 2019 by the United States Patent and Trademark Office. Assignee: Google LLC. `https://patents.google.com/patent/US10229672B1/en`.

[76] Attention-based sequence transduction neural networks. Patent US10452978B2 granted in 2019 by the United States Patent and Trademark Office. Assignee: Google LLC. `https://patents.google.com/patent/US10452978B2/en`.

[77] System and method for automatic speech signal processing. Patent US8442821B1 granted on May 21, 2013 by the United States Patent and Trademark Office. Assignee: Google LLC. `https://patents.google.com/patent/US8442821B1/en`.

[78] System and method for automatic speech recognition using multiple hypotheses and confidence scoring. Patent US8775177B1 granted on July 29, 2014 by the United States

Patent and Trademark Office. Assignee: Google LLC. https://patents.google.com/patent/US8775177B1/en.

[79] J. M. Pedro, B. Eberman. A new algorithm for robust speech recognition: the delta vector taylor series approach. *Fifth European Conference on Speech Communication and Technology, EUROSPEECH 1997, Rhodes, Greece, September 22-25, 1997*, strony 2599–2602. ISCA, 1997.

[80] M. Plumpe, T. Quatieri, D. Reynolds. Modeling of the glottal flow derivative waveform with application to speaker identification. *IEEE Transactions on Speech and Audio Processing*, 7(5):569–586, 1999.

[81] N. V. Prasad, S. Umesh. Improved cepstral mean and variance normalization using bayesian framework. *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, strony 156–161, 2013.

[82] T. Quatieri. *Discrete-Time Speech Signal Processing: Principles and Practice*. Pearson Education, 2008.

[83] L. Rabiner, M. Cheng, A. Rosenberg, C. McGonegal. A comparative performance study of several pitch detection algorithms. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(5):399–418, 1976.

[84] L. Rabiner, B. Juang. *Fundamentals of Speech Recognition*. Prentice-Hall Signal Processing Series: Advanced monographs. PTR Prentice Hall, 1993.

[85] L. Rabiner, R. Schafer. *Theory and Application of Digital Speech Processing*. Prentice Hall Press, 1978.

[86] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, P. Alku. Hmm-based speech synthesis utilizing glottal inverse filtering. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19:153 – 165, 02 2011.

[87] J. Rajnoha, P. Pollak. Modified feature extraction methods in robust speech recognition. *2007 17th International Conference Radioelektronika*, strony 1–4, 2007.

[88] B. Resch, M. Nilsson, A. Ekman, W. Kleijn. Estimation of the instantaneous pitch of speech. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15:813 – 822, 04 2007.

[89] D. Reynolds. *Gaussian Mixture Models*, strony 659–663. Springer US, Boston, MA, 2009.

[90] H. Sak, A. Senior, F. Beaufays. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, strony 338–342, 01 2014.

[91] H. Sak, A. Senior, K. Rao, O. Irsoy, A. Graves, F. Beaufays, J. Schalkwyk. Learning acoustic frame labeling for speech recognition with recurrent neural networks. *ICASSP*, strony 4280–4284, 04 2015.

[92] K. Shinoda, C.-H. L. Structural map speaker adaptation using hierarchical priors. *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, strony 381–388, 1997.

[93] K. Shinoda, C.-H. L. Unsupervised adaptation using structural bayes approach. *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, wolumen 2, strony 793–796 vol.2, 1998.

[94] M. Skowronski, J. Harris. Human factor cepstral coefficients. *The Journal of the Acoustical Society of America*, 112, 11 2002.

[95] Y. Stylianou. Applying the harmonic plus noise model in concatenative speech synthesis. *IEEE Transactions on Speech and Audio Processing*, 9(1):21–29, 2001.

[96] Y. Suk, S. Choi, H. Lee. Cepstrum third-order normalization method for noisy speech recoginition. *IEEE Electronic Letters*, 35(7):527 – 528, 1999.

[97] Y. H. Suk, S. H. Choi. A cepstral pdf normalization method for noise robust speech recognition. S. Lin, X. Huang, redaktorzy, *Advances in Computer Science, Environment, Ecoinformatics, and Education*, strony 34–39, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[98] E. Trentin, M. Gori. A survey of hybrid ann/hmm models for automatic speech recognition. *Neurocomputing*, 37(1):91–126, 2001.

[99] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin. Attention is all you need. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, strona 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.

