Magdalena Kaniewska

Zespolona pulsacja chwilowa w analizie i konwersji głosu

Rozprawa doktorska

Promotor:
prof. dr hab. inż. Ewa Hermanowicz
Wydział Elektroniki, Telekomunikacji
i Informatyki
Politechnika Gdańska

Gdańsk, 2011

Podziękowania

Pragnę złożyć serdeczne podziękowania pani profesor Ewie Hermanowicz za nieocenioną pomoc na każdym etapie pracy, doktorowi Mirosławowi Rojewskiemu za inspirację i konstruktywną krytykę, a także kierownikowi Katedry Systemów Multimedialnych, profesorowi Andrzejowi Czyżewskiemu za stworzenie możliwości pracy i rozwoju oraz kolegom za chętnie udzielaną pomoc.

Wykaz ważniejszych skrótów

(w porządku alfabetycznym)

AM	–	modulacja amplitudy (od ang. *Amplitude Modulation*)
Ave	–	uśrednianie, filtracja uśredniająca (ang. *Averaging*)
CC	–	współrzędne kartezjańskie (ang. *Cartesian Coordinates*)
DC	–	składowa stała (ang. *Direct Current Component*)
DFT	–	dyskretna transformacja Fouriera (ang. *Discrete Fourier Transformation*)
EMD	–	od ang. *Empirical Mode Decomposition*
FIR	–	skończona odpowiedź impulsowa (ang. *Finite Impulse Response*)
FM	–	modulacja częstotliwości (od ang. *Frequency Modulation*)
GE	–	błąd gruby (ang. *Gross Error*)
HNM	–	model harmoniczne+szum (od ang. *Harmonic+Noise Model*)
HT	–	transformata Hilberta (od ang. *Hilbert Transform*)
HVD	–	od ang. *Hilbert Vibration Decomposition*
IA	–	amplituda chwilowa (ang. *Instantaneous Frequency*)
IB	–	chwilowa szerokość pasma (od ang. *Instantaneous Bandwidth*)
ICF	–	zespolona pulsacja chwilowa (od ang. *Instantaneous Complex Frequency*)
ICFE	–	estymator zespolonej pulsacji chwilowej (od ang. *Instantaneous Complex Frequency Estimator*)
IF	–	częstotliwość chwilowa (od ang. *Instantaneous Frequency*)
IFE	–	estymator pulsacji chwilowej (ang. *Instantaneous Frequency Estimator*)
IIR	–	nieskończona odpowiedź impulsowa (ang. *Infinite Impulse Response*)
IPA	–	Międzynarodowy Alfabet Fonetyczny (ang. *International Phonetic Alphabet*)
LP	–	predykcja liniowa (ang. *Linear Prediction*)

LPF	–	filtr dolnoprzepustowy (ang. *Low-Pass Filter*)
LPSD	–	predykcja liniowa w dziedzinie częstotliwości (ang. *Linear Prediction in Spectral Domain*)
MF	–	filtr medianowy (ang. *Median Filter*)
MP	–	minimalnofazowy (ang. *Minimum-Phase*)
MPE	–	obwiednia minimalnofazowa (ang. *Minimum-Phase Envelope*)
MPEE	–	estymator obwiedni minimalnofazowej (ang. *Minimum-Phase Envelope Estimator*)
PC	–	współrzędne biegunowe (ang. *Polar Coordinates*)
PIF	–	dodatnia pulsacja chwilowa (ang. *Positive Instantaneous Frequency*)
PIFP	–	fazor dodatnioskrętny, o zawsze dodatniej pulsacji chwilowej (ang. *Positive Instantaneous Frequency Phasor*)
PIFPE	–	estymator fazora dodatnioskrętnego (ang. *Positive Instantaneous Frequency Estimator*)
RMS	–	wartość skuteczna (ang. *Root Mean Square*)
Sa	–	próbki (ang. *Samples*)
V-KB	–	bifaktoryzacja Voelckera-Kumaresana

Wykaz ważniejszych oznaczeń

Notacja matematyczna:

$\in$	–	przynależność do zbioru
*	–	„w wykładniku" – zespolone sprzężenie, np. x^*
		„w podstawie" – operator splotu liniowego, np. $x * y$
'	–	„w wykładniku" – pochodna funkcji, np. $f'(x)$
$\lvert\ \rvert$	–	wartość bezwzględna funkcji lub liczby
Re( )	–	część rzeczywista liczby zespolonej
Im( )	–	część urojona liczby zespolonej
arg( )	–	argument liczby zespolonej
Arg( )	–	argument główny liczby zespolonej
ln( )	–	zespolony logarytm (analityczny, Cauchy'ego)
Ln( )	–	zespolony logarytm główny liczby zespolonej
exp( )	–	eksponenta
sgn( )	–	signum (znak) liczby rzeczywistej
$H_T\{\ \}$	–	transformata Hilberta
j	–	jednostka urojona
R	–	zbiór liczb rzeczywistych
$\mathbf{R}_+$	–	zbiór liczb rzeczywistych dodatnich
C	–	zbiór liczb zespolonych

Sygnały, ich transformaty i parametry:

F	–	częstotliwość [Hz]
Ω	–	pulsacja [rad/s], $\Omega = 2\pi F$
t	–	zmienna czasu [s]
$\delta(t)$	–	delta Diraca

$x(t)$ – sygnał rzeczywisty z czasem ciągłym

$y(t)$ – sygnał rzeczywisty, będący transformatą Hilberta sygnału $x(t)$

$X(\Omega)$ – transformata Fouriera sygnału $x(t)$

$X(s)$ – transformata Laplace'a sygnału $x(t)$

$u(t)$ – sygnał zespolony z czasem ciągłym, równoważnik hilbertowski sygnału rzeczywistego $x(t)$

$a(t)$ – amplituda chwilowa (obwiednia)

$\lambda(t)$ – logarytm amplitudy chwilowej (logobwiednia)

$\gamma(t)$ – fazor FM

$\varphi(t)$ – faza chwilowa

$f(t)$ – częstotliwość chwilowa

$\omega(t)$ – pulsacja chwilowa $\omega(t) = 2\pi f(t)$

$\sigma(t)$ – względna chwilowa prędkość promieniowa

$p(t)$ – zespolona faza chwilowa

$s(t)$ – zespolona pulsacja chwilowa

$a_{mp}(t)$ – obwiednia minimalnofazowa sygnału analitycznego

$\gamma_{pif}(t)$ – fazor dodatnioskrętny

$\varphi_{mp}(t)$ – faza chwilowa obwiedni minimalnofazowej

$\varphi_{pif}(t)$ – faza chwilowa fazora dodatnioskrętnego

$\omega_{mp}(t)$ – pulsacja chwilowa obwiedni minimalnofazowej

$\omega_{pif}(t)$ – pulsacja chwilowa fazora dodatnioskrętnego

$s_{mp}(t)$ – zespolona pulsacja chwilowa obwiedni minimalnofazowej

$s_{pif}(t)$ – zespolona pulsacja chwilowa fazora dodatnioskrętnego

T_S – okres próbkowania [s]

F_S – szybkość próbkowania [Sa/s], $F_S = 1/T_S$

n – numer próbki sygnału dyskretnego ($n = 0,1,..., N-1$)

f – częstotliwość znormalizowana względem częstotliwości próbkowania [1/Sa], $f = F/F_S$

ω – pulsacja znormalizowana względem częstotliwości próbkowania [rad/Sa], $\omega = 2\pi F/F_S = 2\pi f$

$x[n]$ – sygnał dyskretnoczasowy (dyskretny)

$X(\omega)$ – dyskretnoczasowa transformata Fouriera sygnału $x[n]$

$X(z)$ – transformata Z sygnału $x[n]$

$a[n]$ – amplituda chwilowa (obwiednia)

$\lambda[n]$ – logarytm amplitudy chwilowej (logobwiednia)

$\varphi[n]$ – faza chwilowa [rad]

$\omega[n]$ – pulsacja chwilowa [rad/Sa]

$s[n]$ – zespolona pulsacja chwilowa

$\sigma[n]$ – składowa promieniowa zespolonej pulsacji chwilowej [Np/Sa]

$a_{mp}[n]$ – obwiednia minimalnofazowa sygnału analitycznego

$\gamma_{pif}[n]$ – fazor dodatnioskrętny

$\varphi_{mp}[n]$ – faza chwilowa obwiedni minimalnofazowej

$\varphi_{pif}[n]$ – faza chwilowa fazora dodatnioskrętnego

$\omega_{mp}[n]$ – pulsacja chwilowa obwiedni minimalnofazowej

$\omega_{pif}[n]$ – pulsacja chwilowa fazora dodatnioskrętnego

$s_{mp}[n]$ – zespolona pulsacja chwilowa obwiedni minimalnofazowej

$s_{pif}[n]$ – zespolona pulsacja chwilowa fazora dodatnioskrętnego

F_0 – częstotliwość podstawowa sygnału mowy [Hz]

F1, *F2*, F3, … – formanty mowy

$F_1, F_2, F_3, \ldots$ – częstotliwości rezonansowe formantów mowy

$B_1, B_2, B_3, \ldots$ – szerokości pasm formantów mowy

Charakterystyki systemów (czasowe, częstotliwościowe i operatorowe):

$h(t)$ – odpowiedź impulsowa systemu analogowego

$H(\Omega)$ – charakterystyka częstotliwościowa systemu analogowego

$H(s)$ – transmitancja operatorowa systemu analogowego

$H^{-1}(s)$ – transmitancja systemu odwrotnego do systemu o transmitancji $H(s)$ ($H^{-1}(s)H(s)=1$)

$h[n]$ – odpowiedź impulsowa systemu dyskretnego

$H(\omega)$ – charakterystyka częstotliwościowa systemu dyskretnego

$H(z)$ – transmitancja systemu dyskretnego

$h_T(t)$ – odpowiedź impulsowa idealnego transformatora Hilberta

$H_T(\Omega)$ charakterystyka częstotliwościowa idealnego transformatora Hilberta

$h_T[n]$ – odpowiedź impulsowa przyczynowego filtru FIR, aproksymującego idealny transformator Hilberta

$h_A(t)$ odpowiedź impulsowa idealnego zespolonego filtru Hilberta

$H_A(\Omega)$ – charakterystyka częstotliwościowa idealnego zespolonego filtru Hilberta

$h_{A,F_c,B}[n]$ – odpowiedź impulsowa zespolonego filtru Hilberta o częstotliwości środkowej F_c i szerokości pasma B

$H_{A,F_c,B}(\omega)$ – charakterystyka częstotliwościowa zespolonego filtru Hilberta o częstotliwości środkowej F_c i szerokości pasma B

Spis treści

1. WPROWADZENIE 4

1.1. CELE I ZAKRES PRACY 7

2. GŁOS I MOWA 9

2.1. MECHANIZM GENEROWANIA GŁOSU 9

2.2. PODSTAWOWE POJĘCIA Z ZAKRESU FONETYKI I FONOLOGII 11

2.3. WŁAŚCIWOŚCI SYGNAŁU MOWY 13

2.3.1. Ton krtaniowy 16

2.3.2. Trakt głosowy 17

2.4. PRZEGLĄD WYBRANYCH MODELI MOWY 19

3. KONCEPCJA ZESPOLONEJ PULSACJI CHWILOWEJ 24

3.1. DEFINICJA CZĘSTOTLIWOŚCI CHWILOWEJ 24

3.2. RÓWNOWAŻNIK ANALITYCZNY SYGNAŁU RZECZYWISTEGO 26

3.2.1. Transformacja Hilberta 27

3.2.2. Sygnał analityczny Gabora 30

3.2.3. Reprezentacja sygnału analitycznego przez wskaz 31

3.2.4. Reprezentacja AM·FM 32

3.3. IF SYGNAŁÓW WIELOKOMPONENTOWYCH 35

3.3.1. Wybrane metody dekompozycji sygnałów wielokomponentowych 37

3.4. INNE DEFINICJE IF 39

3.5. DEFINICJA ZESPOLONEJ PULSACJI CHWILOWEJ 43

3.5.1. Interpretacja zespolonej pulsacji chwilowej 45

4. BIFAKTORYZACJA VOELCKERA-KUMARESANA 47

4.1. MINIMALNOFAZOWOŚĆ, MAKSYMALNOFAZOWOŚĆ I MIESZANOFAZOWOŚĆ 47

4.2. FILTR PRZYCZYNOWY JAKO KASKADA FILTRÓW MINIMALNOFAZOWEGO I WSZECHPRZEPUSTOWEGO 51

4.3. BIFAKTORYZACJA V-K SYGNAŁU ANALITYCZNEGO 51

4.4. PORÓWNANIE FAKTORYZACJI V-K I FAKTORYZACJI AM·FM 54
4.5. DYSKRETNA IMPLEMENTACJA ANALIZATORA AM· PIF 55
4.6. TESTOWANIE ALGORYTMU ANALIZATORA 59

5. ZESPOLONA PULSACJA CHWILOWA W ANALIZIE GŁOSU 74

5.1. WŁAŚCIWOŚCI CZYNNIKÓW V-KB ORAZ ICH PARAMETRÓW CHWILOWYCH NA PRZYKŁADZIE POLSKICH GŁOSEK 74
5.1.1. Miara minimalnofazowości głosek 74
5.1.2. Analiza polskich głosek 79
5.1.3. IF a częstotliwość podstawowa 85
5.2. ESTYMACJA CZĘSTOTLIWOŚCI PODSTAWOWEJ 91
5.2.1. Klasyfikacja mowy na dźwięczną i bezdźwięczną 93
5.2.2. Estymacja prawdopodobnych częstotliwości podstawowych 95
5.2.3. Wybór poprawnej estymaty częstotliwości podstawowej 99
5.2.4. Eksperymenty 101
5.2.4.1. Ocena poprawności klasyfikacji mowy na dźwięczną i bezdźwięczną 102
5.2.4.2. Ocena wyników estymacji częstotliwości podstawowej 105
5.3. EKSTRAKCJA FORMANTÓW MOWY 108
5.3.1. Metoda Feldmana dekompozycji sygnałów wielokomponentowych 109
5.3.2. Adaptacja metody HVD dla analizy mowy 110
5.3.3. Eksperymenty 115
5.3.3.1. Ocena poprawności estymacji częstotliwości środkowych formantów 117

6. KONWERSJA GŁOSU W OPARCIU O CZYNNIKI V-KB I ICH PARAMETRY CHWILOWE 121

6.1. GŁOŚNOŚĆ, WYSOKOŚĆ I BARWA GŁOSU 121
6.2. GŁOS – CECHY DYSTYNKTYWNE MÓWCY 124
6.3. MOŻLIWOŚCI MODYFIKACJI GŁOSU ZA POMOCĄ ICF 125
6.3.1. Proponowane modyfikacje ICF 127
6.3.1.1. Synteza sygnału mowy po modyfikacjach 130
6.3.2. Modyfikacje sygnału mowy 131
6.3.2.1. Testy odsłuchowe 136

6.3.3. Modyfikacje ICF poszczególnych formantów 138
6.3.3.1. Testy odsłuchowe 139
7. PODSUMOWANIE 144
BIBLIOGRAFIA 148
ZAŁĄCZNIK A – ZAWARTOŚĆ PŁYTY CD 160

1. Wprowadzenie

Komunikacja werbalna jest podstawowym i najdoskonalszym sposobem porozumiewania się ludzi, pozwalającym na wyrażanie myśli, poglądów i uczuć, wymianę informacji i doświadczeń, dając tym samym podstawy do rozwoju cywilizacji, techniki i kultury. Komunikacja werbalna polega na odpowiednim użyciu dźwięków i języka w celu przekazania treści. Wiążą się z tym trzy główne pojęcia, które, choć używane na co dzień, często są ze sobą mylone. Tymi pojęciami są: głos, mowa i język.

Głos jest to dźwięk generowany przez człowieka za pomocą aparatu mowy, do którego należą: płuca, przepona i mięśnie brzucha, tchawica, krtań oraz jamy: gardłowa, ustna i nosowa. Głos uczestniczy w wytwarzaniu mowy głośnej (nie szeptu), ale nie zawsze się z nim wiąże. Dla przykładu, niemowlęta generują głos, choć nie mają jeszcze zdolności generowania mowy. O wytwarzaniu głosu możemy też mówić w przypadku zwierząt. W niniejszej pracy skupimy się jednak na głosie wyłącznie w kontekście mowy.

Mowa jest narzędziem do przekazania komunikatu. Jej generowanie rozpoczyna się od sformułowania tego komunikatu w umyśle mówcy. Następnym etapem jest wytworzenie głosu oraz odpowiednia artykulacja tak, by generowane dźwięki były zgodne z pewną przyjętą konwencją, a więc zrozumiałe dla słuchacza. W generowaniu mowy ważna jest także prozodia, czyli brzmieniowe właściwości mowy nakładające się na głoskowy, sylabiczny i wyrazowy ciąg wypowiedzi. Do właściwości tych należą: akcenty, intonacja i iloczas (sposób różnicowania głosek i sylab ze względu na długość ich trwania, który może służyć różnicowaniu znaczeń wyrazów – ta funkcja iloczasu zanikła w języku polskim, a także może stanowić podstawę rytmizacji). Prozodia pozwala rozróżniać funkcje i ważność poszczególnych wyrazów czy zdań w wypowiedzi, a także pomaga w przekazywaniu emocji.

Mowa oznacza również używanie języka w procesie porozumiewania się. Język jest ukształtowanym społecznie systemem budowania wypowiedzi, składającym się ze znaków oraz reguł, według których tworzymy i łączymy te znaki. Porozumiewanie się za pomocą mowy wymaga, by mówiący i słuchający używali tego samego języka.

Mowę można analizować na kilku poziomach: semantycznym, czyli dotyczącym treści wypowiedzi, osobniczym, który pozwala zidentyfikować osobę mówiącą, prozodycznym czy emocjonalnym. Analizować można również sam głos, czyli dźwięk, pomijając pozostałe aspekty mowy, takie jak treść czy prozodia. Tak jak wszystkie inne dźwięki, głos

charakteryzowany jest przez trzy atrybuty, odnoszące się do sposobu percepcji: głośność, wysokość i barwę.

Jeśli do analizy mowy chcemy zaprząc narzędzia cyfrowego przetwarzania sygnałów (CPS), musimy mieć jej reprezentację sygnałową. Tym właśnie jest sygnał mowy, zarejestrowany przez mikrofon, który zamienia energię fali akustycznej na energię elektryczną, a otrzymany sygnał poddawany jest procesowi próbkowania i kwantyzacji. Obecnie za pomocą narzędzi CPS możemy, wykorzystując sygnał mowy, analizować właściwie wszystkie aspekty mowy: od głosu po treść wypowiedzi. Analiza głosu polega przede wszystkim na badaniu jego właściwości widmowych, które mają największy wpływ na jego brzmienie. Poprzez analizę głosu możemy uzyskać wiele informacji o jego właścicielu, od weryfikacji jego tożsamości po ocenę jego stanu emocjonalnego.

Wykorzystując CPS można również modyfikować sygnał mowy tak, by uzyskać inne brzmienie głosu, zachowując treść wypowiedzi, prozodię i emocje bez zmian. Takie przetwarzanie nazywać będziemy konwersją lub transformacją głosu. Wyniki konwersji głosu mogą być różne, od delikatnej zmiany jego barwy po zmiany brzmienia w stopniu, który uniemożliwia rozpoznanie mówcy.

Jak podkreśla Roark [RO06], jednym z najlepiej zakorzenionych pojęć w dziedzinie badania głosu jest częstotliwość. Ekstrakcja „parametrów częstotliwościowych" stała się właściwie synonimem analizy głosu. Jednak, w przeciwieństwie do parametrów takich jak amplituda lub energia, pomiar czy estymacja „parametrów częstotliwościowych" nie jest zadaniem łatwym i jednoznacznie zdefiniowanym. Po pierwsze, pojęcie częstotliwości można odnieść do różnych parametrów sygnału mowy. Po drugie, istnieje bardzo wiele diametralnie różnych, a jednak nie dyskwalifikujących się nawzajem, metod opisu głosu w kategoriach częstotliwości i wciąż powstają nowe. Oznacza to, że problem ten, choć od lat podejmowany przez wielu naukowców, pozostaje otwarty, i że wciąż jest na tym polu miejsce dla nowych badań.

Podstawowym modelem stosowanym w przetwarzaniu sygnału mowy jest model „źródło–filtr", który opisuje proces generowania mowy jako liniową filtrację pobudzenia za pomocą filtru o zmieniającej się w czasie charakterystyce. Zaletą tego modelu jest możliwość analizowania oddzielnie charakterystyki pobudzenia i filtru. Głównym ograniczeniem klasycznych metod wykorzystujących ten model jest przetwarzanie sygnału mowy w ramkach.

Ramka określa długość odcinka czasu, w którym sygnał jest analizowany. Otrzymany w ten sposób wynik analizy jest zawsze uśrednioną wartością mierzonego parametru. Klasyczne metody nie oddają więc dynamicznych zmian parametrów częstotliwościowych sygnału, jakie zachodzą w obrębie ramki. Tradycyjnie przyjmuje się, że sygnał mowy jest quasi-stacjonarny, a więc zmiany te w obrębie ramki są na tyle nieznaczne, że można je pominąć. Nasuwa się jednak pytanie, jaka jest granica niestacjonarności sygnału, do której użycie klasycznych metod analizy pozostaje uzasadnione i jak interpretować wynik takiej analizy, gdy granica ta zostanie przekroczona.

W świetle tych rozważań uzasadnionym staje się zastosowanie do opisu głosu częstotliwości chwilowej (IF od ang. *Instantaneous Frequency*), która estymuje częstotliwość sygnału w każdej chwili czasu jako pochodną jego fazy chwilowej. Główną zaletą takiego podejścia w stosunku do metod klasycznych jest możliwość lepszego odzwierciedlenia dynamicznych zmian fazy i częstotliwości sygnału. Przy tym IF równie dobrze nadaje się do opisu sygnałów stacjonarnych i okresowych, jak i niestacjonarnych i aperiodycznych. IF znalazła zastosowanie również w analizie i przetwarzaniu sygnału mowy [BO04] [HA94] [KU03b] [RA00] [RE07]. Jednak, jak podkreśla Roark [RO06], wiele możliwości IF pozostaje na tym polu wciąż nieodkrytych i niewykorzystanych, w porównaniu z innymi dziedzinami nauki, w których stosuje się ją z dużym powodzeniem.

W niniejszej pracy proponujemy zastosowanie do analizy i konwersji głosu pokrewnego do IF parametru chwilowego, zespolonej pulsacji chwilowej (ICF od ang. *Instantaneous Complex Frequency*). ICF, oprócz informacji o częstotliwości chwilowej, niesie również informację o chwilowej szerokości pasma sygnału (IB od ang. *Instantaneous Bandwidth*), która do tej pory nie była wykorzystywana do opisu głosu. Co więcej, ICF stanowi pełną reprezentację analizowanego sygnału, tzn. dysponując jej przebiegiem można bezinercyjnie odtworzyć reprezentowany przez nią sygnał. Zaproponowane w rozprawie podejście wykorzystuje ponadto faktoryzację sygnału na obwiednię minimalnofazową i fazor dodatnioskrętny, nazywaną dalej bifaktoryzacją Voelckera-Kumaresana (V-KB). Jest to reprezentacja sygnału analitycznego alternatywna dla szeroko stosowanej reprezentacji AM·FM, która faktoryzuje sygnał na obwiednię rzeczywistą AM i fazor FM. Wymienione tu narzędzia pozwalają na potokowe przetwarzanie głosu (próbka po próbce) i opracowanie algorytmów działających on-line.

1.1. Cele i zakres pracy

Zasadniczym celem rozprawy jest analiza głosu w kategoriach sygnałowych, widziana przez pryzmat ogólnej teorii Voelckera-Kumaresana zespolonej modulacji sygnałów analitycznych oraz znalezienie związków pomiędzy tym nowym opisem, a parametrami klasycznego modelu „źródło-filtr" i widmem fourierowskim. Wyniki tej analizy są podstawą dla osiągnięcia drugiego celu, jakim jest konwersja głosu za pomocą modyfikacji ICF czynników bifaktoryzacji V-K. Założeniem dla opracowanej metody konwersji było uzyskanie jak najbardziej naturalnego głosu i satysfakcjonującej jakości dźwięku. Osiągnięcie tych celów pozwoli potwierdzić następującą tezę:

Teza: Zespolona pulsacja chwilowa jako reprezentacja sygnału mowy daje nowe, dotychczas nieznane możliwości jego analizy, a proste modyfikacje zespolonej pulsacji chwilowej czynników bifaktoryzacji Voelckera-Kumaresana sygnału mowy pozwalają na konwersję głosu mówcy.

Rozdz. 2 rozprawy poświęcono omówieniu zagadnień z zakresu generowania i właściwości głosu, który jest przedmiotem opisywanych tu badań. Przedstawiono również krótko znane z literatury wybrane modele i metody analizy głosu, do których odnosimy się w dalszych częściach pracy. Ponadto przytoczono i wyjaśniono używane w rozprawie podstawowe pojęcia z zakresu fonetyki i fonologii.

W rozdz. 3 przytoczono zaczerpnięte z literatury definicje, interpretacje i sposoby estymacji IF oraz ICF. Przedstawiono również podstawy teoretyczne, stojące za pojęciami IF i ICF, dotyczące transformacji Hilberta, sygnału analitycznego Gabora oraz reprezentacji AM·FM. Oddzielny podrozdział poświęcono problemowi częstotliwości chwilowej sygnałów wielokomponentowych, do których zalicza się sygnał mowy.

W rozdz. 4 omówiono bifaktoryzację V-K oraz przedstawiono sposób jej cyfrowej implementacji. Przypomniano również krótko, co oznaczają stosowane w tej części pracy pojęcia minimalnofazowości, maksymalnofazowości oraz mieszanofazowości w odniesieniu do systemów i sygnałów. Ostatni podrozdział stanowi opis wykonanych w MATLABie symulacji, które pozwalają przyjrzeć się właściwościom czynników bifaktoryzacji V-K oraz ich ICF.

Rozdz. 5 opisuje możliwości zastosowania bifaktoryzacji V-K oraz ICF w analizie głosu. Pokazano w nim nowy sposób opisu głosu za pomocą ICF obwiedni minimalnofazowej i fazora dodatnioskrętnego sygnału mowy. Omówiono także zaproponowane w pracy algorytmy estymacji częstotliwości podstawowej i ekstrakcji formantów głosu wraz z przeprowadzonymi eksperymentami.

W rozdz. 6 zaprezentowano możliwości konwersji głosu za pomocą modyfikacji ICF czynników bifaktoryzacji V-K. W zaproponowanej metodzie wykorzystano opisane w rozdz. 5 algorytmy estymacji częstotliwości podstawowej i ekstrakcji formantów. Uzyskane efekty brzmieniowe zostały poddane ocenie w testach odsłuchowych, których wyniki przedyskutowano.

Rozdz. 7 stanowi podsumowanie rozprawy i ocenę opracowanych metod pod względem skuteczności, zakresu ich stososowalności i przydatności w praktycznych aplikacjach.

2. Głos i mowa

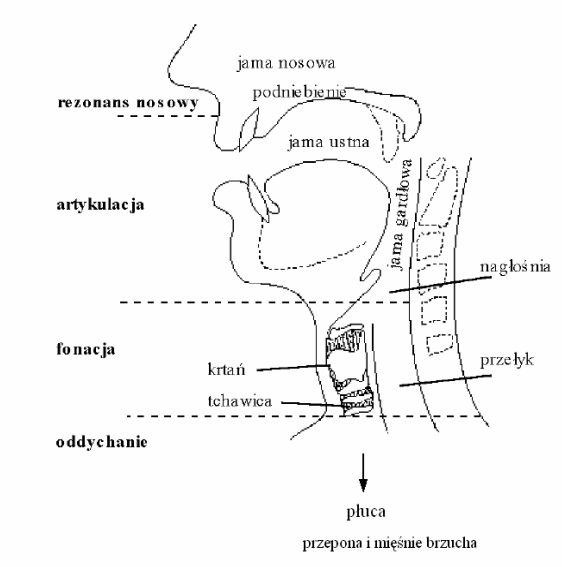
Celem niniejszego rozdziału jest przybliżenie podstawowych pojęć, związanych z głosem, mową i sygnałem mowy, których będziemy używać dalej w pracy. Przede wszystkim omówimy zagadnienia, dotyczące mechanizmu generowania głosu i mowy przez człowieka, jak również właściwości sygnału mowy w dziedzinie czasu i częstotliwości. Krótko przedstawione zostaną także podstawowe pojęcia z zakresu fonetyki i fonologii oraz wybrane modele stosowane w przetwarzaniu mowy.

2.1. Mechanizm generowania głosu

Głos jest falą akustyczną, generowaną przez ludzkie narządy mowy. Proces emisji głosu jest niezwykle złożony. Zaangażowanych jest w niego wiele narządów i mięśni. Przekrój przez najważniejsze narządy, biorące w nim udział oraz ich funkcje przedstawia rys. 2.1.

Jak wynika z rys. 2.1 emisję głosu można podzielić na trzy główne etapy: oddychanie, w którym udział biorą płuca, przepona i mięśnie brzucha, fonację, odbywającą się w krtani oraz artykulację, za którą odpowiadają gardło, jama nosowa i ustna wraz z wargami, zębami i językiem. Całość koordynowana jest przez ośrodkowy układ nerwowy. Dla inżyniera zajmującego się przetwarzaniem sygnału mowy najważniejsze jest zrozumienie etapów fonacji i artykulacji. W tym celu warto przedstawić proces mówienia jako operację filtracji akustycznej, w której pobudzeniem jest strumień powietrza wydobywający się z dolnych narządów głosowych, natomiast filtr stanowi trakt głosowy, zaczynający się na wyjściu krtani i kończący na ustach (odgałęzieniem traktu głosowego jest jama nosowa).

Fonacją nazywa się proces generowania przez krtań pobudzenia quasi-okresowego zwanego tonem krtaniowym (choć ciąg impulsów krtaniowych nie jest w rzeczywistości tonem, nazwa ta przyjęła się i jest ogólnie stosowana). Najważniejsze w tym procesie są fałdy głosowe, zawierające mięśnie i więzadła głosowe, których drgania prowadzą do powstania tonu krtaniowego. Przechodzący przez krtań słup powietrza przecinany jest przez drgające więzadła głosowe, których ruch powoduje rytmiczne zamykanie i otwieranie szpary głośni, znajdującej się pomiędzy fałdami głosowymi, przepuszczając lub zatrzymując wydychane powietrze. Powoduje to cykliczne narastanie i opadanie ciśnienia powietrza.



Rys. 2.1. Narządy mowy i ich funkcje (opisane na rys. pogrubioną czcionką).

Tak powstały ton krtaniowy stanowi pobudzenie dla głosek dźwięcznych. W mowie bezdźwięcznej pobudzeniem jest przepływający strumień powietrza o charakterze szumowym, a ton krtaniowy nie występuje. W tym przypadku więzadła głosowe pozostają w spoczynku, natomiast trakt głosowy jest w różnych miejscach zaciśnięty, co decyduje o rodzaju wymawianej głoski bezdźwięcznej. W przypadku generowania niektórych głosek pobudzenia dźwięczne i bezdźwięczne występują jednocześnie. Przykładem jest głoska /z/ w słowie „zero” lub /w/ w słowie „widmo”.

Powstałe pobudzenie poddawane jest filtracji, w wyniku której kształtowane są pożądane dźwięki mowy. Proces ten nazywany jest artykulacją i polega na odpowiednich zmianach kształtu traktu głosowego. Narządy biorące udział w artykulacji nazywa się artykulatorami. Funkcję artykulacyjną pełni przede wszystkim jama ustna, a w szczególności położenie języka względem podniebienia, ułożenie warg, żuchwy i zębów. Jama nosowa bierze udział w artykulacji głosek nosowych.

Artykulacja głosek związana jest z występowaniem w sygnale mowy tzw. formantów, przede wszystkim formantów niższych. Formant, którego formalną definicję podamy dalej, powstaje w wyniku wystąpienia rezonansu na danej częstotliwości. Rezonatorami w trakcie głosowym są: rezonatory dolne, czyli jama podgłośniowa, tchawica, oskrzela i klatka piersiowa, o małej możliwości przestrajania kształtu i rezonatory górne, czyli kolejno krtań, wpływająca na barwę dźwięku, jama gardłowa oraz ustna i nosowa, które, jak już wspomniano

są powodem występowania formantów niższych. Zagadnienie to zostanie omówione szerzej w dalszej części rozdziału.

2.2. Podstawowe pojęcia z zakresu fonetyki i fonologii

Ponieważ w dalszej części pracy często używane będą pojęcia z zakresu fonetyki i fonologii, w niniejszym podrozdziale zostaną one krótko opisane. Fonetyka i fonologia to nauki, zajmujące się warstwą brzmieniową języka. Fonetyka bada sposób powstawania dźwięków mowy i relacje zachodzące pomiędzy nimi, natomiast fonolodzy zajmują się badaniem dźwięków językowych pod względem ich funkcji dla znaczenia wyrazu.

Podstawowym pojęciem z zakresu fonetyki jest głoska – najmniejszy, niepodzielny element mowy, który daje się wyodrębnić za pomocą słuchu. Podstawowym podziałem głosek jest rozróżnienie na samogłoski i spółgłoski. Samogłoski charakteryzują się tym, że przy ich wymawianiu trakt głosowy nie jest w żadnym miejscu zaciśnięty. Z tego względu nazywa się je głoskami otwartymi. W polskiej mowie występuje sześć samogłosek: a, e, i, o, u, y. Przy wymawianiu spółgłosek trakt głosowy jest mocno zwężony lub zaciśnięty przynajmniej w jednym miejscu. Z tego względu spółgłoski nazywane są głoskami zamkniętymi. Przy tworzeniu wszystkich samogłosek pobudzeniem jest ton krtaniowy, są to więc głoski dźwięczne. Dalszego podziału samogłosek można dokonać ze względu na położenie języka w jamie ustnej i kształt warg w czasie ich wymawiania [WWW2]:

1) położenie języka w poziomie
   - głoski przednie: e, i, y
   - głoski środkowe: a
   - głoski tylne: o, u

2) położenie języka w pionie
   - głoska niska: a
   - głoski średnie: e, o
   - głoski wysokie: i, y, u

3) ułożenie warg
   - głoski okrągłe: o, u

- głoski płaskie: i, y, e
- głoska obojętna: a

Z kolei każdą spółgłoskę charakteryzuje pięć cech [WWW2]:

1) występowanie pobudzenia dźwięcznego
   - głoski dźwięczne, np. b, d, l, m, z
   - głoski bezdźwięczne, np. f, h, s, sz
2) położenie podniebienia miękkiego
   - głoski nosowe: m, n, ń
   - głoski ustne – pozostałe
3) położenie środkowej części języka wobec podniebienia twardego
   - głoski twarde, np. b, m, p, t
   - głoski miękkie, np. ź, ń
4) miejsce artykulacji tzn. punkt największego zbliżenia narządów w jamie ustnej
   - głoski wargowe np. b, m, p
   - głoski przedniojęzykowe, np. d, t, d, s, z
   - głoski środkowojęzykowe, np. j, ś, ź
   - głoski tylnojęzykowe: k, g, ch
5) stopień zbliżenia narządów mowy:
   - głoski zwarte, np. b, d, p, t
   - głoski szczelinowe, np. f, w, s, z
   - głoski zwarto-szczelinowe: c, dz, cz, ć, dź
   - głoski półotwarte, np. m, n, l, r

Spółgłoski półotwarte charakteryzują się tym, że w trakcie ich artykulacji zwarciu narządów w jednym miejscu jamy ustnej towarzyszy jednoczesne otwarcie w innym miejscu lub dochodzi do zbliżenia narządów mowy, ale nie do powstania szczeliny. W pierwszym przypadku mamy do czynienia z tak zwanymi półspółgłoskami. Wszystkie te głoski są dźwięczne i tym różnią się od innych spółgłosek, że nie mają bezdźwięcznych odpowiedników. Ze względu na tę właściwość noszą nazwę sonarnych. W drugim przypadku powstają tak zwane półsamogłoski, które są niezgłoskotwórczymi odpowiednikami samogłosek, np. „j” jest artykulacyjnie spółgłoskowym odpowiednikiem samogłoski „i”. W logopedii wyróżnia się ponadto głoski

dentalizowane (zwane również sybilantami), których wymawianie wymaga zbliżenia siekaczy górnych i dolnych. Dzieli się je na:

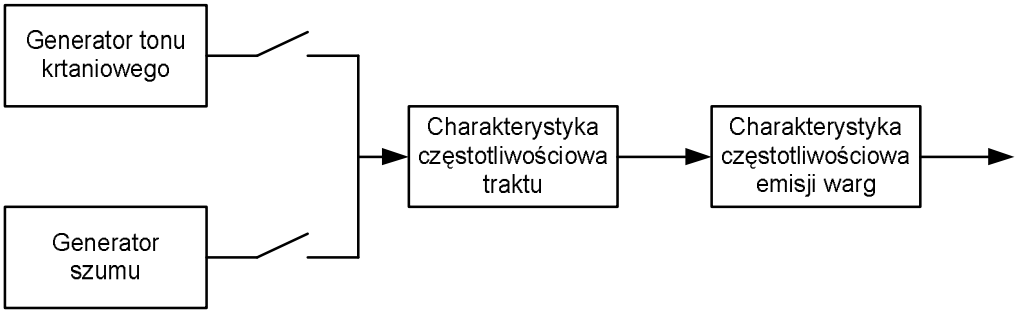
- głoski szumiące: sz, ż , cz, dż
- głoski syczące: s, z, c, dz
- głoski ciszące: ś, ź, ć, dź

Każda głoska jest fizyczną realizacją jakiegoś fonemu. Fonem jest pojęciem abstrakcyjnym, stosowanym w fonologii. Jest to zbiór fonologicznie relewantnych (ważnych) cech głoski (wymienionych powyżej), czyli cech istotnych dla funkcji komunikatywnej języka. Fonem nie posiada znaczenia, ale jest nośnikiem jego zmiany (np. zmiana fonemu /sz/ na /s/ w słowie "*kosz*" spowodowałaby zmianę jego znaczenia) [WWW2]. Ze względu na pewne czynniki, takie jak akcent i płeć mówcy czy efekty koartykulacji (wypowiadania głosek w ciągu fonetycznym), dany fonem może mieć wiele realizacji akustycznych, zwanych alofonami. Alofony oprócz cech fonologicznie relewantnych zawierają również szereg cech nierelewantnych, które zależą od mówcy oraz sąsiednich głosek w wyrazie [DE93].

Do zapisania fonemów używa się Międzynarodowego Alfabetu Fonetycznego IPA (ang. *Internetional Phonetic Alphabet*). Jest to system transkrypcji fonetycznej przyjęty przez Międzynarodowe Towarzystwo Fonetyczne [WWW1] jako ujednolicony sposób przedstawiania głosek wszystkich języków. Jednak, aby ułatwić czytanie pracy, nie będziemy z niego korzystać. Dla zapisania fonemów używać będziemy polskich znaków.

2.3. Właściwości sygnału mowy

W podrozdz. 2.1 przyjęto, że proces mówienia jest operacją filtracji akustycznej, w której pobudzeniem jest strumień powietrza wydobywający się z dolnych narządów głosowych, a filtrem – trakt głosowy. Schemat zastępczy układu wytwarzania mowy można więc przedstawić tak, jak na rys. 2.2. Pobudzenie traktu głosowego stanowi ton krtaniowy lub szum. Trakt głosowy filtruje pobudzenie zmieniając jego widmo bieżące zgodnie z charakterystyką częstotliwościową odpowiadającą aktualnemu kształtowi traktu głosowego. Wyróżnić można również drugi etap filtracji, zgodnie z charakterystyką częstotliwościową emisji, która związana jest głównie z ułożeniem warg.