[100] O. Viikki, K. Laurila. Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech Communication*, 25(1):133–147, 1998.

[101] T. Waaramaa, A. Laukkanen, M. Airas, P. Alku. Perception of emotional valences and activity levels from vowel segments of continuous speech. *Journal of Voice*, 24(1):30–38, 2010.

[102] J. Walker, P. Murphy. A review of glottal waveform analysis, 01 2005.

[103] L. Welling, H. Ney, S. Kanthak. Speaker adaptive modeling by vocal tract normalization. *IEEE Transactions on Speech and Audio Processing*, 10(6):415–426, 2002.

[104] D. Wong, J. Markel, A. Gray. Least squares glottal inverse filtering from the acoustic speech waveform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(4):350–355, 1979.

[105] P. Woodland. Speaker adaptation for continuous density hmms: A review. *ITRW on Adaptation Methods for Speech Recognition*, 01 2001.

[106] H. Yin, V. Hohmann, C. Nadeu. Acoustic features for speech recognition based on gammatone filterbank and instantaneous frequency. *Speech Communication*, 53(5):707–715, 2011. Perceptual and Statistical Audition.

[107] D. Yu, L. Deng. *Automatic Speech Recognition: A Deep Learning Approach*. Springer Publishing Company, Incorporated, wydanie 1st, 2016.

[108] L. Yu, F. Xu, Y. Qu, K. Zhou. Speech emotion recognition based on multi-dimensional feature extraction and multi-scale feature fusion. *Applied Acoustics*, 216:109752, 2024.

[109] A. Zambrzycka. Adaptacja w systemach automatycznego rozpoznawania mowy., 2020.

[110] J. Zarzycki. *Cyfrowa filtracja ortogonalna sygnałów losowych*. WNT, 1998.

[111] Y. Zhang, W. Chan, N. Jaitly. Very deep convolutional networks for end-to-end speech recognition. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, strony 4845–4849, 2017.

[112] Y. Zhang, M. Pezeshki, P. Brakel, S. Zhang, Y. Bengio, A. Courville. Towards end-to-end speech recognition with deep convolutional neural networks. *Proc. Interspeech 2016*, 01 2017.

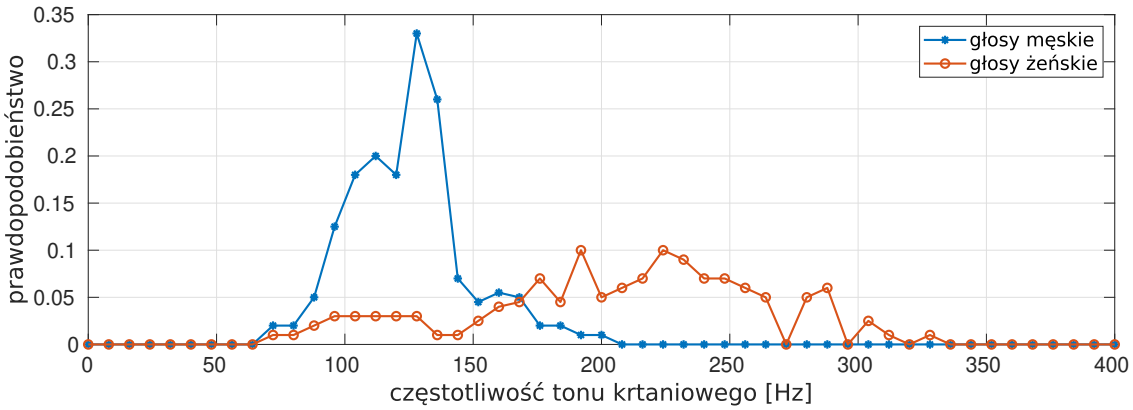
[113] T. P. Zieliński. Cyfrowe przetwarzanie sygnałów : od teorii do zastosowań.

[114] Y. Zouhir, M. Zarka, K. Ouni, L. Ouni. Power wavelet cepstral coefficients (pwcc): An accurate auditory model-based feature extraction method for robust speaker recognition. *IEEE Access*, PP:1–1, 01 2025.