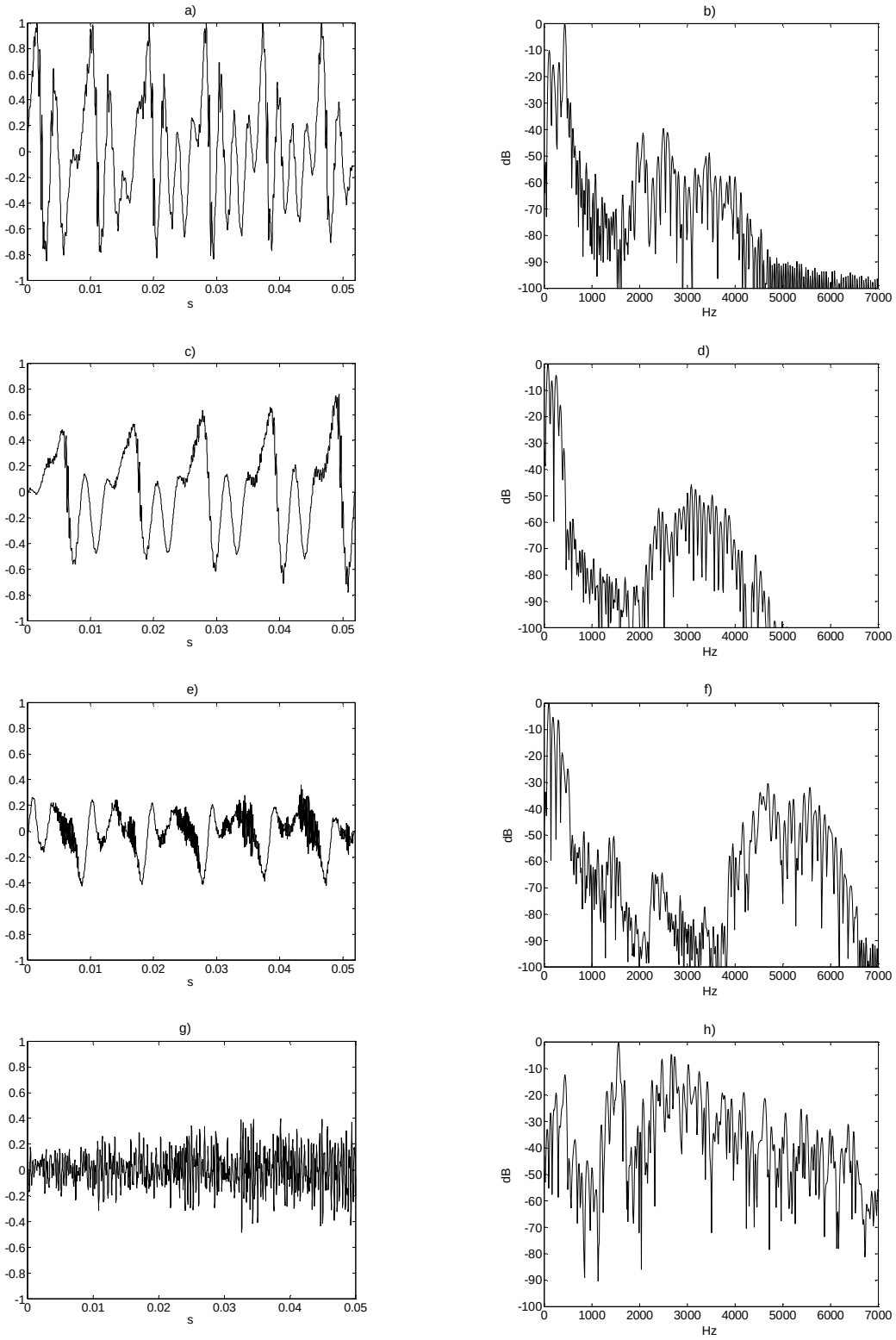


Rys. 2.2. Schemat zastępczy układu wytwarzania mowy.

Często jednak te dwa filtry modeluje się jako jeden. Na wyjściu tego układu otrzymujemy głos, który, dla celów przetwarzania, zamieniany jest za pomocą mikrofonu z fali akustycznej na postać elektryczną (sygnał mowy). Tradycyjnie przyjmuje się, że opisana filtracja jest liniowa. Warto również zauważyć, że zarówno pobudzenie jak i charakterystyka częstotliwościowa traktu głosowego zmieniają się w czasie, a sygnał mowy jest sygnałem niestacjonarnym (można jednak mówić o jego quasi-stacjonarności). Z tego względu nie można również mówić o okresowości sygnału mowy. Sygnał mowy, w odniesieniu do głosek dźwięcznych, jest quasi-okresowy, bliski okresowemu w krótkich przedziałach czasu.

W poprzednim podrozdziale przedstawiony został podział głosek. Wiemy już, że samogłoski są głoskami otwartymi, a spółgłoski zamkniętymi (lub półotwartymi). Zwężenia występujące w trakcie głosowym podczas wypowiadania spółgłosek powodują, że mają one mniejszą moc średnią niż samogłoski. Ponadto, wszystkie samogłoski są dźwięczne, podczas gdy pobudzenie dla spółgłosek może mieć różny charakter. Rys. 2.3. przedstawia oscylogramy i widma samogłoski /e/ oraz trzech spółgłosek (/j/, /sz/, /z/). Głoski zostały wycięte z jednej frazy wypowiadanej przez mężczyznę.

W przebiegach czasowych i w widmach samogłoski oraz spółgłosek /j/ i /z/ wyraźnie widać quasi-okresowość sygnału, wynikającą z cyklicznego pobudzenia. Można również zauważyć, że w głosce /z/ jednocześnie z pobudzeniem dźwięcznym występuje pobudzenie szumowe, objawiające się dużym udziałem wysokich częstotliwości (powyżej 4 kHz) w widmie sygnału. Bezdźwięczna spółgłoska /sz/ ma natomiast charakter wyraźnie szumowy. W widmach wszystkich głosek można zauważyć rezonanse formującego je filtru na pewnych częstotliwościach. Są to formanty traktu głosowego, które zostaną dokładniej omówione w dalszej części rozdziału.



Rys. 2.3. Oscylogramy (lewa kolumna) oraz widma (prawa kolumna) głosek: /e/ (pierwszy wiersz), /j/ (drugi wiersz), /z/ (trzeci wiersz) oraz /sz/ (czwarty wiersz).

Warto również zwrócić uwagę na inną cechę sygnału mowy – w widmach głosek, zwłaszcza dźwięcznych, widać, że składowe na niższych częstotliwościach mają wyższą amplitudę niż składowe wysokoczęstotliwościowe. Szczególnie dobrze uwidocznione jest to w widmie samogłoski /e/ na rys. 2.3, najmniej – w widmie głoski szumowej /sz/. Można więc powiedzieć, że sygnał mowy, w szczególności głoski dźwięczne, są sygnałami prawie minimalnofazowymi. Wynika to z pewnością z większego tłumienia wyższych częstotliwości w powietrzu, ale także z charakterystyki traktu głosowego i pobudzenia. Cecha minimalnofazowości sygnału mowy zostanie bardziej szczegółowo omówiona w rozdz. 5.

2.3.1. Ton krtaniowy

Ton krtaniowy stanowi pobudzenie dla wszystkich głosek dźwięcznych. Jak już wcześniej wspomniano, powstaje on w wyniku przecinania słupa powietrza z płuc przez drgające więzadła głosowe. Powoduje to cykliczne narastanie i opadanie ciśnienia powietrza. Zatem ton krtaniowy jest faktycznie prawie okresowym ciągiem impulsów. Okres tonu krtaniowego równy jest przedziałowi czasu pomiędzy kolejnymi chwilami zamknięcia głośni. Jego odwrotnością jest częstotliwość podstawowa F_0, nazywana często formantem *F0* (jest to nazewnictwo umowne, gdyż formant *F0* nie jest związany z żadnym rezonansem traktu głosowego).

Częstotliwość podstawowa jest jednym z głównych parametrów opisujących sygnał mowy. Jest ona powiązana z wysokością głosu odbieraną przez słuchacza. F_0 jest wielkością zmieniającą się w czasie (zdarza się, że odstęp między kolejnymi impulsami tonu krtaniowego jest różny w każdym kolejnym okresie), jednak dla każdego mówcy można wyznaczyć zakres tych zmian. Zależy on od rozmiarów krtani, gęstości tkanki więzadeł głosowych oraz zakresu zmian ich długości i naprężenia. Dwie pierwsze właściwości związane są z anatomiczną budową krtani i nie ma możliwości ich zmiany. Można natomiast regulować długość i naprężenie strun głosowych zmieniając w ten sposób częstotliwość tonu krtaniowego. F_0 u małych dzieci może zawierać się w paśmie nawet 4 oktaw, podczas gdy u dorosłego człowieka zakres ten zmniejsza się do 1.5 oktawy (większy w głosach śpiewaczych szkolonych – do 3 oktaw). Częstotliwość podstawowa mowy może przyjmować wartości od ok. 90 Hz do ok. 500 Hz (dla śpiewu zakres ten zwiększa się do ok. 1 kHz) i jest ona niższa

dla mężczyzn (ok. 90-250 Hz) niż dla kobiet (ok. 120-500 Hz) [DE93]. Oprócz charakterystycznego zakresu F_0 każdy mówca ma również „naturalną częstotliwość mowy", czyli taką częstotliwość podstawową, której statystycznie używa najczęściej. Zmiany F_0 wynikają głównie z akcentów, intonacji oraz emocji mówcy.

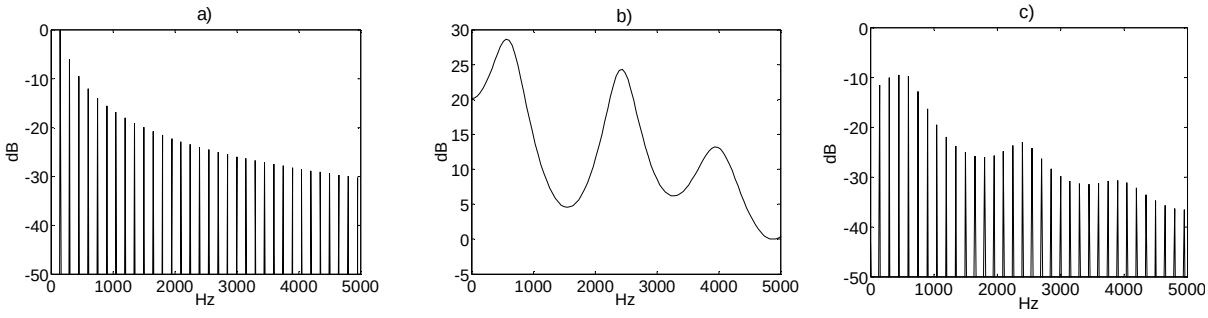
2.3.2. Trakt głosowy

Jak już wspomniano, w procesie generowania mowy pobudzenie filtrowane jest przez trakt głosowy. Zwany jest on czasem traktem głosowo-nosowym, ze względu na odgałęzienie, które stanowi jama nosowa. W tej pracy używana będzie jednak jego krótsza nazwa. Trakt głosowy jest połączeniem kilku komór. Kształty tych komór ulegają zmianom w procesie artykulacji. Każda z nich odpowiedzialna jest za powstanie formantu w wynikowym sygnale mowy. Pojęcie formantu zostało po raz pierwszy zdefiniowane przez Fanta w 1960 roku [FA60] (cytowane w [DE93]) jako maksimum w widmie amplitudowym głosu. Jest ono zazwyczaj utożsamiane z częstotliwością rezonansową traktu głosowego, czyli z częstotliwością, dla której występuje maksimum charakterystyki amplitudowej traktu. Tożsamość ta jednak nie zawsze występuje, np. przy generowaniu dźwięków o wysokiej częstotliwości podstawowej, ok. 1 kHz (sopran) w trakcie głosowym może wystąpić rezonans na niższej częstotliwości, który nie będzie miał odzwierciedlenia w strukturze formantowej wygenerowanego dźwięku. Zawsze prawdziwe jest stwierdzenie, że położenie formantów jest ściśle związane z częstotliwościami rezonansowymi traktu głosowego i zależy od jego kształtu. Słowa formant używa się również do określenia nie samego maksimum widma amplitudowego, ale zakresu widma, w którym to maksimum występuje, ograniczonego przez występujące przed nim i za nim minima. Częstotliwość, na której występuje maksimum nazywamy wtedy częstotliwością rezonansową. Zamiast częstotliwości rezonansowej można mówić również o częstotliwości środkowej formantu, czyli częstotliwości środkowej zajmowanego przez formant pasma (częstotliwość rezonansowa i środkowa mogą, ale nie muszą się pokrywać). Formant charakteryzowany jest również przez szerokość pasma. Rys. 2.4 przedstawia poglądowo sposób powstawania formantów w sygnale mowy. Pokazano na nim widmo amplitudowe tonu krtaniowego, charakterystykę amplitudową traktu głosowego oraz wynikowe widmo amplitudowe sygnału mowy. Rys. 2.4 pokazuje, że widmo tonu

krtaniowego ma charakter prawie minimalnofazowy – składowe niskoczęstotliwościowe mają amplitudę wyższą niż składowe na wyższych częstotliwościach. Dodatkowo trakt głosowy wzmacnia bardziej niższe częstotliwości, więc wynikowy sygnał mowy również ma charakter minimalnofazowy.

Na rys. 2.4. widoczne są trzy rezonanse traktu głosowego oraz trzy odpowiadające im formanty sygnału mowy. W literaturze [DE93] formanty mowy oznaczane są jako *F1*, *F2*, *F3*,…, począwszy od najniższej częstotliwości. Teoretycznie w każdym dźwięku istnieje nieskończenie wiele formantów, jednak w praktyce znajduje się ich maksymalnie pięć. Wynika to stąd, że ze względu na fizyczne możliwości narządów mowy, szerokość pasma mowy ludzkiej jest ograniczona do 7-8 kHz.

Częstotliwości rezonansowe ($F_1, F_2, F_3, \ldots$) oraz szerokości pasm formantów ($B_1, B_2, B_3, \ldots$) zależą zarówno od wypowiadanej głoski, jak i od cech indywidualnych mówcy. Jak pokazują badania [DE93], wypowiadanie konkretnych fonemów wpływa głównie na częstotliwości rezonansowe formantów *F1* i *F2* (dla niektórych głosek również *F3*). Wynika to stąd, że formanty te są ściśle związane z ułożeniem artykulatorów (z miejscami zwężenia się traktu głosowego oraz ich szerokością). Formant *F3* i wyższe zależą głównie od długości traktu głosowego i ich częstotliwości rezonansowe zmieniają się niewiele podczas wypowiadania różnych głosek. Z tego względu to niższe formanty wykorzystywane są w rozpoznawaniu fonemów [DE93]. Również szerokości pasm formantów różnią się dla różnych głosek, jednak różnice te nie są aż tak znaczące jak różnice w częstotliwościach środkowych, gdyż zależą głównie od cech osobniczych. Warto również wspomnieć, że szerokości pasm formantów rosną wraz ze wzrostem ich częstotliwości środkowych.



Rys. 2.4. Widmo amplitudowe tonu krtaniowego (a), charakterystyka amplitudowa traktu głosowego (a) oraz widmo amplitudowe sygnału mowy (c).

2.4. Przegląd wybranych modeli mowy

Najczęściej stosowane metody analizy mowy bazują na liniowym modelu „źródło-filtr", który wynika bezpośrednio z mechanizmu generowania głosu, rozumianego jako liniowa filtracja akustyczna (podr. 2.1 oraz 2.3). W modelu tym głos jest sygnałem na wyjściu filtru liniowego o zmieniającej się charakterystyce, pobudzanego przez quasi-okresowy ciąg impulsów dla mowy dźwięcznej lub szum dla mowy bezdźwięcznej. Transmitancję $H(z)$ tego filtru estymuje się najczęściej za pomocą predykcji liniowej LP (ang. *Linear Prediction*) [DE93] [RA07] [TA88], przy czym przyjmuje się transmitancję o stałym liczniku z wielomianem *p*-tego stopnia w mianowniku, nie posiadającą zer (poza *z*=0), a więc zwykle mówi się skrótowo o modelu biegunowym (ang. *all-pole model*):

$$H(z) = \frac{G}{1 - \sum_{i=1}^{p} a_i z^{-i}} \tag{2.1}$$

W (2.1) *G* jest wzmocnieniem filtru, *p* jest stopniem mianownika transmitancji, określającym również rząd predykcji, a a_i są współczynnikami mianownika transmitancji, zwanymi również współczynnikami predykcji liniowej (LPC od ang. *Linear prediction Coefficients*). Ten model autoregresji (AR) dobrze reprezentuje rezonansowy charakter traktu głosowego (bieguny transmitancji odpowiadają rezonansom traktu), a jednocześnie sprowadza problem znalezienia współczynników predykcji do rozwiązania układu równań liniowych. Na podstawie współczynników LP oblicza się również inne parametry: PARCORy (od ang. *Partial Correlation*) [TU00], LSF (ang. *Linear Spectral Frequencies*) [KL03] [GA07] czy LSF w skali barkowej.

Alternatywną metodą dla wyznaczania charakterystyki filtru reprezentującego trakt głosowy jest metoda cepstralna [DE93] [RA07] [CZ01]. W metodzie tej splot pobudzenia z odpowiedzią impulsową traktu głosowego zamienia się na sumę za pomocą przekształcenia homomorficznego, dzięki czemu możliwe jest rozdzielenie tych dwóch przebiegów. Stosowane przekształcenie homomorficzne polega na obliczeniu widma $X[k]$ sygnału mowy $x[n]$ za pomocą dyskretnej transformacji Fouriera (DFT od ang. *Discrete Fourier*

*Transform*ation), jego zlogarytmowaniu i przekształceniu za pomocą transformacji kosinusowej (lub DFT) jak pokazuje poniższy wzór

$$\hat{X}_c(i) = \sum_{k=0}^{K-1} \ln\big(X[k]\big) \cos\left(\frac{(n-0.5)i\pi}{K}\right) \tag{2.2}$$

W ten sposób uzyskujemy współczynniki $\hat{X}_c(i)$ cepstrum zespolonego. Natomiast przy obliczaniu cepstrum rzeczywistego zamiast widma zespolonego $X[k]$ stosujemy widmo amplitudowe $|X[k]|$. Odpowiedzi impulsowej traku głosowego i pobudzeniu odpowiadają odpowiednio niższe i wyższe współczynniki cepstralne. Odrzucając współczynniki odpowiadające pobudzeniu, a następnie stosując przekształcenie odwrotne do (2.2) możemy uzyskać charakterystykę częstotliwościową traktu. Najczęściej współczynniki cepstrum wyznacza się w skali melowej (MFCC od ang. *Mel Frequency Cepstrum Coefficients*), poprzez zastosowanie w pierwszym kroku, zamiast samej DFT, analizy częstotliwościowej z wykorzystaniem banku filtrów w skali melowej [RA07]. Cepstrum jest także wykorzystywane jako baza dla obliczenia współczynników IPSE (ang. *Improved Power Spectrum Envelope*), które pozwalają z większą dokładnością wyznaczyć lokalne maksima obwiedni widma odpowiadające rezonansom traktu głosowego [TA97].

Modelem często stosowanym w algorytmach przetwarzania mowy jest również model sinusoidalny [QU92], w którym zakłada się, że pobudzenie $e[n]$ jest sumą L sinusoid:

$$e[n] = \sum_{l=1}^{L} a_l[n] \cos\varphi_l[n] \tag{2.3}$$

gdzie l oznacza numer kolejnej sinusoidy, $a_l[n] > 0$ jest amplitudą chwilową, a $\varphi_l[n]$ jest fazą chwilową l-tej sinusoidy. Ponadto

$$\varphi_l[n] = \mathbf{A}\{\omega_l[n]\} + \varphi_{0,l} \tag{2.4}$$

gdzie $\omega_l[n]$ jest zmienną w czasie pulsacją l-tej sinusoidy, $\varphi_{0,l} \in [-\pi, \pi]$ jest jej fazą początkową, a $\mathbf{A}\{\ \}$ oznacza akumulator (n-ta próbka sygnału na wyjściu jest sumą próbek wejściowych o numerach od 0 do n). Jeżeli następnie przyjmiemy, że charakterystyka częstotliwościowa $H(\omega, n)$ filtru reprezentującego trakt głosowy może być wyrażona przez charakterystyki amplitudową $M(\omega, n)$ i fazową $\Psi(\omega, n)$, zmieniające się w czasie dyskretnym reprezentowanym przez n – numer próbki

$$H(\omega, n) = M(\omega, n) \exp[j\Psi(\omega, n)] \tag{2.5}$$

oraz jeżeli przez $M_l[n]$ i $\Psi_l[n]$ oznaczymy

$$M_l[n] = M\left(\omega_l[n], n\right) \tag{2.6}$$

$$\Psi_l[n] = \Psi\left(\omega_l[n], n\right) \tag{2.7}$$

to sygnał mowy $x[n]$ można opisać wzorem

$$x[n] = \sum_{l=1}^{L} A_l[n] \cos\left(\theta_l[n]\right) \tag{2.8}$$

gdzie

$$A_l[n] = a_l[n] M_l[n] \tag{2.9}$$

$$\theta_l[n] = \varphi_l[n] + \Psi_l[n] \tag{2.10}$$

Wzór (2.8) przedstawia sinusoidalny model sygnału mowy – sygnał mowy modelowany jest jako suma sinusoid o zmieniającej się w czasie amplitudzie i pulsacji. W praktyce sygnał mowy przetwarzany jest w ramkach i wtedy we wzorach (2.3)-(2.10) zamiast numeru próbki n stosuje się numer ramki m. Dla każdej ramki oddzielnie liczona jest krótkoczasowa transformata Fouriera. Pulsacje $\omega_l(m)$ i amplitudy $A_l(m)$ kolejnych sinusoid wyznaczane są poprzez znalezienie pików w widmie amplitudowym sygnału mowy w danej ramce. Wadą

tego modelu jest jego duża złożoność (duża liczba L sinusoid składających się na sygnał mowy).