Dodatek A

Estymacja częstotliwości tonu krtaniowego

Wymaganie dotyczącej pełnej znajomości bieżącej wartości okresu tonu podstawowego T_0 wymusza zastosowanie w praktyce prostej i efektywnej metody jego estymacji. Częstotliwość tonu krtaniowego ludzkiego głosu waha się w przedziale od ok. 50 do 200 Hz dla mężczyzn i od ok. 200 do 400 Hz dla kobiet [42]. Granice te nie są sztywne i możliwe są wartości je przekraczające, jednak w literaturze to właśnie taki zakres stanowi podstawę do wyznaczania częstotliwości podstawowej. Unormowane histogramy częstotliwości tonu krtaniowego głosów męskich i żeńskich wykorzystywanej w pracy bazy przedstawiono na rys. A.1 [57].



Rys. A.1: Unormowane histogramy częstotliwości tonu krtaniowego dla głosów damskich (krzywa czerwona) i męskich (krzywa niebieska) wykorzystywanej w pracy bazy.

Zagadnienie estymacji F_0 znajduje szerokie zastosowanie, szczególnie w syntezie mowy, przetwarzaniu sygnału pobudzenia [86] [95], analizie prozodii mowy [40], ekstrakcji melodii, modyfikacji mowy (ang. *speech manipulation*) [68] [47].

A.1. Klasyfikacja algorytmów estymacji

Algorytmy estymacji F_0 (ang. *pitch tracking*) podzielić można na dwie grupy: algorytmy zdarzeniowe, w których okres podstawowy T_0 wyznaczany jest na podstawie pomiaru odległości między charakterystycznymi elementami sygnału, nazywanymi zdarzeniami, związanymi ze zwieraniem się strun głosowych [45][88], oraz algorytmy blokowe, które zakładają lokalną stacjonarność sygnału tj. niezmienność częstotliwości podstawowej wewnątrz fragmentu sygnału o ustalonej długości. W systemach ARM powszechnie wykorzystywane są algorytmy blokowe. Wśród nich wyróżnić można [38] [57]:

- **algorytmy widmowe**, które oparte są na reprezentacjach czasowo-częstotliwościowych sygnału i transformacji falkowej [45]. Klasyczne metody z tej grupy okazały się nieskuteczne głównie ze względu na problemy związane z niemożnością jednoczesnej poprawy rozdzielczości czasowej i częstotliwościowej, co wiąże się z powstawaniem licznych artefaktów i błędów estymacji. Jednym z efektywnych i niezawodnych podejść reprezentujących tę grupę jest metoda wykorzystująca dystrybucję Wignera-Ville'a (WVD) [60] [7], a także jej zoptymalizowana i ulepszona wersja pseudodystrybucji Wignera-Ville'a (PWVD) [56]. Charakteryzują się one dużą rozdzielczością, opierają się na transformacji falkowej oraz filtracji cepstrum w celu eliminacji błędów estymacji.
- **algorytmy cepstralne**, wykorzystujące fakt, że cepstrum mocy ma właściwości separacji informacji zawartych w splecionych wielkościach - składowych modelu sygnału mowy [74]. Dla systemu liniowego inercyjnego cepstrum mocy wyrazić można za pomocą sumy komponentów reprezentujących ton krtaniowy i kanał głosowy. Informacje dotyczące tych komponentów zajmują różne zakresy w dziedzinie pseudoczasu (ang. quefrequency), stąd estymacja F_0 polega na detekcji maksimum lokalnego odpowiadającego częstotliwości tonu krtaniowego [72] [57].
- **algorytmy wykorzystujące prognozę LPC**, w których oblicza się funkcję autokorelacji sygnału błędu prognozy liniowej. Jedną ze znanych metod estymacji reprezentujących tę grupę jest SIFT (ang. Simplified Inverse Filtering Technique) [61], w której podstawą do obliczenia częstotliwości F_0 jest maksimum lokalne autokorelacji poszukiwane w przedziale opóźnień od 2.5 do 20 ms, co odpowiada zakresowi częstotliwości od 50 do 400 Hz.
- **algorytmy korelacyjne**, w których podstawę do estymacji częstotliwości F_0 stanowić mogą: (i) funkcja korelacji wzajemnej, (ii) funkcja autokorelacji, (iii) unormowany współczynnik korelacji, (iv) modyfikacje ww. podejść
- **metody wykorzystujące modele głębokich sieci neuronowych i algorytmy uczenia maszynowego**, które w ostatnim czasie, ze względu na dynamiczny rozwój tej dziedziny nauki są popularne i efektywne. Do tej grupy zaliczyć można metodę CREPE (ang. *Convolutional Representation for Pitch Estimation*) [48], wykorzystującą konwolucyjne sieci neuronowe oraz metodę SPICE (ang. *Self-supervised Pitch Estimation*) [24], wykorzystującą uczenie samona-

dzorowane (ang. *self-supervised*), w której model uczy się reprezentacji sygnału na podstawie struktury danych bez konieczności ich wcześniejszej segmentacji i etykietyzacji.