Rozwinięciem modelu sinusoidalnego jest model HNM (ang. *Harmonic+Noise Model*) [ST98] [BA96] [KU03A] [KU04], w którym dokonywana jest dekompozycja sygnału mowy synchronicznie z częstotliwością podstawową (tzn. długość i położenie analizowanych segmentów sygnału zależy od okresu tonu krtaniowego). Dla ramek reprezentujących głoski dźwięczne widmo sygnału mowy dzielone jest na dwa pasma. Wyższe pasmo obejmuje częstotliwości, dla których nie występują żadne składowe harmoniczne. Niższe pasmo modelowane jest za pomocą sumy sinusoid, natomiast wyższe (jak również ramki bezdźwięczne) za pomocą szumu białego modyfikowanego przez filtr, którego transmitancja nie posiada zer.

Ograniczenia modelu liniowego [PO95], głównie pominięcie nieliniowych zjawisk występujących w procesie emisji głosu oraz założenie lokalnej stacjonarności sygnału mowy (w obrębie ramki przetwarzania), skłoniły Maragosa i in. do stworzenia nowego, nieliniowego modelu mowy [HA94] [MA95]. W modelu tym pojedynczy, l-ty formant reprezentowany jest przez zmodulowany amplitudowo i częstotliwościowo sygnał $r_l[n]$:

$$r_l[n] = a_l[n]\cos\varphi_l[n] \tag{2.11}$$

Reprezentację taką nazywamy reprezentacją AM·FM (od ang. *Amplitude Modulation – Frequency Modulation*). Amplituda chwilowa $a_l[n] > 0$ jest czynnikiem AM, natomiast $\cos\varphi_l[n]$ jest czynnikiem FM rozpatrywanego formantu. Sygnał mowy modelowany jest jako suma sygnałów AM·FM

$$x[n] = \sum_{l=1}^{L} r_l[n] \tag{2.12}$$

gdzie L jest liczbą formantów. Zauważmy podobieństwo tego modelu do modelu sinusoidalnego – sygnał mowy jest w obu przypadkach modelowany jako suma sygnałów zmodulowanych amplitudowo i częstotliwościowo. Jednak model Maragosa i in [MA95]. jest znacznie mniej złożony, gdyż za pomocą sygnałów AM·FM modeluje się całe formanty, a nie

pojedyncze harmoniczne w widmie sygnału mowy. Ponadto Maragos i in. rezygnują z liniowego modelu „źródło-filtr", co daje możliwość obserwacji nieliniowych i zmiennych w czasie zjawisk, występujących w procesie generowania mowy. Czynniki AM i FM mają prostą interpretację fizyczną, a jednocześnie unika się rozwiązywania trudnego problemu rozplotu pobudzenia i odpowiedzi impulsowej traktu głosowego. Dodatkowo model AM·FM pozwala na zbadanie znaczenia modulacji amplitudy i fazy dla percepcji mowy [PO95].

Z modelem zaproponowanym przez Maragosa i in. ściśle związane jest pojęcie częstotliwości chwilowej, które zostanie omówione szczegółowo w kolejnym rozdziale. W dalszej części pracy przedstawione zostaną również szczegóły reprezentacji AM·FM.

3. Koncepcja zespolonej pulsacji chwilowej

Najbardziej intuicyjną definicją częstotliwości jest ta sformułowana dla sygnałów okresowych, która mówi, że częstotliwość jest liczbą cykli występujących w jednostce czasu. Jednak taka definicja stwarza problemy estymacji częstotliwości już dla sygnałów quasi-periodycznych, a dla sygnałów aperiodycznych czyni ją niemożliwą. Dlatego nieodzowna jest bardziej ogólna definicja częstotliwości jako szybkości zmian fazy w czasie. Tej drugiej definicji odpowiada właśnie częstotliwość chwilowa, która stanowi pochodną fazy po czasie.

W dalszej części tego rozdziału dokładniej wyjaśnimy koncepcję częstotliwości chwilowej (IF od ang. *Instantaneous Frequency*) oraz przytoczymy różne próby jej zdefiniowania i interpretacji. Opierając się na definicji IF przedstawimy także pojęcie zespolonej pulsacji chwilowej, która jest głównym narzędziem wykorzystywanym w niniejszej pracy.

3.1. Definicja częstotliwości chwilowej

Pojęcie częstotliwości chwilowej pojawiło się w teorii sygnałów już w latach 30. XX wieku. Jak wiele innych pojęć z tej dziedziny oryginalnie odnosiło się ono do modulacji częstotliwości FM wykorzystywanej w telekomunikacji. Pierwszą formalną definicję IF przypisuje się Carsonowi i Fry’owi [CA37] (cytowane w [B092a]), którzy w 1937 roku rozważyli zmodulowany częstotliwościowo sygnał $u_m(t)$

$$u_m(t) = \exp(j(\omega_0 t + \beta \int_0^t m(\tau) d\tau) \tag{3.1}$$

W powyższym wzorze ω_0 jest stałą częstotliwością nośnej, β jest indeksem modulacji, a $m(t)$ reprezentuje przesyłaną wiadomość i ma wymiar częstotliwości kątowej (autorzy [CA37] pominęli w (3.1) fazę początkową sygnału, dlatego my również jej nie uwzględniamy). Dla sygnału $u_m(t)$ Carson i Fry zdefiniowali pulsację chwilową $\omega(t)$ i częstotliwość chwilową $f(t)$ jako, odpowiednio

$$\omega(t) = \omega_0 + \beta m(t) \tag{3.2}$$

i

$$f(t) = \frac{1}{2\pi}(\omega_0 + \beta m(t)) = f_0 + \frac{\beta}{2\pi} m(t) \tag{3.3}$$

W 1946 roku Van der Pol podszedł do problemu IF inaczej, wychodząc od prostego zapisu rzeczywistego sygnału sinusoidalnego [PO46]

$$x(t) = a\cos(2\pi ft + \theta) \tag{3.4}$$

w którym a jest amplitudą, f częstotliwością, a θ fazą początkową. Argument funkcji kosinus jest oczywiście fazą sygnału. Następnie Van der Pol zdefiniował modulację fazy jako

$$\theta(t) = \theta_0[1 + \mu g(t)] \tag{3.5}$$

W powyższym wzorze $g(t)$ jest sygnałem modulującym, a μ współczynnikiem modulacji [PO46]. W konsekwencji faza chwilowa $\varphi(t)$ staje się przebiegiem zmiennym w czasie

$$\varphi(t) = 2\pi ft + \theta(t) \tag{3.6}$$

Analogicznie do (3.5) można zdefiniować modulację częstotliwości jako

$$f(t) = f_0[1 + \mu g(t)] \tag{3.7}$$

Jednak, jak zauważył autor [PO46], proste podstawienie (3.7) do (3.4) prowadziłoby do nieścisłości, gdyż uzyskana w ten sposób faza nie zgadzałaby się z (3.6). Zamiast tego Van der Pol zaproponował zapisanie sygnału $x(t)$ jako

$$x(t) = a\cos(\int_0^t 2\pi f(\tau)d\tau + \theta) \tag{3.8}$$

Na podstawie (3.8) Van der Pol wyznaczył częstotliwość chwilową $f(t)$ jako pochodną argumentu funkcji kosinus, a więc pochodną fazy $\varphi(t)$, podzieloną przez 2π, by jej mianem był herc, dochodząc tym samym do następującej definicji IF

$$f(t) = \frac{1}{2\pi}\frac{d\varphi(t)}{dt} = \frac{1}{2\pi}\varphi'(t) \tag{3.9}$$

gdzie prim oznacza pochodną.

W 1948 roku, korzystając z dotychczasowych prac dotyczących IF oraz z teorii Gabora [GA46], który zdefiniował analityczny równoważnik sygnału rzeczywistego, Ville [VI48] (cytowane w [BO92a]) podał definicję IF sygnału $x(t) = a(t)\cos\varphi(t)$ jak następuje

$$f(t) = \frac{1}{2\pi}\frac{d}{dt}\arg u(t) \tag{3.10}$$

Tu, w (3.10), $u(t) = a(t)\exp(j\varphi(t))$ jest analitycznym równoważnikiem rzeczywistego sygnału $x(t)$, który dokładniej opiszemy w dalszej części pracy. Ta definicja, prosta i intuicyjna, przyjęła się powszechnie i jest obecnie najczęściej stosowana. Ville [VI48] doszedł do wniosku, że najprostszym sposobem na wyznaczenie IF sygnału rzeczywistego jest skorzystanie z odpowiadającego mu sygnału zespolonego. Dowiódł także, że średnia IF po czasie równa się dokładnie średniej częstotliwości w widmie sygnału $u(t)$. Był to argument przemawiający za przyjęciem zaproponowanej przez niego definicji IF, gdyż pozwalał powiązać IF z fourierowskim widmem sygnału.

3.2. Równoważnik analityczny sygnału rzeczywistego

Jak pokazano w podrozdz. 3.1, wygodnie jest definiować IF rzeczywistego sygnału $x(t)$ wykorzystując jego zespoloną reprezentację $u(t)$. Przede wszystkim dla sygnału zespolonego można jednoznacznie wyznaczyć fazę $\varphi(t)$, której pochodna stanowi IF, podczas gdy dla sygnału rzeczywistego $x(t) = a(t)\cos\varphi(t)$ istnieje nieskończenie wiele par

przebiegów $(a(t),\ \cos\varphi(t))$, których iloczyn daje $x(t)$, nie można więc jednoznacznie określić $\varphi(t)$. Zastosowanie sygnału zespolonego dla wyznaczenia IF jest więc uzasadnione. Trudność polega jednak na znalezieniu odpowiedniej reprezentacji zespolonej sygnału $x(t)$, takiej by wyznaczona na jej podstawie IF byłą szukaną częstotliwością chwilową sygnału rzeczywistego. U podstaw rozwiązania tego problemu leży sformułowana już w 1743 roku przez Leonharda Eulera (cytowane w [HA07]) tożsamość

$$e^{jz} = \cos(z) + j\sin(z) \tag{3.11}$$

wiążąca funkcje trygonometryczne $\cos(\ )$ i $\sin(\ )$ z funkcją wykładniczą argumentu urojonego. W 150 lat później Kennely i Steinmetz (cytowane w [HA07]) wykorzystali wzór Eulera do zdefiniowania sinusoidy zespolonej (kompleksoidy)

$$e^{j\omega t} = \cos(\omega t) + j\sin(\omega t), \quad \omega = 2\pi f \tag{3.12}$$

Jednak dopiero wprowadzona na początku XX wieku transformacja Hilberta pozwoliła Gaborowi [GA46] zdefiniować w 1946 roku równoważnik analityczny, uznawany powszechnie za jedyną bezsporną reprezentację zespoloną sygnału rzeczywistego.

3.2.1. Transformacja Hilberta

Transformacja (przekształcenie) Hilberta, tak jak każda inna transformacja jednowymiarowa, przekształca funkcję jednej zmiennej (najczęściej rzeczywistej, np. czasu) na, w ogólności zespoloną, funkcję innej, często zespolonej zmiennej, co można krótko zapisać jako

$$x(t) \Leftrightarrow X(z) \tag{3.13}$$

gdzie $x(t)$ i $X(z)$ stanowią parę transformat. Dla transformacji Hilberta (HT od ang. *Hilbert Transformation*) zmienna *z*, tak samo jak *t*, jest zmienną czasu. HT przekształca więc jedną

funkcję czasu w inną funkcję czasu, stąd często dla transformaty Hilberta stosuje się oznaczenie będące funkcją czasu, np. $y(t)$

$$y(t) = H_T\{x(t)\} \tag{3.14}$$

gdzie symbol operatora $H_T\{\ \}$ jest oznaczeniem idealnej transformacji Hilberta

$$H_T\{x(t)\} = \frac{1}{\pi}\int_{-\infty}^{+\infty}\frac{x(\tau)}{t-\tau}d\tau \tag{3.15}$$

Przyjęło się również zapisywać parę transformat Hilberta wykorzystując symbol splotu liniowego

$$y(t) = \frac{1}{\pi t} * x(t) \tag{3.16}$$

$$x(t) = -\frac{1}{\pi t} * y(t) \tag{3.17}$$

gdzie $h_T(t) = \frac{1}{\pi t}$ jest odpowiedzią impulsową filtru zwanego idealnym transformatorem Hilberta, którego charakterystyka częstotliwościowa $H_T(\Omega)$ (zespolona, a ściślej urojona) ma postać

$$H_T(\Omega) = \begin{cases} j & \Omega < 0 \\ 0 & \Omega = 0 \\ -j & \Omega > 0 \end{cases} \tag{3.18}$$

Znając $H_T(\Omega)$ można zapisać widmo $Y(\Omega)$ sygnału $y(t) = H_T\{x(t)\}$, będącego transformatą Hilberta sygnału $x(t)$ o widmie $X(\Omega)$ jako

$$Y(\Omega) = -j\,\mathrm{sgn}(\Omega)X(\Omega) \tag{3.19}$$

gdzie $\mathrm{sgn}(\ )$, czyli signum (znak), jest zdefiniowana dla $x \in \mathbf{R}$ jako

$$\mathrm{sgn}(x) = \begin{cases} x/|x|, & x \neq 0 \\ 0, & x = 0 \end{cases} \tag{3.20}$$

W tabeli 3.1 przedstawiono inne wybrane właściwości HT. Dokładny ich opis wraz z wyprowadzeniami znaleźć można w literaturze [HA95].

TAB. 3.1. WYBRANE WŁAŚCIWOŚCI TRANSFORMATY HILBERTA

L.p.	*Właściwość*	*Sygnał*	*Transformata Hilberta*
1	Oznaczenia	$x(t)$	$y(t) = H_T\{x(t)\}$
2	Transformaty sygnałów sinusoidalnych	$\cos(\omega t)$ $\sin(\omega t)$	$\sin(\omega t)$ $-\cos(\omega t)$
3	Liniowość	$ax_1(t) + bx_2(t)$	$ay_1(t) + by_2(t)$
4	Skalowanie w dziedzinie czasu	$x(at), \quad a > 0$	$y(at)$
5	Przesunięcie na skali czasu	$x(t-a)$	$y(t-a)$
6	Parzystość (e)/nieparzystość (o)	$x_{1e}(t) + x_{2o}(t)$	$y_{1o}(t) + y_{2e}(t)$
7	Pochodna po czasie	$x'(t)$	$y'(t) = \frac{1}{\pi t} * x'(t)$
8	Splot	$x_1(t) * x_2(t) = -y_1(t) * y_2(t)$	$x_1(t) * y_2(t) = y_1(t) * x_2(t)$
9	Energia	$\int_{-\infty}^{+\infty} x^2(t)dt$	$\int_{-\infty}^{+\infty} y^2(t)dt = \int_{-\infty}^{+\infty} x^2(t)dt$
10	Iloczyn sygnałów, których widma nie nakładają się (Tw. Bedrosiana)	$x_1(t)x_2(t)$ $x_1(t)$ – sygnał dolnopasmowy $x_2(t)$ – sygnał górnopasmowy	$y_1(t)y_2(t)$ $y_1(t) = H_T\{x_1(t)\}$ $y_2(t) = H_T\{x_2(t)\}$
11	Dwukrotna transformacja	$H_T\{x(t)\}$	$-x(t)$

Należy zaznaczyć, że właściwości 6 i 7 są prawdziwe, gdy sygnał $x(t)$ nie zawiera składowej stałej, która nie jest przenoszona przez transformator Hilberta.

3.2.2. Sygnał analityczny Gabora

Gabor [GA46] wykorzystał transformację Hilberta do zdefiniowania równoważnika analitycznego sygnału rzeczywistego. Założył on, że częścią rzeczywistą sygnału zespolonego $u(t)$ powinien być dany sygnał rzeczywisty $x(t)$ oraz, że widmo sygnału zespolonego $U(\Omega)$ powinno być takie jak widmo sygnału rzeczywistego $X(\Omega)$ dla dodatnich częstotliwości i równe zeru dla częstotliwości ujemnych. Z założeń tych wynikało, że

$$u(t) = 2\frac{1}{2\pi}\int_0^{\infty} X(\omega)e^{j\omega t}d\omega \tag{3.21}$$

Mnożenie przez 2 we wzorze (3.21) gwarantuje, że część rzeczywista $u(t)$ jest równa $x(t)$, a nie $x(t)/2$. Po wykonaniu odpowiednich podstawień i przekształceń [CO95] równości (3.21), otrzymujemy

$$u(t) = x(t) + \frac{j}{\pi}\int_{-\infty}^{+\infty} \frac{x(\tau)}{t-\tau}d\tau \tag{3.22}$$

Porównując (3.22) z (3.15) można zauważyć, że część urojona $u(t)$ w (3.22) jest transformatą Hilberta sygnału $x(t)$

$$u(t) = x(t) + jH_T\{x(t)\} \tag{3.23}$$

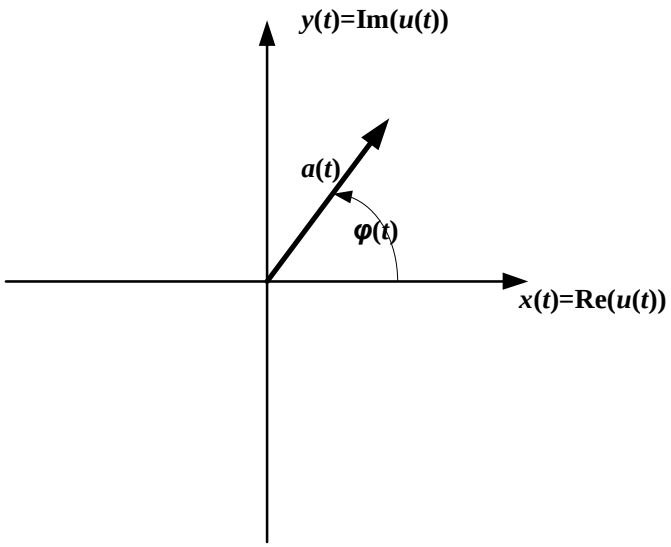
Zdefiniowany przez (3.23) sygnał jest funkcją analityczną, tzn. na zadanym obszarze spełnia równanie Cauchy’ego-Riemanna (dowód podaje Hahn [HA95]). Z tego względu zdefiniowany przez Gabora sygnał zespolony nazywany jest sygnałem analitycznym lub równoważnikiem analitycznym sygnału rzeczywistego.

3.2.3. Reprezentacja sygnału analitycznego przez wskaz

Jak podaje Hahn [HA95] atrybuty chwilowe (amplitudę, fazę, częstotliwość), można przystępnie, a przede wszystkim jednoznacznie zdefiniować, a także graficznie zinterpretować, wykorzystując postać wykładniczą (biegunową) sygnału analitycznego $u(t)$

$$u(t) = a(t)e^{j\varphi(t)} \tag{3.24}$$

Graficznie sygnał $u(t)$ można przedstawić jako wektor zależny od czasu – wskaz (fazor) o zmiennej długości wirujący na płaszczyźnie zespolonej $(x(t), y(t))$ ze zmienną prędkością, jak pokazuje rys. 3.1. Płaszczyzna zespolona nazywana jest płaszczyzną Gaussa, a trajektoria kreślona przez wirujący wskaz nazywa się diagramem Agranda.



Rys. 3.1. Graficzna reprezentacja sygnału $u(t)$.