A.2. Algorytmy korelacyjne estymacji częstotliwości podstawowej

Zarówno algorytmy cepstralne, jak i te, wykorzystujące podejście LPC, są bezużyteczne w warunkach znacznego zaszumienia sygnału. Spośród wymienionych powyżej metod estymacji, dużą popularnością w literaturze cieszą się **algorytmy korelacyjne**.

Jeden z takich algorytmów wykorzystuje zestaw funkcji grzebieniowych, mających niezerowe wartości tylko w wielokrotnościach hipotetycznej częstotliwości F_0 z zakresu od 50 do 400 Hz. Z racji, że dla reprezentacji widmowej dźwięcznego fragmentu mowy charakterystyczne są zafalowania (ang. ripples) rozłożone w miarę równomiernie, w wielokrotnościach częstotliwości tonu krtaniowego, za podstawę do wyznaczenia estymaty F_0 wykorzystuje się częstotliwość funkcji grzebieniowej, dla której unormowany współczynnik korelacji widma sygnału z nią samą jest maksymalny [55].

Inny algorytm bazuje na okienkowaniu sygnału mowy za pomocą dwóch przylegających do siebie okien prostokątnych o długości k próbek i obliczaniu unormowanego współczynnika korelacji pomiędzy nimi. Jeżeli długość okna k będzie równa okresowi podstawowemu T_0, to podobieństwo między analizowanymi fragmentami sygnału będzie bardzo silne [64].

W celu zwiększenia dokładności wyznaczania okresu T_0, szczególnie w warunkach mniejszych częstotliwości próbkowania i, co za tym idzie, gorszej rozdzielczości analizy, stosuje się inteprolację. Algorytmy korelacyjne są również bardzo podatne na tzw. błędy grube. Gdy wartości okresu podstawowego są małe, to maksimów lokalnych korelacji w interesującym nas przedziale opóźnień może być więcej i nie zawsze największa wartość maksimum odpowiada F_0. Biorąc pod uwagę niestacjonarność sygnału i możliwą małą wartość częstotliwości próbkowania, maksimum korelacji może wystąpić w wielokrotnościach okresu podstawowego, co skutkuje błędnym oszacowaniem częstotliwości F_0.

Drugim problemem występującym podczas estymacji jest występowanie wartości maksymalnej korelacji dla opóźnienia $k = \frac{T_0}{2}$, co jest związane z dominującą wartością amplitudy pierwszego formantu, która dla niektórych fonemów może wystąpić w okolicy częstotliwości 200 Hz, a więc w pobliżu potencjalnych wartości częstotliwości tonu krtaniowego. W celu wyeliminowania ww. zjawisk stosuje się filtrację dolnoprzepustową sygnału, a także ustalenie wartości progowej, powyżej której dane maksimum lokalne korelacji uznaje się za związane z tonem krtaniowym. W tym celu stosuje się także modyfikacje powszechnie znanych algorytmów korelacyjnych. Jedną z nich jest zastosowanie ważonej autokorelacji, w której wagą jest odwrotność funkcji różnicowej AMDF (ang. Average Magnitude Difference Function), posiadającej minima lokalne dla dokładnie tych samych argumentów, dla których autokorelacja przyjmuje warto-

ści maksymalne. Taki zabieg powoduje zredukowanie błędów estymacji wielokrotności F_0 [49] [73].

Kolejnym powszechnie stosowanym, relatywnie prostym i wydajnym algorytmem, który obrano w tej pracy za podstawę do wyznaczania częstotliwości F_0, jest **algorytm YIN** [12]. Zaletą tego podejścia jest brak górnej granicy zakresu, w którym poszukuje się potencjalnych wartości F_0, co zapewnia dobrą jakość estymacji dla głosów wysokich, charakteryzujących się dużymi wartościami częstotliwości tonu krtaniowego. Wykorzystywana w pracy w eksperymentach numerycznych wersja algorytmu opiera się na znormalizowanej funkcji różnicowej $d'_t(\tau)$:

$$d'_t(\tau) = \begin{cases} 1 & \text{dla } \tau = 0, \\ \frac{d_t(\tau)}{1/\tau \sum_{j=1}^{\tau} d_t(j)} & \text{w przeciwnym przypadku,} \end{cases} \tag{A.1}$$

gdzie $d_t(\tau)$ jest funkcją różnicową:

$$d_t(\tau) = \sum_{j=1}^{W} (x_j - x_{j+\tau})^2, \tag{A.2}$$

a W jest długością okna analizy sygnału. Wybór takiego rozwiązania zapewnia normalizację funkcji $d_t(\tau)$ i jednocześnie skutecznie kompensuje część istotnych błędów estymacji szczególnie w sytuacji bliskości częstotliwości podstawowej F_0 oraz częstotliwości pierwszego formantu. Dalej stosuje się także progowanie oraz interpolację w celu redukcji błędów związanych z ograniczeniami wynikającymi z próbkowania sygnału. Zaletą tego algorytmu jest także jego odporność na przypadki estymacji podwojonej wartości częstotliwości podstawowej.

Ulepszoną statystycznie wersją tego algorytmu jest PYIN (Probabilistic YIN) [62], za pomocą którego, w odróżnieniu od pierwotnej wersji algorytmu, nie wyznacza się jednego estymatora częstotliwości F_0, lecz kilka potencjalnych wartości wraz z powiązanymi z nimi prawdopodobieństwami. Wynika to z faktu, że algorytm YIN zakłada stałą wartość parametru progowania, a PYIN wykorzystuje rozkład prawdopodobieństwa tych wartości. Prawdopodobieństwa te są dalej wykorzystywane w modelach HMM do obliczenia wiarygodnego estymatora odpornego na błędy.

Dodatek B

Opis pozostałych algorytmów wykorzystywanych w pracy

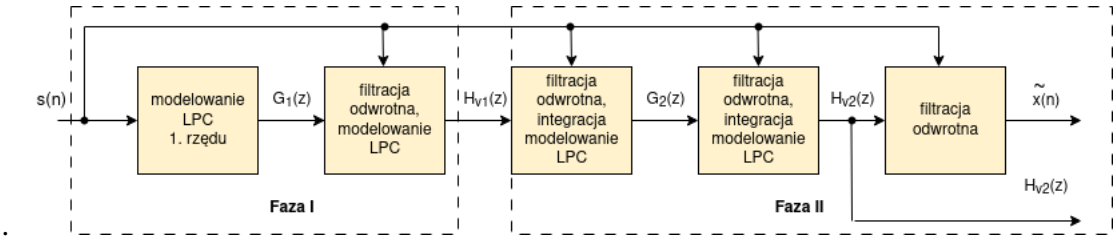
B.1. Detekcja dźwięcznych ramek sygnału mowy

Jednym z zagadnień w systemach ARM jest detekcja dźwięcznych fragmentów sygnału mowy. Najczęściej decyzję o dźwięczności ramki i wartości częstotliwości F_0 podejmuje się na podstawie unormowanej funkcji autokorelacji [57]. Podejście takie wiąże się z dużą złożonością obliczeniową, dlatego w celu redukcji tej złożoności, warto wcześniej podjąć decyzję o dźwięczności, a autokorelację liczyć tylko dla ramek dźwięcznych. W literaturze opisano wiele metod detekcji dźwięczności ramek sygnału mowy [57][83]. Wśród nich efektywne i popularne są metody bazujące na energii oraz liczbie przejść sygnału przez zero [4]. Nie zawsze są one jednak efektywne. W związku z tym, w niniejszej pracy zaproponowano nieco inne podejście, w którym energie w pasmach wyznacza się na podstawie obwiedni rozfiltrowanych sygnałów, a uśrednianie i normowanie gwarantują uniezależnienie się od poziomu sygnału. W zaproponowanej metodzie używa się filtrów quasi-kwadraturowych [59]. W kolejnym kroku uśrednia się obwiednie w czterech pasmach częstotliwościowych i dla każdej ramki sygnału wyznacza się energię w danym podpaśmie. Te cztery parametry stanowią postawę podejmowania decyzji o dźwięczności. Podejście to szczegółowo opisano w pracy [26].