Postać wykładniczą (3.24) uzyskuje się z postaci algebraicznej $u(t) = x(t) + jy(t)$, gdzie $y(t)$ zdefiniowano w (3.14 – 3.16), na podstawie wzorów

$$a(t) = \sqrt{x^2(t) + y^2(t)} = |u(t)| \tag{3.25}$$

$$\varphi(t) = \arg(u(t)) \tag{3.26}$$

Stosowana w powyższym zapisie zmienna w czasie nieujemna wielkość $a(t)$ nazywa się amplitudą chwilową (IA od ang. *Instantaneous Amplitude*) sygnału $u(t)$ i określa chwilową wartość długości wskazu, natomiast $\varphi(t)$ jest jego fazą chwilową i określa chwilową wielkość kąta wskazu. Znając definicję fazy chwilowej można następnie zdefiniować pulsację chwilową $\omega(t)$ sygnału $u(t)$, będącą prędkością kątową obracającego się na płaszczyźnie ($x(t)$, $y(t)$) wskazu (rys. 3.1), jako pochodną po czasie fazy $\varphi(t)$

$$\omega(t) = \frac{d\varphi(t)}{dt} \tag{3.27}$$

Stąd jednoznaczną definicją częstotliwości chwilowej $f(t)$ wyrażanej w hercach [Hz] jest (por. (3.9))

$$f(t) = \frac{\omega(t)}{2\pi} = \frac{1}{2\pi}\frac{d\varphi(t)}{dt} \tag{3.28}$$

Jednostką pulsacji chwilowej są, tak jak i dla pulsacji sygnałów okresowych, radiany na sekundę [rad/s]. Korzystając z graficznej interpretacji z rys. 3.1 można zauważyć, że dodatnia wartość IF oznacza obrót wskazu w kierunku dodatnim, tj. przeciwnym do ruchu wskazówek zegara. Zmiana znaku IF odpowiada zmianie kierunku obrotu wskazu.

3.2.4. Reprezentacja AM·FM

Dla definicyjnego wyznaczenia IF konieczna jest znajomość fazy sygnału, stąd sygnał rzeczywisty, którego IF szukamy, zapisujemy w postaci

$$x(t) = a(t)\cos\varphi(t) \tag{3.29}$$

Prawa strona równości (3.29) jest tak zwaną reprezentacją AM·FM sygnału $x(t)$. Traktuje ona $x(t)$ jako sygnał zmodulowany amplitudowo i częstotliwościowo. W powyższym zapisie przebieg $a(t)$, zwany amplitudą chwilową lub obwiednią sygnału, jest czynnikiem AM, natomiast $\cos\varphi(t)$ jest czynnikiem FM, gdyż pochodna fazy chwilowej $\varphi'(t)$ zgodnie z (3.27) jest zmodulowaną pulsacją chwilową $\omega(t)$. Jak już wcześniej wspomniano, jednoznaczne wyznaczenie czynników AM i FM dla sygnału rzeczywistego nie jest możliwe, gdyż istnieje nieskończenie wiele par przebiegów $(a(t), \cos\varphi(t))$, których iloczyn daje $x(t)$. Problem znalezienia właściwych przebiegów $a(t)$ i $\cos\varphi(t)$ nazywany jest często dekompozycją AM·FM sygnału, my jednak będziemy posługiwać się terminem faktoryzacja lub bifaktoryzacja, gdyż dwa poszukiwane przebiegi są czynnikami iloczynu (ang. *factors of product*), natomiast nazwa dekompozycja może mylnie wskazywać, że poszukiwane są składowe superpozycji (ang. *components of superposition*).

Vakman [VA78] [VA96] [VA98] argumentuje, że ponieważ istnieje wiele sposobów wyznaczenia amplitudy i fazy sygnału rzeczywistego, uzasadnione jest zdefiniowanie pewnych warunków, które powinna spełniać para $(a(t), \cos\varphi(t))$, aby faktoryzacja taka miała sens (by możliwa była jej fizyczna interpretacja). Vakman proponuje trzy takie warunki (tzw. postulaty Vakmana):

1) niewielka zmiana wartości sygnału $x(t)$ powinna powodować odpowiednio małą zmianę przebiegu $a(t)$;
2) wymnożenie sygnału $x(t)$ przez stałą $c > 0$ nie powinno zmieniać fazy $\varphi(t)$. Powinno natomiast powodować wymnożenie amplitudy $a(t)$ przez tę samą stałą;
3) IF oraz IA wyznaczone dla sygnału sinusoidalnego $a_0 \cos(\omega_0 t + \varphi_0)$ powinny wynosić, odpowiednio, $\omega_0/(2\pi)$ oraz a_0.

Jednocześnie Vakman udowadnia, że sygnał $u(t) = x(t) + jT\{x(t)\}$, gdzie $T\{\ \}$ jest operacją, za pomocą której z rzeczywistego sygnału uzyskujemy sygnał zespolony, spełnia powyższe warunki tylko wtedy, gdy $T\{\ \}$ jest transformacją Hilberta, a $u(t)$ jest równoważnikiem analitycznym sygnału $x(t)$.

Znalezienie równoważnika analitycznego sygnału rzeczywistego $x(t) = a(t)\cos\varphi(t)$ jest najczęściej stosowanym sposobem bifaktoryzacji AM·FM. Zakłada się, że równoważnik analityczny ma tę samą amplitudę i fazę co rzeczywisty sygnał $x(t)$, tzn.

$$u(t) = x(t) + jH\{x(t)\} = a(t)e^{j\varphi(t)} \tag{3.30}$$

Bifaktoryzację AM·FM będziemy prościej zapisywać zastępując czynnik $e^{j\varphi(t)}$ pojedynczym symbolem $\gamma(t)$, oznaczającym fazor FM:

$$u(t) = a(t)\gamma(t) \tag{3.31}$$

Biorąc pod uwagę wzór Eulera (3.11) oraz definicję zespolonego przebiegu harmonicznego Kennely'ego i Steinmetza (3.12), można stwierdzić, że prawa równość w (3.30) będzie prawdziwa tylko wtedy, gdy dla $x(t) = a(t)\cos\varphi(t)$, $H_T\{x(t)\} = a(t)\sin\varphi(t)$. Ponieważ $x(t)$ jest iloczynem przebiegów $a(t)$ i $\cos\varphi(t)$, aby zweryfikować (3.30) należy skorzystać z twierdzenia Bedrosiana [BE63], które mówi, że jeżeli pasma zajmowane przez widma dwóch sygnałów są rozłączne (nie nakładają się) to transformata Hilberta ich iloczynu jest równa iloczynowi czynnika „dolnopasmowego” i transformaty Hilberta czynnika „górnopasmowego”. Oznacza to, że jeżeli widma: $U_1(\Omega)$ i $U_2(\Omega)$, sygnałów, odpowiednio, $u_1(t)$ i $u_2(t)$ spełniają warunek

$$\begin{aligned} U_1(\Omega) &= 0 \quad \text{dla} \quad |\Omega| > a \\ U_2(\Omega) &= 0 \quad \text{dla} \quad |\Omega| < b \end{aligned} \tag{3.32}$$

gdzie $b \geq a \geq 0$, to transformatę Hilberta ich iloczynu wyraża wzór

$$H_T\{u_1(t)u_2(t)\} = u_1(t)H_T\{u_2(t)\} \tag{3.33}$$

Bedrosian [BE63] wykazał również, że (3.33) jest spełnione, gdy oba sygnały, $u_1(t)$ i $u_2(t)$, są analityczne. Później twierdzenie Bedrosiana zostało rozszerzone [NU66]. Dla dwóch sygnałów $u_1(t)$ i $u_2(t)$, ogólnie zespolonych, prawdziwa jest równość (3.33), jeżeli ich widma spełniają warunek

$$\begin{aligned} U_1(\Omega) &= 0 \quad \text{dla} \quad \Omega < -a \\ U_2(\Omega) &= 0 \quad \text{dla} \quad \Omega < a \end{aligned} \tag{3.34}$$

gdzie $a > 0$. Warto zauważyć, że w późniejszej literaturze można znaleźć wiele przykładów innych sformułowań warunków koniecznych i wystarczających dla spełnienia twierdzenia Bedrosiana [RI66][CA73] [BR74][BR86][XU06][TA09].

3.3. IF sygnałów wielokomponentowych

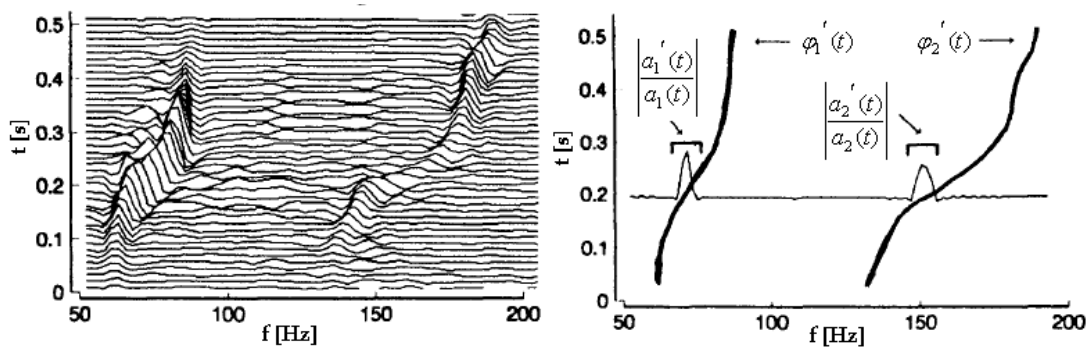
Analizując rozkłady czasowo-częstotliwościowe wielu naturalnych sygnałów, takich jak głos czy nawoływanie niektórych zwierząt, można zauważyć wyraźnie wyodrębniające się w ich strukturze liczne komponenty. W sygnale mowy są to formanty. Takie sygnały nazwano w literaturze sygnałami wielokomponentowymi (ang. *multicomponent*). Korzystając z definicji Cohena [CO92], sygnałami wielokomponentowymi nazywać będziemy sygnały, będące sumą dwóch lub więcej monokomponentów postaci $a(t)e^{j\varphi(t)}$, spełniające ponadto warunek, że szerokości pasm komponentów są mniejsze niż odstępy między nimi na osi częstotliwości. Sygnały jednokomponentowe charakteryzują się tym, że pomiędzy sąsiednimi przejściami sygnału przez zero występuje tylko jedno ekstremum oraz tym, że przebieg ich IF nie zmienia znaku. Przykładami sygnałów jednokomponentowych są sygnały wąskopasmowe i sygnały o wolnej modulacji częstotliwościowej. IF sygnałów wielokomponentowych przyjmuje wartości i dodatnie, i ujemne, a pomiędzy kolejnymi przejściami przez zero sygnału może wystąpić więcej niż jedno ekstremum. Podstawowym przykładem sygnału wielokomponentowego jest sygnał składający się z dwóch komponentów

$$u(t) = u_1(t) + u_2(t) = a_1(t)e^{j\varphi_1(t)} + a_2(t)e^{j\varphi_2(t)} \tag{3.35}$$

zilustrowany na rys. 3.2 [CO92]. Chwilowe szerokości pasm IB (ang. *Instantaneous Bandwidth*) poszczególnych komponentów Cohen definiuje jako $\left|a_l'(t)/a_l(t)\right|$, gdzie l jest numerem komponentu, natomiast odstęp między nimi jest różnicą ich IF. Sygnał (3.35) będzie więc, według definicji Cohena, sygnałem wielokomponentowym przy spełnieniu warunku

$$\left|\frac{a_1'(t)}{a_1(t)}\right|,\quad \left|\frac{a_2'(t)}{a_2(t)}\right| \quad << \left|\varphi_2'(t)-\varphi_1'(t)\right| \tag{3.36}$$

Warto tu zauważyć, że ponieważ wszystkie wielkości w (3.36) są funkcjami czasu, sygnał może być sygnałem wielokomponentowym w pewnej chwili czasu, a w innej nie.



Rys. 3.2. Sygnał wielokomponentowy, składający się z dwóch komponentów [CO92]: rozkład czas-częstotliwość sygnału zobrazowany jako „wodospad” (lewy rys.) oraz przebiegi IF obu składowych (prawy rys.).

Wielu autorów twierdzi, że wyznaczanie IF ma sens tylko dla sygnałów jednokomponentowych, gdyż tylko dla takich sygnałów IF można interpretować jako średnią częstotliwość w każdej chwili czasu [WE98]. Aby temu sprostać, konieczna jest dekompozycja sygnału i obliczenie IF oddzielnie dla każdego komponentu. Takie podejście przyjęło się szczególnie w analizie sygnału mowy (zgodnie z modelem Maragosa i in. [HA94] [MA95], o którym pisaliśmy w podrozdz. 2.4), gdyż komponentami sygnału mowy są formanty, uzasadniona jest więc indywidualna analiza każdego komponentu [HA94][MA95][LU96] [KU03B][RA00][KA08a][KA08b][BO04].

W tym miejscu warto zaznaczyć, że choć analiza pojedynczych formantów rzeczywiście znajduje uzasadnienie w strukturze widmowej mowy, tu nie zgadzamy się z założeniem, że IF obliczona dla sygnału wielokomponentowego nie ma sensu fizycznego. W

rozdziale dotyczącym analizy mowy przedstawimy interpretację IF obliczonej dla sygnału mowy o paśmie ograniczonym do 8 kHz, a więc zawierającego wszystkie istotne formanty.

3.3.1. Wybrane metody dekompozycji sygnałów wielokomponentowych

Jedną z najczęściej stosowanych metod dekompozycji sygnałów wielokomponentowych, zwłaszcza w analizie nieliniowych i niestacjonarnych sygnałów z zakresu mechaniki, akustyki czy sejsmologii, jest adaptacyjna metoda EMD (od ang. *Empirical Mode Decomposition*) [HU98]. Polega ona na iteracyjnym znajdowaniu tzw. funkcji IMF (od ang. *Intrinsic Mode Function*) i odejmowaniu ich od analizowanego sygnału, aż do uzyskania residuum, które jest przebiegiem stałym lub funkcją trendu. IMF-y, reprezentujące poszczególne komponenty sygnału, są funkcjami spełniającymi dwa warunki: a) pomiędzy każdym minimum i maksimum funkcji znajduje się przejście przez zero, oraz b) średnia obwiedni zdefiniowanej przez lokalne maksima (tzw. obwiedni górnej) i obwiedni zdefiniowanej przez lokalne minima (obwiedni dolnej) jest w każdej chwili czasu równa zeru [HU98]. Algorytm rozpoczyna się od wyznaczenia obwiedni dolnej i górnej poprzez aproksymowanie ich wielomianami 3-go stopnia oraz obliczenia ich średniej. Średnia ta jest następnie odejmowana od sygnału. Postępujemy w ten sposób aż przebieg uzyskany po odjęciu średniej stanie się funkcją IMF (tzn. będzie spełniał dwa wymienione wyżej warunki). Całą procedurę powtarzamy aż do otrzymania wszystkich IMF-ów. Dla każdego komponentu można następnie obliczyć IA i IF. Metodę EMD można wykorzystać w analizie sygnału mowy, jednak należy zaznaczyć, że znalezione monokomponenty nie odpowiadają formantom mowy [RO06], jak to zakłada model Maragosa i in. [HA94][MA95].

Metodę alternatywną, znacznie mniej złożoną obliczeniowo, a bazującą na transformacie Hilberta i częstotliwości chwilowej, zaproponował Feldman [FE06] [FE08][FE11][BR11]. Jest to metoda HVD (od ang. *Hilbert Vibration Decomposition*). Jest ona również metodą iteracyjną. Każda iteracja składa się z dwóch zasadniczych etapów: estymacji IF komponentu o największej mocy, a następnie wyznaczenia obwiedni tego komponentu. Wyodrębniony w ten sposób komponent jest następnie odejmowany od sygnału podanego na wejście algorytmu. Szczegółowo metoda HVD zostanie opisana w podrozdz. 5.3

niniejszej pracy, w którym przedstawimy sposób jej wykorzystania dla znajdowania formantów mowy.

Gianfelici i in. [GI05] [GI07] podeszli w sposób odwrotny do powyższego problemu dekompozycji, opierając swoją metodę na estymacji IA poszczególnych komponentów. Opisany przez nich algorytm IHT (od ang. *Iterated Hilbert Transform*) estymuje najpierw obwiednię amplitudową sygnału analitycznego, reprezentującego rzeczywisty sygnał wielokomponentowy, a następnie oblicza jej składowe: wolno- i szybkozmienną. Iteracyjnie obliczany jest sygnał analityczny, którego część rzeczywistą stanowi znaleziona w poprzednim kroku składowa szybkozmienna, a część urojoną – jej transformata Hilberta. Dla tak uzyskanego przebiegu znowu estymuje się obwiednię amplitudową oraz jej składowe: wolno- i szybkozmienną. Procedura ta jest powtarzana, aż uzyskana składowa szybkozmienna ma na tyle małą energię, że można ją pominąć. Obliczone w każdej iteracji składowe wolnozmienne obwiedni amplitudowych stanowią IA poszczególnych komponentów. Chwilowe fazy komponentów są natomiast obliczane jako odpowiednia kombinacja faz uzyskanych w każdej iteracji [GI07]. Gianfelici i in. pokazali w [GI05][GI07], że taka metoda dekompozycji może być wykorzystana dla znajdowania formantów sygnału mowy.

Najprostszym podejściem do dekompozycji sygnału mowy na komponenty odpowiadające formantom jest zastosowanie banku filtrów adaptacyjnych, których częstotliwości środkowe adaptują się do częstotliwości środkowych formantów. Takie podejście zaproponowali m.in. Maragos i in. [MA93] [HA94] oraz Kumaresan i Rao [KU99] [RA00]. Maragos i in. wykorzystali w swojej pracy filtry Gabora o szerokościach pasm równych odpowiednio 800 Hz dla formantów o częstotliwościach środkowych poniżej 1000 Hz i 1100 Hz dla pozostałych. Częstotliwości środkowe filtrów ustalane są wstępnie na podstawie analizy sygnału mowy inną metodą, np. LPC. Następnie estymowane są IF przebiegów na wyjściach filtrów. Stanowią one częstotliwości środkowe filtrów w kolejnej iteracji. Procedurę tą powtarza się do czasu, gdy w kolejnych dwóch iteracjach częstotliwości środkowe filtrów nie zmieniają się o więcej niż 5 Hz. Maragos i in. rozważają również możliwość wyboru różnych szerokości pasm filtrów w kolejnych iteracjach na podstawie odstępu między częstotliwościami środkowymi sąsiednich filtrów. Kumaresan i Rao [KU99][RA00] proponują natomiast użycie banku filtrów rezonansowych o przestrajanych częstotliwościach rezonansowych. Filtry te poprzedza bank filtrów samozerowych (których

transmitancje mają same zera). Filtr samozerowy znajdujący się w *i*-tej gałęzi ma tłumić częstotliwości rezonansowe filtrów rezonansowych znajdujących się w pozostałych gałęziach. Aby ustalić początkowe parametry banku filtrów Kumaresan i Rao [RA07] proponują procedurę, w której najpierw stosowany jest pojedynczy filtr rezonansowy, którego częstotliwość rezonansowa dostrajana jest do częstotliwości, dla której energia widmowa sygnału wielokomponentowego jest największa. Następnie wprowadzana jest druga gałąź filtracji z filtrem samozerowym, tłumiącym częstotliwość filtru rezonansowego w pierwszej gałęzi oraz drugim filtrem rezonansowym dostrajającym się znowu do częstotliwości, dla której energia widmowa jest największa. W ten sposób wprowadza się kolejne gałęzie aż do uzyskania założonej liczby gałęzi.

Metody dekompozycji wykorzystujące banki filtrów adaptacyjnych sprawdzają się w analizie głosu, ponieważ, generalnie, poszczególne formanty są od siebie dobrze odseparowane na osi częstotliwości. Problemem mogą być głoski, w których formant F2 „zachodzi" na formant F1 (np. samogłoski tylne /o/ i /u/). Innym problemem jest odpowiednie dobranie szerokości pasm filtrów tak, by pokrywały one całe pasmo formantu, a jednocześnie nie nachodziły na sąsiednie formanty.

3.4. Inne definicje IF

Zaproponowana przez Ville'a definicja IF jest szeroko przyjęta i stosowana, budzi jednak pewne dyskusje wśród niektórych autorów [BO92a][BO92b][LO96][MA74][OL00] [VA96][WO99][CO99][HA03]. Dotyczą one przede wszystkim problemów związanych z interpretacją IF zdefiniowanej przez Ville'a. Rozpatrywane są jej zalety i wady w odniesieniu do konkretnych zastosowań, a także sens fizycznej interpretacji IF dla konkretnych sygnałów. Celem niniejszego podrozdziału nie jest podważanie zasadności definicji IF zapisanej w (3.10). Chcemy jednak przedstawić i ustosunkować się do prac autorów, poszukujących innych definicji, które w niektórych zastosowaniach mogą się okazać bardziej celne.