B.2. Algorytm filtracji odwrotnej IAIF (ang. Iterative Adaptive Inverse Filtering)

W drugiej z zaproponowanych w niniejszej pracy metod odpornej parametryzacji wykorzystywany jest algorytm filtracji odwrotnej IAIF [86], za pomocą którego możliwa jest estymacja sygnału pobudzenia $\tilde{x}(n)$ oraz odpowiedzi impulsowej filtru traktu głosowego $\tilde{v}(n)$. W rozdziale

przedstawiono opis niniejszego algorytmu. Całą procedurę filtracji odwrotnej można podzielić na dwie fazy, co zaprezentowano na schemacie blokowym z rys. B.1.



Rys. B.1: Schemat algorytmu filtracji odwrotnej IAIF.

Sygnał wejściowy $s(n)$ poddawany jest najpierw filtracji górnoprzepustowej przy użyciu filtru FIR o częstotliwości granicznej f_c ok. 70 Hz. Zabieg ten ma na celu skasowanie niskoczęstotliwościowych szumów otoczenia zarejestrowanych przez mikrofon. Jest to standardowa procedura preprocessingu w algorytmach filtracji odwrotnej, które wymagają użycia wysokiej jakości mikrofonów o dolnej granicy pasma przenoszenia zaczynającej się już od ok. 5-20 Hz.

W pierwszej fazie zakłada się, że wpływ sygnału pobudzenia na widmo mowy można zamodelować procesem AR niskiego rzędu. Kasując ten wpływ z oryginalnego sygnału uzyskuje się parametryczny model traktu głosowego, który służy dalej do eliminacji wpływu formantów. W pierwszym kroku, za pomocą analizy LPC 1. rzędu obliczany jest zgrubny estymator filtru odwrotnego o transmitancji $G_1(z) = 1 + az^{-1}$, modelującego wpływ pobudzenia i właściwości emisyjnych ust mówcy. Udział ten jest kasowany z oryginalnego sygnału mowy $s(n)$ poprzez filtrację odwrotną. Sygnał wyjściowy takiego filtru poddawany jest dalej analizie LPC rzędu p w celu obliczenia zgrubnego estymatora filtru traktu głosowego o funkcji transmitancji $H_{vt1}(z) = 1 + \sum_{k=1}^{p} a(k)z^{-k}$.

W drugiej fazie wykorzystuje się modele AR wyższych rzędów, za pomocą których można uzyskać dokładniejsze oszacowanie sygnału pobudzenia. Wpływ transmitancji kanału głosowego jest eliminowany z widma oryginalnego sygnału poprzez filtrację odwrotną. Z racji, że komponent $r(n)$ z modelu Fanta można aproksymować układem różniczkującym, wpływ transmisji mowy przez usta mówcy eliminuje się z sygnału mowy za pomocą operacji całkowania. Sygnał wynikowy poddawany jest dalej analizie LPC rzędu g, dzięki czemu możliwe jest dokładniejsze oszacowanie wpływu pobudzenia. W tym kroku możliwa jest analiza LPC rzędu wyższego niż w pierwszej fazie algorytmu, ponieważ po skasowaniu z sygnału udziału kanału głosowego model parametryczny nie będzie zawierał fałszywych maksimów charakterystyki amplitudowej podobnych do formantów. Wpływ pobudzenia i właściwości emisyjnych jest eliminowany przy pomocy filtru odwrotnego o transmitancji $G_2(z)$ oraz operacji całkowania, a analiza LPC rzędu p służy do wyznaczenia wiarygodnego modelu $H_{v2}(z)$ kanału głosowego. Ostatnią operacją jest filtracja odwrotna, w wyniku której otrzymuje się estymator sygnału pobudzenia $\tilde{x}(n)$

Dodatek C

Błędy rozpoznawania par fonemów

W rozdziale zaprezentowano wartości $E_{f,g}$ czyli lokalny FER w rozpoznawaniu jeden-do-jeden dla wszystkich par badanych fonemów. Górna wartość w każdej komórce tabeli oznacza błąd rozpoznawania związany z metodą referencyjną HFCC, a dolna - z jedną z metod odpornych. Analogicznie do wcześniejszych rozważań, dla każdej z metod przygotowano 2 tabele: jedną dotyczącą samogłosek, drugą dotyczącą spółgłosek. Podobnie, kolorem zielonym oznaczono przypadki redukcji błędu, zaś kolorem czerwonym przypadki jego zwiększenia.

Tab. C.1: Wartości błędu $E_{f,g}$, określającego częstość pomyłek pomiędzy parami fonemów. Rozpoznawanie samogłosek w metodzie PS-HFCC.

fonem / fonem	i	y	e	a	o	u
i		1.45 1.46	0.43 0.50	0.00 0.00	0.00 0.13	1.53 0.73
y	3.46 3.53		21.60 18.75	0.00 0.00	0.19 0.00	0.14 0.14
e	1.30 0.55	8.28 7.84		5.77 4.58	0.42 0.44	0.12 0.05
a	0.00 0.00	0.06 0.00	6.52 5.34		4.01 3.81	0.01 0.00
o	0.00 0.02	0.26 0.09	1.47 0.76	5.82 4.44		1.54 1.20
u	0.87 1.44	0.34 0.36	0.50 0.03	0.12 0.00	6.91 4.62	
j	29.99 29.59	15.09 15.09	19.08 16.93	0.00 0.00	0.19 0.00	0.44 0.39
ł	0.66 1.30	0.08 0.00	0.50 0.35	0.66 0.00	14.37 11.65	26.59 26.78
r	0.60 1.05	7.78 7.47	16.79 14.94	5.09 4.09	10.02 6.79	4.13 3.00
l	2.79 3.86	9.34 6.89	3.58 3.12	1.40 0.55	4.80 3.76	3.49 2.66
m	1.61 3.74	0.22 0.51	1.13 0.28	0.00 0.06	1.51 0.85	5.65 4.70
n	3.23 4.82	0.74 0.69	0.82 0.39	0.74 0.13	0.94 0.56	2.78 2.62
n'	10.42 14.47	0.96 1.01	3.25 2.10	0.00 0.00	0.67 0.62	5.32 4.51
$n,$	2.67 3.96	0.91 0.48	3.89 3.67	1.27 0.67	6.30 6.63	4.48 5.53

Tab. C.2: Wartości błędu $E_{f,g}$, określającego częstość pomyłek pomiędzy parami fonemów. Rozpoznawanie spółgłosek w metodzie PS-HFCC.

fonem / fonem	j	ł	r	l	m	n	n'	$n,$
i	5.13 5.49	0.10 0.04	0.02 0.00	0.16 0.15	0.67 0.90	1.43 1.29	2.13 2.22	0.92 1.16
y	6.08 6.78	0.00 0.03	2.24 2.15	1.01 1.67	0.14 0.17	0.44 0.40	0.19 0.32	0.79 0.55
e	2.61 1.80	0.01 0.04	1.38 1.99	0.13 0.18	0.02 0.02	0.15 0.07	0.22 0.13	0.53 0.59
a	0.01 0.00	0.00 0.00	0.33 0.37	0.05 0.02	0.06 0.00	0.05 0.00	0.00 0.00	0.17 0.15
o	0.03 0.00	1.94 1.67	2.51 2.57	0.25 0.33	0.07 0.00	0.12 0.01	0.07 0.02	1.32 1.10
u	0.34 0.43	10.15 9.87	2.40 2.69	3.18 2.85	1.37 1.67	2.99 2.49	1.40 1.18	4.36 4.36
j		0.12 0.00	0.06 0.13	0.06 0.13	0.75 0.72	0.50 0.92	1.50 1.38	0.62 1.25
ł	0.00 0.00		0.66 0.43	0.08 0.09	0.25 0.35	0.25 0.43	0.58 0.78	2.97 2.87
r	0.40 0.63	0.04 0.00		6.41 5.78	1.08 1.52	1.96 2.07	0.28 0.34	1.12 1.43
l	0.35 1.01	0.09 0.00	16.49 18.55		2.09 1.56	3.84 4.59	1.13 1.10	1.57 1.29
m	0.27 0.40	0.48 0.34	1.13 1.36	0.91 1.13		20.22 16.58	6.99 6.85	9.30 9.45
n	0.08 0.26	0.16 0.13	1.72 2.19	1.55 2.15	9.44 10.63		6.99 6.76	12.31 10.76
n'	1.26 0.86	0.67 0.47	0.74 0.93	0.30 0.39	8.57 9.81	18.11 13.77		7.39 8.72
$n,$	0.54 0.38	0.45 1.48	1.36 1.24	0.50 0.62	8.47 7.72	20.43 17.49	7.20 7.01	