Dość powszechnie przyjęła się opinia, że IF nie sprawia problemów interpretacyjnych dla sygnałów wąskopasmowych [BO92a], choć Hahn [HA95] polemizuje z tym twierdzeniem, argumentując, że przy zastosowaniu odpowiedniego przetwarzania wstępnego, można otrzymać sensowne wyniki również dla sygnałów szerokopasmowych. Wielu autorów [CO95]

[JO90][LO98][OL98a][OL99][WE98] problemy w interpretacji IF tłumaczy tym, że analizowane sygnały mogą składać się z więcej niż jednego komponentu postaci $a(t)\cos\varphi(t)$, czyli są sygnałami wielokomponentowymi, o których mowa była w podrozdz. 3.3. Inni, np. Loughlin i Tacer [LO96], podważają sensowność bifaktoryzacji AM·FM za pomocą transformacji Hilberta, a tym samym zastosowanie analitycznego równoważnika sygnału rzeczywistego dla wyznaczenia jego IF, co omówiono w p. 3.2.4. Wymienione problemy skłoniły autorów działających na tym polu do poszukiwania innych sposobów zdefiniowania parametrów chwilowych sygnału, z których wybrane opisane zostaną w tym punkcie. Odniesiemy się do nich również w dalszej części pracy.

W p. 3.2.4 wspomniano o postulatach Vakmana, w świetle których transformacja Hilberta jest najlepszym sposobem otrzymywania bifaktoryzacji AM·FM. Loughlin i Tacer [LO96] zaproponowali, by do postulatów Vakmana dodać kolejny warunek:

4) jeżeli pasmo sygnału jest ograniczone, to wartości IF powinny znaleźć się w tych samych granicach jak pasmo.

Ponadto rozszerzyli oni pierwszy warunek Vakmana uznając, że jeśli wartości sygnału są ograniczone, to amplituda $a(t)$ również powinna być ograniczona. Jednocześnie Loughlin i Tacer [LO96] wykazali, że równoważnik analityczny Gabora może nie spełniać warunku czwartego, gdyż pochodna fazy sygnału zespolonego (niekoniecznie analitycznego), może przyjmować wartości spoza pasma sygnału. Ponadto, gdy w przebiegu $x(t)$ występują nieciągłości, amplituda sygnału analitycznego jest nieograniczona nawet, jeżeli wartości $x(t)$ są ograniczone. Sygnał analityczny może więc nie spełniać uzupełnionego warunku pierwszego i czwartego. Jednocześnie Loughlin i Tacer zaproponowali metodę wyznaczania IF jako pierwszego momentu rozkładu czasowo-częstotliwościowego sygnału dla częstotliwości dodatnich (sposób liczenia takiego rozkładu można znaleźć w literaturze [LO94]). W swojej definicji wykorzystali teorię dodatnich rozkładów czasowo-częstotliwościowych, opisanych przez Cohena i Poscha [CO85]

$$\omega(t) = 2\int_{0}^{+\infty} \omega P(\omega|t)d\omega \tag{3.37}$$

gdzie $P(\omega|t)$ to unormowany rozkład czasowo-częstotliwościowy dla $\omega \in (-\infty, \infty)$, a całka liczona jest jednostronnie, by otrzymać niezerowy wynik dla sygnałów rzeczywistych. Zaproponowany sposób estymacji IF wynika ze sposobu, w jaki autorzy [LO96] interpretują częstotliwość chwilową jako średnią częstotliwość fourierowską chwilowego widma mocy w każdej chwili czasu. Znając IF, która stanowi czynnik FM analizowanego sygnału, przeprowadza się koherentną demodulację, by wyznaczyć czynnik AM. Demodulacja ta polega na przemnożeniu sygnału przez kosinus i sinus fazy $\varphi_F(t)$, zdefiniowanej jako

$$\varphi_F(t) = \int_{-\infty}^{t} \omega(\tau) d\tau \tag{3.38}$$

po której następuje filtracja dolnoprzepustowa. W ten sposób uzyskuje się składowe: synfazową $a_I(t)$ i kwadraturową $a_Q(t)$ czynnika AM

$$a_I(t) = \int_{-\infty}^{+\infty} x(\tau) \cos\varphi_F(\tau) h_{lp}(t,\tau) d\tau \tag{3.39}$$

$$a_Q(t) = \int_{-\infty}^{+\infty} x(\tau) \sin\varphi_F(\tau) h_{lp}(t,\tau) d\tau \tag{3.40}$$

We wzorach (3.39) i (3.40) $h_{lp}(t,\tau)$ jest zmieniającą się w czasie odpowiedzią impulsową filtru dolnopasmowego, którego częstotliwość odcięcia równa jest $\omega(t)$, a wzmocnienie w paśmie przepustowym wynosi 2. Taka demodulacja kwadraturowa prowadzi do uzyskania obwiedni zespolonej $a_c(t) = a_I(t) + j a_Q(t) = a(t) e^{j\varphi_A(t)}$. Czynnik AM, czyli $a(t)$, można obliczyć jako pierwiastek z sumy kwadratów składowych $a_I(t)$ i $a_Q(t)$. Ponieważ amplituda $a_c(t)$ jest przebiegiem zespolonym, sygnał $u(t)$ zapisuje się jako

$$u(t) = a(t) e^{j(\varphi_A(t) + \varphi_F(t))} \tag{3.41}$$

Całkowita faza sygnału wynosi więc $\varphi_A(t) + \varphi_F(t)$. Modulację częstotliwości sygnału określa $\omega_F(t) = \varphi'_F(t)$ i ta wielkość uznawana jest przez Louglina i Tacera za częstotliwość chwilową.

Inne podejście do problemu definicji i interpretacji IF zaprezentowali w swoich pracach Oliveira i Barroso [OL00][OL98a][OL98b][OL99]. Przyjęli oni, że $u(t) = a(t)e^{j\varphi(t)}$ reprezentuje sygnał zespolony jako wersję sygnału $a(t)$ heterodynowanego do częstotliwości $\varphi'(t)$. Wyznaczenie $\varphi'(t)$, czyli IF, polega więc na znalezieniu częstotliwości, wokół której zlokalizowane jest widmo $u(t)$. Ponieważ $a(t)$ jest sygnałem rzeczywistym o widmie amplitudowym symetrycznym względem częstotliwości zerowej, chwilowe widmo amplitudowe $|U_t(\omega)|$ (przez $U_t(\omega)$ rozumiemy rozkład czasowo-częstotliwościowy) sygnału $u(t)$, będącego wynikiem heterodynowania na częstotliwość $\varphi'(t)$ powinno być symetryczne względem $\varphi'(t)$. Jeśli tak nie jest, oznacza to, że $u(t)$ jest w rzeczywistości wynikiem heterodynowania sygnału zespolonego, a więc model, w którym heterodynowany jest sygnał rzeczywisty $a(t)$, nie jest poprawną reprezentacją. Prawidłowo model ten należałoby zapisać w postaci [OL00]

$$u(t) = a(t)e^{j\varphi_A(t)}e^{j\varphi_F(t)} \tag{3.42}$$

W powyższym modelu sygnał zespolony $a(t)e^{j\varphi_A(t)}$ jest heterodynowany do częstotliwości $\varphi'(t)$. W zapisie (3.42) faza sygnału $u(t)$ jest rozdzielana na dwie składowe: $\varphi_A(t)$ i $\varphi_F(t)$. Według Oliveiry i Barroso [OL00], składowa $\varphi_A(t)$ odpowiada za brak symetrii widma sygnału $u(t)$, natomiast $\varphi_F(t)$ reprezentuje proces heterodynowania i to ta składowa powinna być zróżniczkowana w celu otrzymania IF. Taka interpretacja sugeruje, że IF powinna być częstotliwością $\nu(t)$, dla której uzyskujemy największą symetrię chwilowego widma $U_t(\omega)$ sygnału $u(t)$, tzn. jeśli zdefiniujemy część parzystą widma chwilowego jako

$$U_{t,\nu}^{(e)}(\omega) = \frac{1}{2}\left[U_t(\omega) + U_t(2\nu - \omega)\right] \tag{3.43}$$

Wówczas IF powinna być częstotliwością, dla której część parzysta ma największą energię. Podobnie, można zdefiniować część nieparzystą jako

$$U_{t,\nu}^{(o)}(\omega) = \frac{1}{2}\left[U_t(\omega) - U_t(2\nu - \omega)\right] \tag{3.44}$$

i przyjąć, że IF jest częstotliwością, dla której część nieparzysta ma najmniejszą energię. Stąd IF można obliczyć jako

$$\nu(t) = \arg\min_{\nu}\left[\int_{-\infty}^{+\infty}\left(U_t(\omega) - U_t(2\nu - \omega)\right)^2 d\omega\right] \tag{3.45}$$

Zauważmy, że w obu przytoczonych wyżej definicjach IF, faza chwilowa sygnału rozdzielana jest na dwie składowe: $\varphi_A(t)$ i $\varphi_F(t)$, a IF jest pochodną składowej $\varphi_F(t)$. Louglin i Tacer [LO96] interpretują jednak IF jako średnią częstotliwość fourierowską chwilowego widma mocy w każdej chwili czasu, natomiast Oliveira i Barroso [OL00] jako częstotliwość, dla której uzyskujemy największą symetrię chwilowego widma sygnału $u(t)$. Przykłady innych definicji, a także sposobów interpretacji i estymacji IF można znaleźć w literaturze, m.in. w [BO92b][SA00] [PO97] [MA93].

3.5. Definicja zespolonej pulsacji chwilowej

W latach 50-tych wprowadzona została nowa wielkość chwilowa, tzw. zespolona pulsacja chwilowa (ICF od ang. *Instantaneous Complex Frequency*). Pierwszy pisał o niej Linden w 1958 roku [LI58]. Niedługo potem temat ten podjął Hahn [HA59], dając swoją pracą najistotniejszy wkład w rozwój koncepcji ICF. Pierwotnie ICF miało na celu ułatwienie opisu i analizy oscylatorów analogowych. Koncepcja ta wywodzi się od powszechnie wykorzystywanej w rachunku operatorowym w teorii obwodów, w teorii sterowania i w automatyce pulsacji zespolonej, będącej argumentem transformacji Laplace'a $s = \sigma + j\omega$. Wyprowadzenie definicji ICF [HA95] można zacząć od wyznaczenia fazy chwilowej na podstawie definicji (3.27) pulsacji chwilowej

$$\varphi(t) = \int_0^t \omega(\tau)d\tau + \varphi_0 \tag{3.46}$$

gdzie φ_0 jest fazą początkową, tj. w chwili $t = 0$. Hahn zdefiniował także pojęcie względnej chwilowej prędkości promieniowej $\sigma(t)$ (zakładając $a(t) > 0$)

$$\sigma(t) = \frac{1}{a(t)} \frac{da(t)}{dt} = \frac{a'(t)}{a(t)} \tag{3.47}$$

która określa względną prędkość zmian amplitudy chwilowej. To pozwala zapisać amplitudę chwilową w postaci wykładniczej

$$a(t) = a_0 \exp(\int_0^t \sigma(\tau) d\tau) \tag{3.48}$$

gdzie a_0 jest amplitudą początkową. Na podstawie (3.46) i (3.48) można zapisać sygnał analityczny w postaci

$$u(t) = a_0 \exp(j\varphi_0) \exp(\int_0^t [\sigma(\tau) + j\omega(\tau)] d\tau) \tag{3.49}$$

gdzie funkcja pod całką definiuje ICF

$$s(t) = \sigma(t) + j\omega(t) \tag{3.50}$$

a stała $a_0 e^{j\varphi_0}$ jest amplitudą zespoloną w chwili $t = 0$. Alternatywnie definicję ICF można wyprowadzić definiując najpierw zespoloną fazę chwilową jako

$$p(t) = \ln u(t) = \ln a(t) + j\varphi(t) \tag{3.51}$$

Jak widać, część rzeczywista zespolonej fazy chwilowej stanowi logarytm amplitudy chwilowej (nazywany poziomem chwilowym lub logobwiednią sygnału i wyrażany w

neperach [Np]), natomiast część urojona to faza chwilowa $\varphi(t) = \arg\big(u(t)\big)$. ICF można określić analogicznie do definicji IF jako pochodną po czasie zespolonej fazy chwilowej

$$s(t) = \frac{d}{dt}\ln[u(t)] = \frac{a'(t)}{a(t)} + j\varphi'(t) = \sigma(t) + j\omega(t) \tag{3.52}$$

To wyprowadzenie daje więc taki sam wynik jak w (3.50).

Jak można zauważyć, ICF jest uzależnioną od czasu pulsacją zespoloną $s = \sigma + j\omega$. Część urojona ICF jest pulsacją chwilową sygnału, natomiast część rzeczywista – względną chwilową prędkością promieniową, zdefiniowaną przez (3.47).

3.5.1. Interpretacja zespolonej pulsacji chwilowej

W p. 3.2.3 pokazaliśmy, że dla interpretacji wartości chwilowych sygnału zespolonego wygodnie jest przedstawić go w postaci wskazu wirującego na płaszczyźnie Gaussa. Obecnie powrócimy do tej reprezentacji, by omówić właściwości zespolonej pulsacji chwilowej sygnałów AM·FM [ŚW10]. Zaczniemy jednak od najprostszego sygnału zespolonego, czyli czystej kompleksoidy o stałej amplitudzie a_0 i częstotliwości ω_0 (brak modulacji). W tym przykładzie sygnał zespolony jest wskazem o stałej długości równej a_0, obracającym się ze stała prędkością ω_0 (w kierunku dodatnim czyli przeciwnym do ruchu wskazówek zegara, gdy $\omega_0 > 0$). Obrazem ICF takiego sygnału na płaszczyźnie zespolonej ($\sigma = \mathrm{Re}(s), \omega = \mathrm{Im}(s)$) jest punkt nie zmieniający swojego położenia. Współrzędna rzeczywista tego punktu jest równa zeru, co wskazuje na brak modulacji amplitudy AM (długość wskazu nie zmienia się w czasie). Współrzędna urojona ICF wynosi ω_0, czyli jest to częstotliwość sinusoidy zespolonej. Niezmienność części urojonej ICF wskazuje na brak modulacji częstotliwości FM.

Jeżeli wprowadzimy modulację amplitudy kompleksoidy, to wskaz wirujący na płaszczyźnie Gaussa ma nadal stałą (nie zmieniającą się) prędkość kątową ω_0, natomiast jego długość zmienia się. Zespolona pulsacja chwilowa takiego sygnału ma stałą (nie zmieniającą się) część urojoną równą ω_0 (brak modulacji FM). Modulacja AM, a dokładniej względna prędkość zmian amplitudy, jest odzwierciedlona w części rzeczywistej ICF. Gdy amplituda

chwilowa (długość wskazu) rośnie, wtedy $\sigma(t) > 0$, a gdy maleje to $\sigma(t) < 0$. Mówimy zatem, że $\sigma(t)$ jest, odpowiednio, chwilowym logarytmicznym inkrementem wzmocnienia lub dekrementem tłumienia sygnału zespolonego [ŚW10].

Jako ostatni przykład rozważymy sygnał o stałej amplitudzie chwilowej i modulowanej częstotliwości. Teraz wskaz na płaszczyźnie zespolonej ma stałą długość, natomiast zmienia się prędkość jego obrotu. Część rzeczywista ICF sygnału FM jest równa zeru w każdej chwili czasu (brak modulacji AM). Część urojona ICF pokazuje zmiany prędkości obrotowej wskazu. Gdy wskaz obraca się szybciej, to $\omega(t)$ rośnie, a gdy wolniej, to $\omega(t)$ maleje. Dodatkowo, wartość $\omega(t)$, dodatnia albo ujemna, wskazuje na kierunek obrotu wskazu na płaszczyźnie Gaussa w kierunku, odpowiednio, przeciwnym lub zgodnym z ruchem wskazówek zegara. Powyższe rozważania, szerzej przedstawione w [ŚW10], a także w [SU06], pozwalają na interpretację ICF jako indykatora modulacji AM i FM (ściśle, część rzeczywista ICF jest indykatorem modulacji AM, a urojona – FM).

Ponadto Hahn, jak również później Cohen, powiązali $|\sigma(t)|$ z chwilową szerokością pasma sygnału IB, wyrażoną w Np/s. O definicji pasma chwilowego według Cohena [CO89] [CO95] pisaliśmy już w podrozdz. 3.3. Cohen argumentuje, że IB, która charakteryzuje rozrzut częstotliwości wokół częstotliwości średniej w danej chwili czasu, może być utożsamiana z warunkowym odchyleniem standardowym $\sigma_{\omega|t}(t)$ rozkładu czas-częstotliwość (np. spektrogramu). Pokazuje on też, że dla większości rozkładów czasowo-częstotliwościowych warunkowe odchylenie standardowe wynosi (por.(3.47) i (3.50))

$$\sigma_{\omega|t}(t) = \left|\frac{a'(t)}{a(t)}\right| \tag{3.53}$$

a więc równe jest modułowi części rzeczywistej ICF. ICF przenosi więc informację nie tylko o częstotliwości chwilowej sygnału, ale również o chwilowej szerokości jego pasma. Warto również dodać, że ICF stanowi pełną reprezentację sygnału zespolonego, tzn. dysponując jej przebiegiem możemy odtworzyć reprezentowany przez nią sygnał zespolony (z dokładnością do fazy początkowej).

4. Bifaktoryzacja Voelckera-Kumaresana

W rozdziale tym opisana zostanie faktoryzacja sygnału analitycznego na obwiednię minimalnofazową i fazor dodatnioskrętny –sposób reprezentacji sygnałów analitycznych alternatywny do faktoryzacji AM·FM. Wywodzi się ona z teorii faktoryzacji w dziedzinie częstotliwości, czyli z reprezentacji filtru przyczynowego za pomocą kaskady filtrów minimalnofazowego i wszechprzepustowego [OP89], i z prac Voelckera [VO66a][VO66b]. Voelcker jako pierwszy zaproponował, by sygnały zespolone (a więc również analityczne) modelować jako wielomiany lub ilorazy wielomianów, tak jak modeluje się charakterystyki systemów za pomocą transmitancji operatorowej. Dzięki temu wiele teorii opracowanych w odniesieniu do systemów może być wykorzystanych dla analizy sygnałów. Hermanowicz i Rojewski [HE91] badali właściwości obwiedni minimalnofazowej i możliwości jej zastosowania w przetwarzaniu sygnałów pasmowych. Zastosowanie faktoryzacji na obwiednię minimalnofazową i fazor dodatnioskrętny w przetwarzaniu sygnału mowy jako pierwsi zaproponowali Kumaresan i in. [KU99], wykorzystując prace Voelckera [VO66a] [VO66b], stąd będziemy nazywać ją dalej bifaktoryzacją Voelckera-Kumaresana (V-K). Kumaresan i in. zaproponowali ponadto, by, zamiast pary $(a(t),\ \omega(t))$, sygnał analityczny reprezentowały: jego obwiednia, czyli amplituda chwilowa, oraz pulsacja chwilowa fazora dodatnioskrętnego. Zalety takiej reprezentacji sygnału analitycznego przedstawione zostaną w dalszej części rozdziału. W pierwszej części krótko opiszemy właściwości układów i sygnałów minimalnofazowych, maksymalnofazowych i mieszanofazowych oraz przedstawimy teorię faktoryzacji w dziedzinie częstotliwości.

4.1. Minimalnofazowość, maksymalnofazowość i mieszanofazowość

W dalszej części pracy często korzystać będziemy z pojęć minimalnofazowości, maksymalnofazowości i mieszanofazowości, warto zatem krótko przypomnieć, co oznaczają te pojęcia w odniesieniu do układów i sygnałów.

Mając dany stabilny i przyczynowy układ o odpowiedzi impulsowej $h(t)$, charakterystyce częstotliwościowej $H(\Omega)=|H(\Omega)|\exp(j\arg[H(\Omega)])$ oraz transmitancji

operatorowej $H(s)$ mówimy, że jest on minimalnofazowy, gdy logarytm jego charakterystyki amplitudowej $|H(\Omega)|$ oraz jego charakterystyka fazowa $\arg[H(\Omega)]$ tworzą parę transformat Hilberta [OP89]:

$$\{\arg[H(\Omega)]\} = H_T\{\ln|H(\Omega)|\} \tag{4.1}$$

lub, co jest równoznaczne

$$\{\arg[H(s)]\} = H_T\{\ln|H(s)|\} \tag{4.2}$$

Powyższe wymaganie jest często nazywane warunkiem minimalnofazowości. Inaczej można sformułować ten warunek zastrzegając, by zera i bieguny transmitancji $H(s)$ znajdowały się w lewej półpłaszczyźnie zmiennej zespolonej *s*, co jest równoznaczne stwierdzeniu, że istnieje przyczynowy i stabilny układ odwrotny o transmitancji $H^{-1}(s)$, takiej że $H^{-1}(s)H(s)=1$ [OP89].

Rozpatrzmy teraz klasę systemów, które mają takie same charakterystyki amplitudowe $|H(\Omega)|$, ale różnią się charakterystykami fazowymi. Do klasy tych układów należy układ minimalnofazowy, którego charakterystykę częstotliwościową oznaczać będziemy od teraz jako $H_{mp}(\Omega)$, transmitancję jako $H_{mp}(s)$, a odpowiedź impulsową jako $h_{mp}(t)$ (mp – ang. *minimum-phase*). Wiemy już, że wszystkie zera i bieguny transmitancji $H_{mp}(s)$ leżą w lewej półpłaszczyźnie zmiennej zespolonej *s*. Transmitancje pozostałych układów w klasie uzyskuje się poprzez przeniesienie zer transmitancji układu minimalnofazowego na prawą półpłaszczyznę (poprzez zmianę znaku ich części rzeczywistej) we wszystkich możliwych kombinacjach. Przeniesienie zera jest równoznaczne ze zwiększaniem co do wartości bezwzględnej ujemnych wartości fazy, nazywanych opóźnieniem fazowym [OP98], czyli im więcej zer znajduje się w prawej półpłaszczyźnie tym większe opóźnienie fazowe wprowadza dany układ. A zatem układ minimalnofazowy wprowadza najmniejsze opóźnienie fazowe spośród wszystkich układów o takiej samej charakterystyce amplitudowej (moduł jego charakterystyki fazowej jest najmniejszy dla wszystkich pulsacji Ω). Stąd nazywa się go

również minimalnie opóźniającym, choć nazwa minimalnofazowy bardziej się upowszechniła. Ponadto spośród odpowiedzi impulsowych układów z omawianej klasy, $h_{mp}(t)$ ma energię skupioną w przedziale czasu najbliższym zeru (bo $h_{mp}(t)$ jest tu najmniej opóźniona).

W omawianej klasie można wyróżnić również układ, dla którego wszystkie zera transmitancji operatorowej leżą w prawej półpłaszczyźnie zmiennej *s*, a zatem wprowadza on największe opóźnienie fazowe i nazywany jest maksymalnie opóźniającym lub, częściej, maksymalnofazowym. Pozostałe układy z tej klasy są mieszanofazowe. Oczywiście, jeśli rozważamy systemy cyfrowe zamiast transmitancji operatorowej używamy transformaty Z odpowiedziu impulsowej $h_{mp}[n]$. Warunkiem minimalnofazowości jest wówczas, by zera transmitancji $H_{mp}(z)$ znajdowały się wewnątrz okręgu jednostkowego.

Oppenheim i Schaffer [OP89] definiują nie tylko system, ale również sygnał minimalnofazowy, czyli taki, którego transformata Z jest minimalnofazowa. My jednak będziemy mówić o minimalnofazowości sygnałów tak, jak robił to Voelcker [VO66a]. Zaproponował on mianowicie, by sygnał analityczny $u(t) = a(t)\exp(j\varphi(t))$ modelować za pomocą wielomianów lub ilorazów wielomianów (czyli tak, jak zapisuje się transmitancje operatorowe układów). W tym celu Voelcker [VO66a] definiuje sygnał $u(z)$ jako

$$u(z) = \int_0^\infty U(\Omega)\exp(j\Omega z)d\Omega \tag{4.3}$$

gdzie $U(\Omega)$ jest charakterystyką częstotliwościową sygnału analitycznego $u(t)$, a z jest zmienną zespoloną $z = r + j\phi$. Reprezentacja $u(z)$ pozwala znaleźć zera i bieguny analizowanego sygnału, tak jak znajduje się zera i bieguny transmitancji operatorowej. Pozwala również skorzystać z zasad dualizmu czasowo-częstotliwościowego, z których wynika m.in., że

1) zespolona transmitancja operatorowa systemu odpowiada zespolonemu sygnałowi okresowemu,
2) przyczynowa odpowiedź impulsowa systemu odpowiada widmu sygnału równemu zeru dla częstotliwości ujemnych (sygnał analityczny),

3) moduł transmitancji operatorowej układu odpowiada amplitudzie chwilowej sygnału, argument transmitancji – fazie chwilowej sygnału, a opóźnienie grupowe – pulsacji chwilowej.

Dalej Voelcker opisuje również klasę sygnałów, które mają taką samą obwiednię amplitudową. Pokazuje, że poprzez zmianę znaku części urojonej jednego lub więcej zer sygnału $u(z)$ (czyli poprzez przesuwanie ich z dolnej na górną półpłaszczyznę zmiennej z lub odwrotnie) można uzyskać nowy sygnał o tej samej obwiedni amplitudowej. Wszystkie możliwe kombinacje takiej manipulacji zerami pozwalają obliczyć wszystkie sygnały z omawianej klasy. Jednym z sygnałów w tej klasie jest sygnał minimalnofazowy, który oznaczymy poprzez $u_{mp}(t)$ i $u_{mp}(z)$. Wszystkie jego zera leżą w dolnej półpłaszczyźnie zmiennej zespolonej z. Drugim sygnałem, który można wyróżnić, jest sygnał maksymalnofazowy. Jego zera leżą w górnej półpłaszczyźnie zmiennej z. Pozostałe sygnały w tej klasie są mieszanofazowe. Jeśli zmienną z zastąpimy zmienną $\alpha = \exp(-j\Omega_0 z)$ (gdzie $\Omega_0 = 2\pi / T_0$, a T_0 jest okresem sygnału $u(t)$) to sygnał minimalnofazowy będzie tym, dla którego wszystkie α_i opowiadające zerom sygnału z_i znajdują się wewnątrz okręgu jednostkowego na płaszczyźnie zmiennej zespolonej α. Voelcker [VO66a] wykazał również, że warunkiem minimalnofazowości sygnału analitycznego $u(t) = a(t)\exp(j\varphi(t))$ jest

$$\{\varphi(t)\} = H_T\{\ln|a(t)|\} \tag{4.4}$$

gdzie $a(t) = |u(t)|$ oraz $\varphi(t) = \arg[u(t)]$. Warunek ten, zgodnie z zasadami dualizmu, jest analogiczny do warunku sformułowanego dla układów minimalnofazowych i przytoczonego w (4.1). Z zasad dualizmu możemy wywnioskować również, że sygnał minimalnofazowy ma najmniejszą spośród wszystkich częstotliwość chwilową (odpowiada to najmniejszemu opóźnieniu grupowemu układu minimalnofazowego) oraz, że energia jego widma skupiona jest wokół elementów o najniższej częstotliwości (tak jak energia odpowiedzi impulsowej systemu minimalnofazowego skupiona jest w przedziale czasu najbliższym zeru).

4.2. Filtr przyczynowy jako kaskada filtrów minimalnofazowego i wszechprzepustowego

Dowolny przyczynowy filtr cyfrowy można przedstawić w postaci kaskady dwóch filtrów, również przyczynowych [OP89], z których jeden jest filtrem minimalnofazowym, tzn. takim, którego wszystkie zera i bieguny transmitancji znajdują się wewnątrz okręgu jednostkowego, a drugi filtrem wszechprzepustowym, tzn. takim, którego charakterystyka amplitudowa jest stała (domyślnie równa 1) dla wszystkich częstotliwości. W dziedzinie częstotliwości można to zapisać za pomocą zależności

$$H(\Omega) = H_{mp}(\Omega)H_{ap}(\Omega) \tag{4.5}$$

gdzie $H(\Omega)$ jest charakterystyką częstotliwościową dowolnego filtru przyczynowego, $H_{mp}(\Omega)$ charakterystyką częstotliwościową filtru minimalnofazowego, a $H_{ap}(\Omega)$ – filtru wszechprzepustowego (*ap* od ang. *all-pass*). Dowód powyższego twierdzenia dla filtrów cyfrowych podają Oppenheim i Schafer w [OP89]. Charakterystyka amplitudowa filtru minimalnofazowego równa jest charakterystyce amplitudowej wyjściowego filtru

$$\left|H(\Omega)\right| = \left|H_{mp}(\Omega)\right| \tag{4.6}$$

Filtr minimalnofazowy charakteryzuje się ponadto tym, że logarytm jego charakterystyki amplitudowej oraz jego charakterystyka fazowa stanowią parę transformat Hilberta, o czym była mowa w podrozdz. 4.1. Znamienne dla filtru wszechprzepustowego jest, oprócz stałej charakterystyki amplitudowej, zawsze dodatnie opóźnienie grupowe.

4.3. Bifaktoryzacja V-K sygnału analitycznego

Zarys teorii przedstawiony w podrozdz. 4.1 i 4.2 posłuży nam do omówienia bifaktoryzacji V-K sygnału analitycznego o ograniczonym paśmie [OP89]. W tym celu skorzystamy z dualizmu czasowo-częstotliwościowego, zamieniając ze sobą dziedziny czasu i częstotliwości. Wybrane zasady dualizmu wymieniliśmy już w podrozdz. 4.1. Zgodnie z tymi

zasadami możemy, analogicznie do filtru przyczynowego, przedstawionego w postaci kaskady dwóch filtrów, przeprowadzić faktoryzację sygnału analitycznego $u(t)$, o widmie $U(\Omega)$, na dwa czynniki: czynnik minimalnofazowy $a_{mp}(t)$, zwany obwiednią minimalnofazową (MPE – ang. *Minimum-Phase Envelope*), o widmie $A_{mp}(\Omega)$ oraz fazor dodatnioskrętny $\gamma_{pif}(t)$ o widmie $\Gamma_{pif}(\Omega)$

$$u(t) = a_{mp}(t)\gamma_{pif}(t) \tag{4.7}$$

Powyższy rozkład nazywać będziemy bifaktoryzacją V-K. Poprzez analogię do filtrów minimalnofazowego i wszechprzepustowego, które są przyczynowe, wiemy, że oba czynniki, $a_{mp}(t)$ i $\gamma_{pif}(t)$, są analityczne, zatem zawsze spełnione jest

$$U(\Omega) = A_{mp}(\Omega) * \Gamma_{pif}(\Omega) \tag{4.8}$$

niezależnie od dolnej pulsacji granicznej widma fazora dodatnioskrętnego, gdyż, zgodnie z zasadami dualizmu, tak jak wynikiem splotu dwóch sygnałów przyczynowych jest sygnał przyczynowy, wynikiem splotu widm dwóch sygnałów analitycznych jest widmo sygnału analitycznego. Oba czynniki bifaktoryzacji można zapisać w postaci wykładniczej

$$a_{mp}(t) = a(t)\exp(j\varphi_{mp}(t)) \tag{4.9}$$

$$\gamma_{pif}(t) = \exp(j\varphi_{pif}(t)) \tag{4.10}$$

gdzie $\varphi_{mp}(t)$ i $\varphi_{pif}(t)$ są fazami chwilowymi sygnałów $a_{mp}(t)$ i $\gamma_{pif}(t)$, natomiast $a(t)$ jest amplitudą chwilową sygnału $a_{mp}(t)$, równą amplitudzie chwilowej sygnału $u(t)$ (analogicznie do (4.6))

$$\left|a_{mp}(t)\right| = \left|u(t)\right| = a(t) \tag{4.11}$$

Dodatkowo, wykorzystując wspomnianą analogię, można również stwierdzić, że faza chwilowa $\varphi_{mp}(t)$ sygnału $a_{mp}(t)$ jest transformatą Hilberta jego logobwiedni (logarytmu amplitudy chwilowej)

$$\varphi_{mp}(t) = H_T\{\ln(a(t)\} \tag{4.12}$$

Fazor dodatnioskrętny $\gamma_{pif}(t)$, analogicznie do charakterystyki amplitudowej filtru wszechprzepustowego, ma amplitudę chwilową równą 1 w każdej chwili czasu, a jego częstotliwość chwilowa, tak jak opóźnienie grupowe filtru, jest zawsze dodatnia (stąd skrót *pif* – ang. *positive instantaneous frequency*). Dlatego fazor dodatnioskrętny oznaczać będziemy skrótem PIFP (ang. *Positive Instantaneous Frequency Phasor*). Warto również zauważyć, że zespolona obwiednia minimalnofazowa, $a_{mp}(t)$, zależy wyłącznie od amplitudy chwilowej analizowanego sygnału $u(t)$ i reprezentowana jest całkowicie przez $a(t)$, natomiast cała informacja o fazorze dotanioskrętnym $\gamma_{pif}(t)$ zawarta jest w jego fazie chwilowej $\varphi_{pif}(t)$.

Dla obu czynników bifaktoryzacji V-K można wyznaczyć zespoloną pulsację chwilową

$$s_{mp}(t) = \sigma(t) + j\omega_{mp}(t) \tag{4.13}$$

$$s_{pif}(t) = j\omega_{pif}(t) \tag{4.14}$$

W powyższych wzorach $s_{mp}(t)$ i $s_{pif}(t)$ to zespolone pulsacje chwilowe obwiedni minimalnofazowej oraz fazora dodatnioskrętnego, $\omega_{mp}(t) = \varphi'_{mp}(t)$ i $\omega_{pif}(t) = \varphi'_{pif}(t)$ to ich pulsacje chwilowe, natomiast $\sigma(t) = a'(t)/a(t)$ jest względną chwilową prędkością promieniową.

4.4. Porównanie faktoryzacji V-K i faktoryzacji AM·FM

W p. 3.2.4 została przedstawiona reprezentacja AM·FM sygnałów, czyli reprezentacja za pomocą iloczynu dwóch czynników, AM oraz FM. Dla zespolonego sygnału $u(t)$ zapisujemy ją jak w (3.31), tj.

$$u(t) = a(t)\gamma(t) \tag{4.15}$$

gdzie $a(t)$ jest czynnikiem AM (obwiednią rzeczywistą), a $\gamma(t)$ jest fazorem FM jak poprzednio. Zasadniczą cechą różniącą bifaktoryzację AM·FM od bifaktoryzacji V-K (4.7) jest to, że oba czynniki bifaktoryzacji V-K są zawsze analityczne. Dla faktoryzacji AM·FM fazor $\gamma(t)$ bywa analityczny tylko wtedy, gdy $a(t)$ i $\gamma(t)$ spełniają założenie twierdzenia Bedrosiana [BE63] (które bardziej szczegółowo omówiliśmy w p. 3.2.4). Wtedy zapisujemy

$$u(t) = a(t)\gamma(t) = a(t)\exp(j\varphi(t)) \tag{4.16}$$

gdzie ponownie $\varphi(t)$ stanowi fazę chwilową sygnału $u(t)$. Zakładając, że spełniona jest prawa równość w (4.16) oraz pamiętając, że $\varphi_{mp}(t) = H_T\{\lambda(t)\}$, gdzie logobwiednia $\lambda(t) = \ln a(t)$ (por. 4.12), otrzymujemy

$$\begin{aligned} u(t) &= a(t)\exp(j\varphi(t)) = \exp(\lambda(t) + j\varphi(t)) = \\ &= \exp(\lambda(t) + j\varphi(t) + j\varphi_{mp}(t) - j\varphi_{mp}(t)) = \\ &= \underbrace{\exp(\lambda(t) + j\varphi_{mp}(t))}_{a_{mp}(t)}\underbrace{\exp(j\varphi(t) - j\varphi_{mp}(t))}_{\gamma_{pif}(t)} \end{aligned} \tag{4.17}$$

Ponadto pamiętamy, że

$$a_{mp}(t) = a(t)\exp(j\varphi_{mp}(t)) = \exp(\lambda(t) + j\varphi_{mp}(t)) \tag{4.18}$$

co, podstawione do (4.17), daje

$$\gamma_{pif}(t) = \exp(j(\varphi(t) - \varphi_{mp}(t))) \tag{4.19}$$

A zatem

$$\varphi_{pif}(t) = \varphi(t) - \varphi_{mp}(t) \tag{4.20}$$

oraz

$$\omega_{pif}(t) = \omega(t) - \omega_{mp}(t) \tag{4.21}$$

Dodatnia pulsacja chwilowa (PIF) fazora dodatnioskrętnego reprezentuje tę część pulsacji chwilowej $\omega(t)$ sygnału $u(t)$, która pozostaje po odjęciu z niej udziału pulsacji chwilowej obwiedni minimalnofazowej. Zatem para $a(t)$ i $\omega_{pif}(t)$ (nazwiemy ją reprezentacją AM·PIF) w pełni reprezentuje sygnał $u(t)$. Oznacza to, że sygnał $u(t)$ można odtworzyć z tych dwóch przebiegów. Tak jak reprezentacja AM·FM, czyli para $(a(t), \omega(t))$, rozkład AM·PIF spełnia trzy cytowane w p. 3.2.4 warunki Vakmana na częstotliwość i amplitudę chwilową, determinujące zasadność takiej reprezentacji sygnału od strony fizycznej [VA96]. Jednocześnie oba przebiegi, $a(t)$ i $\omega_{pif}(t)$, przyjmują zawsze wartości dodatnie.

4.5. Dyskretna implementacja analizatora AM·PIF

Teraz zajmiemy się dyskretną implementacją analizatora sygnałów, a wszczególności reprezentacji AM·PIF. W tym celu należy zdefiniować dyskretne odpowiedniki wprowadzonych wcześniej przebiegów analogowych. Zacznijmy od rzeczywistego dyskretnego ciągu $x[n]$, który powstaje przez spróbkowanie rzeczywistego sygnału analogowego $x(t)$ z szybkością próbkowania F_S [Sa/s], której odpowiada okres próbkowania $T_S = 1/F_S$ [s]

$$x[n] = x(t)|_{t=nT_S} = x(nT_S) \tag{4.22}$$

Analogicznie przez $u[n]$ oznaczymy dyskretny ciąg zespolony powstały przez spróbkowanie zespolonego sygnału $u(t)$

$$u[n] = u(t)\,|_{t=nT_S} = u(nT_S) \tag{4.23}$$

Ponadto, ponieważ $u(t)$ jest równoważnikiem analitycznym analogowego sygnału $x(t)$, to $u[n]$ będzie równoważnikiem hilbertowskim dyskretnego sygnału $x[n]$

$$u[n] = x[n] + jy[n] \tag{4.24}$$

gdzie

$$y[n] = x[n] * h_T[n] \tag{4.25}$$

a $h_T[n]$ jest odpowiedzią impulsową dyskretnego transformatora Hilberta [OP89]. Ciąg $u[n]$ można także zdefiniować inaczej, jako splot ciągu $x[n]$ z odpowiedzią impulsową $h_A[n]$ dyskretnego zespolonego („analitycznego”) filtru Hilberta

$$u[n] = x[n] * h_A[n] \tag{4.26}$$

Charakterystyka amplitudowo-fazowa idealnego zespolonego filtru Hilberta ma postać

$$H_A(\omega) = \begin{cases} 2, & 0 < \omega < \pi \\ 1, & \omega = 0 \\ 0, & -\pi < \omega < 0 \end{cases} \tag{4.27}$$

Ze względu na większą wygodę implementacji, w zaproponowanym dalej algorytmie sygnał hilbertowski, będący zespoloną reprezentacją sygnału rzeczywistego, otrzymujemy

korzystając z definicji (4.26) zamiast z (4.24) i (4.25). Następnie dla ciągu $u[n]$ definiujemy jego dyskretną amplitudę i fazę chwilową, odpowiednio, $a[n]$ i $\varphi[n]$

$$a[n] = |u[n]| \tag{4.28}$$

$$\varphi[n] = \arg(u[n]) \tag{4.29}$$

Zespolona pulsacja chwilowa $s[n]$ sygnału $u[n]$ jest tu przebiegiem uzyskanym przez próbkowanie analogowej $s(t)$

$$s[n] = s(t)\,|_{t=nT_S} = s(nT_S) \tag{4.30}$$

Część urojona przebiegu $s[n]$ jest pulsacją chwilową $\omega[n]$ sygnału $u[n]$, a więc pochodną dyskretnej fazy chwilowej $\varphi[n]$. Część rzeczywista $\sigma[n]$ jest unormowaną pochodną amplitudy chwilowej $a[n]$.