Tab. C.3: Wartości błędu $E_{f,g}$, określającego częstość pomyłek pomiędzy parami fonemów. Rozpoznawanie samogłosek w metodzie IF-HFCC.

fonem / fonem	i	y	e	a	o	u
i		1.45 1.41	0.43 0.13	0.00 0.00	0.00 0.00	1.53 0.99
y	3.46 3.71		21.60 18.02	0.00 0.00	0.19 0.00	0.14 0.12
e	1.30 0.53	8.28 9.72		5.77 4.43	0.42 0.40	0.12 0.12
a	0.00 0.00	0.06 0.01	6.52 5.89		4.01 3.26	0.01 0.00
o	0.00 0.00	0.26 0.08	1.47 0.97	5.82 4.96		1.54 1.82
u	0.87 0.69	0.34 0.20	0.50 0.26	0.12 0.10	6.91 7.92	
j	29.99 31.01	15.09 13.44	19.08 13.23	0.00 0.00	0.19 0.00	0.44 0.55
ł	0.66 0.09	0.08 0.00	0.50 0.44	0.66 0.17	14.37 12.95	26.59 24.67
r	0.60 0.27	7.78 6.83	16.79 14.30	5.09 4.00	10.02 7.42	4.13 4.77
l	2.79 2.26	9.34 8.19	3.58 4.24	1.40 0.56	4.80 2.54	3.49 4.52
m	1.61 1.68	0.22 0.24	1.13 0.90	0.00 0.12	1.51 1.14	5.65 5.05
n	3.23 3.58	0.74 1.16	0.82 0.74	0.74 0.05	0.94 0.79	2.78 2.93
n'	10.42 9.33	0.96 1.44	3.25 2.29	0.00 0.00	0.67 0.59	5.32 3.90
$n,$	2.67 2.29	0.91 0.73	3.89 3.90	1.27 1.09	6.30 9.98	4.48 4.58

Tab. C.4: Wartości błędu $E_{f,g}$, określającego częstość pomyłek pomiędzy parami fonemów. Rozpoznawanie spółgłosek w metodzie IF-HFCC.

fonem / fonem	j	ł	r	l	m	n	n'	$n,$
i	5.13 6.90	0.10 0.09	0.02 0.00	0.16 0.16	0.67 0.74	1.43 0.69	2.13 1.52	0.92 0.78
y	6.08 8.19	0.00 0.00	2.24 2.60	1.01 0.78	0.14 0.21	0.44 0.30	0.19 0.63	0.79 0.21
e	2.61 2.80	0.01 0.01	1.38 2.04	0.13 0.10	0.02 0.14	0.15 0.11	0.22 0.52	0.53 0.52
a	0.01 0.00	0.00 0.00	0.33 0.50	0.05 0.02	0.06 0.02	0.05 0.02	0.00 0.00	0.17 0.28
o	0.03 0.01	1.94 1.99	2.51 2.25	0.25 0.52	0.07 0.09	0.12 0.19	0.07 0.03	1.32 1.24
u	0.34 0.13	10.15 10.55	2.40 2.10	3.18 3.15	1.37 2.37	2.99 2.43	1.40 1.64	4.36 3.71
j		0.12 0.00	0.06 0.07	0.06 0.07	0.75 0.76	0.50 0.76	1.50 1.45	0.62 0.76
ł	0.00 0.00		0.66 0.35	0.08 0.00	0.25 0.17	0.25 0.61	0.58 0.79	2.97 2.36
r	0.40 0.63	0.04 0.04		6.41 7.96	1.08 1.62	1.96 2.70	0.28 0.58	1.12 1.53
l	0.35 0.75	0.09 0.00	16.49 20.53		2.09 1.04	3.84 5.93	1.13 1.04	1.57 0.66
m	0.27 0.24	0.48 0.18	1.13 2.17	0.91 1.74		20.22 18.35	6.99 6.98	9.30 10.11
n	0.08 0.05	0.16 0.19	1.72 2.60	1.55 2.65	9.44 11.38		6.99 5.43	12.31 13.28
n'	1.26 1.53	0.67 0.93	0.74 1.61	0.30 0.85	8.57 9.75	18.11 13.66		7.39 9.84
$n,$	0.54 0.62	0.45 1.72	1.36 0.94	0.50 0.73	8.47 8.48	20.43 18.15	7.20 5.88	