Dla celów implementacji posłużymy się estymatą ICF, zaproponowaną przez Rojewskiego [RO94]

$$s[n] = \mathrm{Ln}\left(\frac{u[n]}{u[n-1]}\right) \tag{4.31}$$

gdzie $\mathrm{Ln}(\ )$ to zespolony logarytm główny, zdefiniowany jako

$$\mathrm{Ln}(z) = \ln|z| + \mathrm{Arg}(z) \tag{4.32}$$

W (4.32) z jest liczbą zespoloną, $|z|$ jej modułem, a $\mathrm{Arg}(z)$ jej argumentem głównym (jego wartość należy do przedziału $[-\pi, \pi)$). Aby nie wprowadzać zbyt dużej liczby różnych oznaczeń w (4.31) nie wyróżniamy estymaty zespolonej pulsacji chwilowej osobnym symbolem, zwłaszcza, że przy odpowiednich warunkach nie różni się ona znacznie od zdefiniowanej w (4.30) dyskretnej ICF. Wybrana estymata (4.31) ma tę zaletę, że pozwala

obliczyć ICF bez konieczności rozwijania fazy lub pulsacji [RO94]. Jeśli sygnał $u[n]$ jest sporo nadpróbkowany, wtedy błędy estymacji części rzeczywistej ICF praktycznie nie wystąpią. Błędy estymacji części urojonej są tym mniejsze im bardziej nadpróbkowany jest sygnał $u[n]$, gdyż wtedy prawdopodobieństwo wykroczenia wartości „prawdziwej” pulsacji chwilowej poza przedział $[-\pi,\pi)$ jest bardzo małe. Analogicznie estymować będziemy zespolone pulsacje chwilowe $s_{mp}[n]$ i $s_{pif}[n]$ oraz pulsacje chwilowe $\omega_{mp}[n]$ i $\omega_{pif}[n]$ przebiegów $a_{mp}[n]$ i $\gamma_{pif}[n]$. Obwiednię minimalnofazową $a_{mp}[n]$ oraz fazor dodatnioskrętny $\gamma_{pif}[n]$ sygnału $u[n]$ obliczać będziemy wprost z definicji, a mianowicie

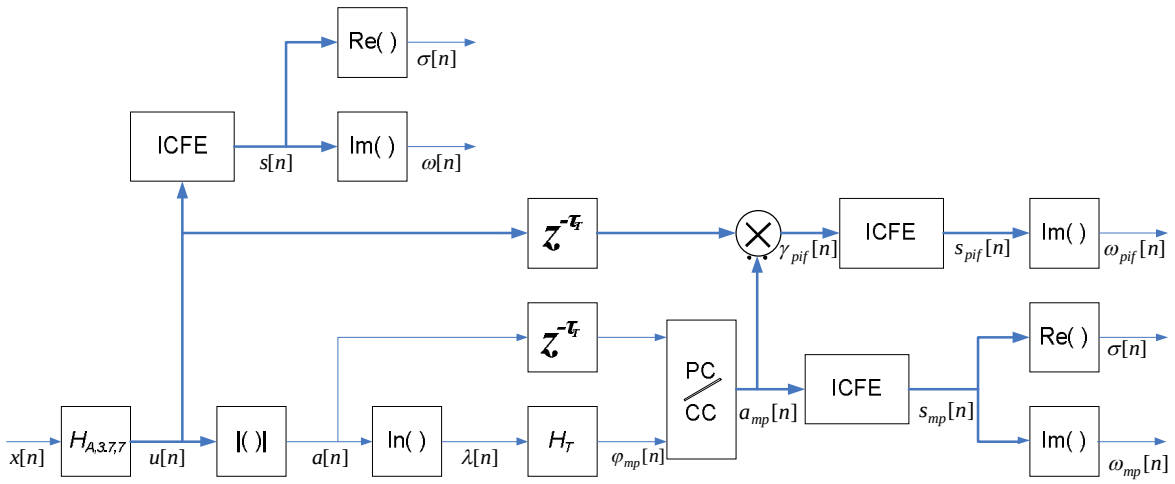
$$\left|a_{mp}[n]\right| = \left|u[n]\right| = a[n] \tag{4.33}$$

$$\varphi_{mp}[n] = \ln\big(a[n]\big) * h_T[n] \tag{4.34}$$

$$\gamma_{pif}[n] = \frac{u[n]}{a_{mp}[n]} \tag{4.35}$$

W (4.34) $h_T[n]$ jest odpowiedzią impulsową przyczynowego filtru FIR aproksymującego idealny transformator Hilberta.

Schemat implementacji zaproponowanego tu potokowego algorytmu przedstawia rys. 4.1.



Rys. 4.1. Schemat proponowanego potokowego analizatora sygnałów.

Na schemacie zastosowano oznaczenia:

- $H_{A,3.7,7}$ – cyfrowy zespolony filtr Hilberta o częstotliwości środkowej 3.7 kHz i szerokości pasma 7 kHz;
- H_T – aproksymator FIR idealnego transformatora Hilberta;
- $z^{-\tau_T}$ – element opóźniający, wyrównujący opóźnienie transformatora Hilberta;
- PC/CC – konwerter współrzędnych biegunowych na współrzędne kartezjańskie;
- ICFE – estymator zespolonej pulsacji chwilowej zgodnie z (4.33);
- Re( ) – część rzeczywista;
- Im( ) – część urojona.

Na wyjściu analizatora mamy do dyspozycji zespolone pulsacje chwilowe $s[n]$, $s_{mp}[n]$ oraz $s_{pif}[n]$ (ich części rzeczywiste i urojone). Można z niego wyprowadzić również przebiegi $a[n]$ lub $\lambda[n]$, $a_{mp}[n]$ i $\gamma_{pif}[n]$. Za jego pomocą możemy więc uzyskać nie tylko stosowane w tej pracy czynniki bifaktoryzacji V-K oraz ich ICF, ale także inne reprezentacje sygnału zespolonego (np. AM·FM, AM·PIF, reprezentację poprzez $s[n]$). Ponieważ w praktycznych systemach na wejściu mamy do czynienia z sygnałami rzeczywistymi, jak sygnał mowy, pierwszym ogniwem algorytmu jest zespolony filtr Hilberta, który pozwala uzyskać zespoloną reprezentację rzeczywistego sygnału $x[n]$. Wykorzystany filtr ma pasmo przenoszenia od 200 Hz do 7.2 kHz, dzięki czemu można ograniczyć pasmo mowy, zachowując jednocześnie wszystkie istotne formanty.

4.6. Testowanie algorytmu analizatora

Dla przetestowania działania algorytmu bifaktoryzacji V-K oraz określenia właściwości czynników tej faktoryzacji i ich zespolonych pulsacji chwilowych przeprowadzono serię symulacji na sygnałach syntetycznych. Do testów wybrano klasę ośmiu sygnałów okresowych, 4-tonowych, mających tę samą obwiednię $a[n]$, zadanych wzorem

$$u_4[n] = \sum_{l=1}^{4} a_l \exp[\, j(2\pi f_l n + \varphi_{0,l})] \tag{4.36}$$

Poszczególne składowe (składniki powyższej sumy) sygnałów w tej klasie mają te same częstotliwości f_l, różnią się natomiast amplitudami a_l oraz fazami początkowymi $\varphi_{0,l}$. Pierwszym rozważanym przez nas sygnałem z tej klasy jest sygnał minimalnofazowy, charakteryzujący się tym, że składowa o najniższej częstotliwości jest dominująca, natomiast jego widmo amplitudowe jest funkcją nierosnącą, tzn.

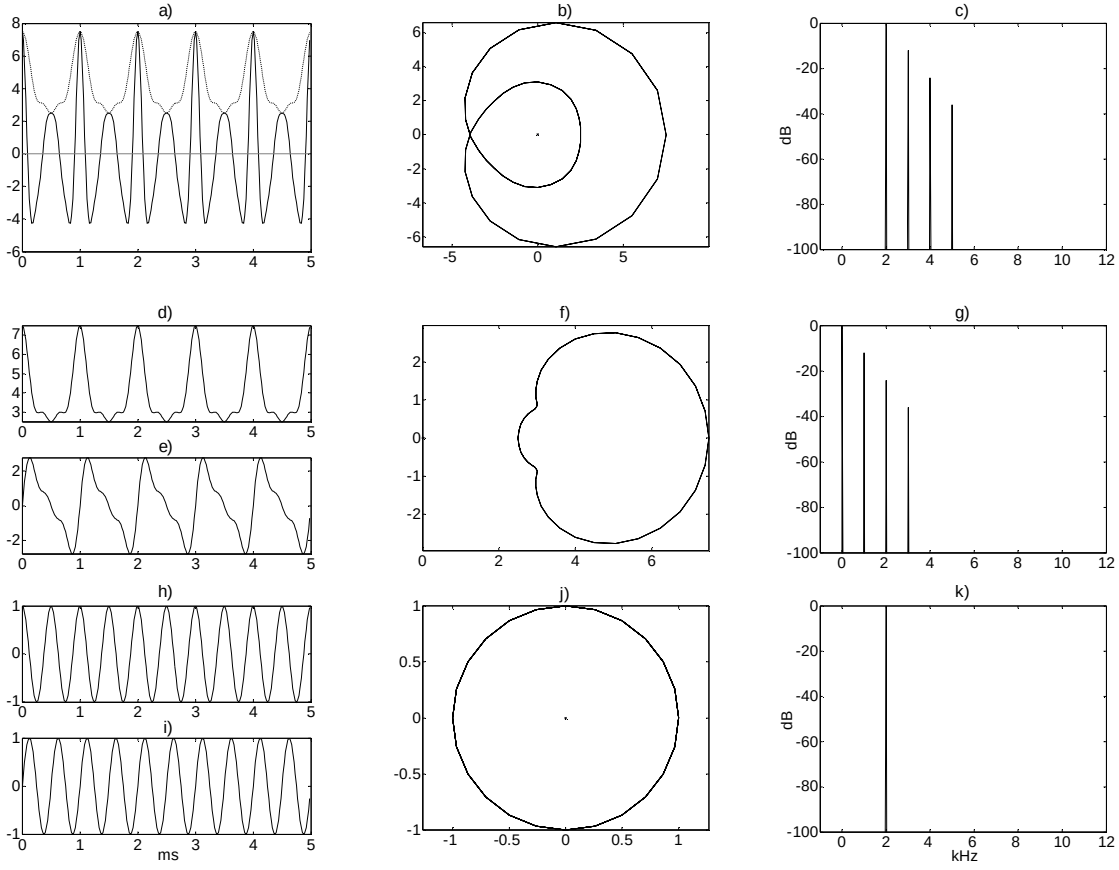
$$\begin{gathered} a_1 > a_2 \geq a_3 \geq a_4 > 0 \quad \text{oraz} \quad a_1 > a_2 + a_3 + a_4 \\ \text{dla} \quad 0 \leq f_1 < f_2 < f_3 < f_4 \end{gathered} \tag{4.37}$$

Jeśli tak jak Voelcker przedstawimy sygnał 4-tonowy w postaci wielomianu, będzie on miał trzy zera, z których dwa są ze sobą sprzężone. Dla sygnału minimalnofazowego wszystkie zera leżą wewnątrz okręgu jednostkowego: $z_1 = 0.5\exp(-j\pi)$, $z_2 = 0.5\exp(-j\pi/2)$, $z_3 = 0.5\exp(j\pi/2)$. Zmieniając położenie poszczególnych zer tak, by znajdowały się poza okręgiem jednostkowym (zmieniając ich amplitudę na odwrotną), otrzymuje się pozostałe siedem sygnałów z omawianej klasy. W szczególności, zmieniając położenia wszystkich zer otrzymujemy sygnał maksymalnofazowy, dla którego wszystkie zera leżą poza okręgiem jednostkowym, dominującym prążkiem jest prążek o najwyższej częstotliwości, a widmo jest niemalejące. Pozostałe sygnały są mieszanofazowe, o pewnym stopniu zawartości minimalnofazowości (0-100%). W dalszej części pracy pokażemy, jak w oparciu o przebiegi ICF obliczać stopień minimalnofazowości sygnałów. Rys. 4.2 – 4.11 przedstawiają wyniki faktoryzacji wszystkich ośmiu sygnałów z omawianej klasy, zaczynając od sygnału minimalnofazowego $u_{4,1}[n]$ (rys. 4.2 – 4.4). W opisie eksperymentów stosujemy oznaczenia analogiczne z wprowadzonymi w podrozdz. 4.4 z dodatkowymi indeksami dolnymi: 4, który odnosi się do sygnałów 4-tonowych oraz *i* (*i*=1,2,3,4,5,6,7,8), który określa kolejne sygnały z omawianej klasy. Wykresy na rys. 4.2 przedstawiają oscylogramy (lewa kolumna) sygnału $u_{4,1}[n]$, jego obwiedni minimalnofazowej i fazora dodatnioskrętnego, ich trajektorie zespolone (środkowa kolumna) oraz periodogramy (prawa kolumna). Na podstawie tych wykresów wnioskujemy, że dla sygnału minimalnofazowego:

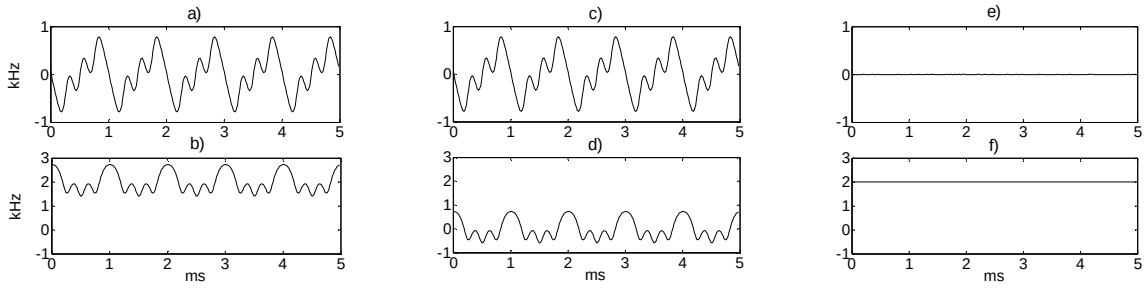
1) periodogram MPE jest taki sam jak periodogram sygnału $u_{4,1}[n]$ tylko przesunięty na osi częstotliwości (prążki znajdują się na częstotliwościach 0, 1000, 2000 i 3000 Hz);

2) PIFP jest zespoloną sinusoidą o częstotliwości równej częstotliwości dominującego prążka sygnału $u_{4,1}[n]$ (2000 Hz), jego amplituda jest stała i równa 1, co wynika z właściwości bifaktoryzacji V-K, opisanych w podrozdz. 4.2;
3) periodogram PIFP to pojedynczy prążek usytuowany na częstotliwości dominującego prążka sygnału $u_{4,1}[n]$ (co jest zgodne z poprzednim punktem), skąd wnioskujemy, że fazor dodatnioskrętny niesie informacje o położeniu widma sygnału.

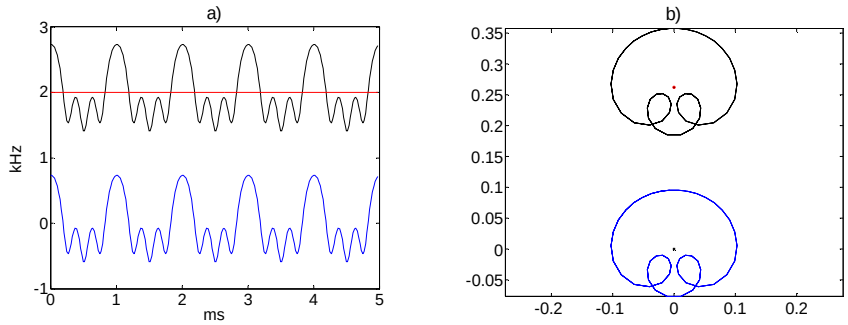
Zespolone pulsacje chwilowe obliczone dla obu czynników faktoryzacji, jak również dla sygnału $u_{4,1}[n]$, pokazano na rys. 4.3. Na rys. 4.4 zobaczyć można ich zestawienie dla dalszych badań. Wartości w hercach uzyskano mnożąc części rzeczywiste i urojone ICF przez współczynnik $F_S / 2\pi$. Na podstawie rys. 4.3 i 4.4 można stwierdzić, że:



Rys. 4.2. Wyniki bifaktoryzacji sygnału 4-tonowego $u_{4,1}[n]$: część rzeczywista sygnału $x_{4,1}[n]$ wraz z jego obwiednią $a_{4,1}[n]$ (a), trajektoria zespolona $u_{4,1}[n]$ (b), periodogram $u_{4,1}[n]$ (c), część rzeczywista (d) i urojona (e) MPE, trajektoria zespolona MPE (f), periodogram MPE (g), część rzeczywista (h) i urojona (i) PIFP, trajektoria zespolona PIFP (j), periodogram PIFP (k).



Rys. 4.3. Przebiegi części rzeczywistych i urojonych zespolonych częstotliwości chwilowych sygnału $u_{4,1}[n]$ (a,b), jego MPE (c,d) oraz PIFP (e,f).



Rys. 4.4. Porównanie przebiegów IF (a) oraz trajektorii ICF (b) sygnału $u_{4,1}[n]$ (kolor czarny), jego MPE (niebieski) i PIFP (czerwony).

1) części rzeczywiste ICF sygnału $u_{4,1}[n]$ oraz jego MPE są sobie równe, natomiast część rzeczywista ICF fazora dodatnioskrętnego jest równa zeru, co wynika oczywiście stąd, że jego amplituda jest stała i równa 1; ta prawidłowość zachodzi dla czynników bifaktoryzacji V-K dowolnego sygnału zespolonego i wynika z właściwości tej faktoryzacji;
2) część urojona ICF fazora dodatnioskrętnego, czyli jego IF jest stała i równa częstotliwości dominującego prążka sygnału $u_{4,1}[n]$ (2000 Hz), nie pokrywa się z częstotliwością środkową pasma sygnału ani też ze środkiem ciężkości widma, obliczonym jako moment unormowany I rzędu;
3) wartość średnia częstotliwości chwilowej sygnału $u_{4,1}[n]$ jest równa częstotliwości chwilowej i widmowej PIFP;
4) części urojone ICF sygnału $u_{4,1}[n]$ oraz jego MPE mają taki sam przebieg, z tym że dla obwiedni minimalnofazowej składowa stała IF równa jest zeru, natomiast dla sygnału $u_{4,1}[n]$ składowa stała IF równa jest IF fazora dodatnioskrętnego;

5) wartość IF żadnego z omawianych przebiegów nie wskazuje na częstotliwość podstawową sygnału $u_{4,1}[n]$. Można jednak zauważyć, że przebiegi $s_{4,1}[n]$ oraz $s_{mp4,1}[n]$ zachowują okresowość sygnału $u_{4,1}[n]$.

Kolejne rysunki przedstawiają wyniki bifaktoryzacji pozostałych siedmiu sygnałów z omawianej klasy. Oznaczymy je jako $u_{4,2}[n]$ (z_1 poza okręgiem jednostkowym), $u_{4,3}[n]$(z_2 poza okręgiem jednostkowym), $u_{4,4}[n]$(z_3 poza okręgiem jednostkowym), $u_{4,5}[n]$ (z_1 i z_2 poza okręgiem jednostkowym), $u_{4,6}[n]$ (z_1 i z_3 poza okręgiem jednostkowym), $u_{4,7}[n]$ (z_2 i z_3 poza okręgiem jednostkowym) oraz $u_{4,8}[n]$ (wszystkie zera poza okręgiem jednostkowym – sygnał maksymalnofazowy). Parametry składowych wszystkich ośmiu sygnałów z omawianej kasy zostały zebrane w tab. 4.1.

TAB. 4.1. PARAMETRY SKŁADOWYCH UŻYTYCH DO TESTÓW SYGNAŁÓW

	a_1	a_2	a_3	a_4	$\varphi_{0,1}$	$\varphi_{0,2}$	$\varphi_{0,3}$	$\varphi_{0,4}$
$u_{4,1}[n]$	4	2	1	0.5	0	0	0	0
$u_{4,2}[n]$	2	4	0.5	1	0	0	0	0
$u_{4,3}[n]$	2	$\sqrt{10}$	2.5	1	0	0.4π	0.2π	0
$u_{4,4}[n]$	2	$\sqrt{10}$	2.5	1	0	-0.4π	-0.2π	0
$u_{4,5}[n]$	1	2.5	$\sqrt{10}$	2	0	0.2π	0.4π	0
$u_{4,6}[n]$	1	2.5	$\sqrt{10}$	2	0	-0.2π	-0.4π	0
$u_{4,7}[n]$	1	0.5	4	2	0	0	0	0
$u_{4,8}[n]$	0.5	1	2	4	0	0	0	0

Porównując wykresy uzyskane dla tych sygnałów widzimy, że:

1) wszystkie przebiegi wyznaczone dla MPE sygnałów z omawianej klasy są identyczne, co wynika oczywiście stąd, że sygnały te mają takie same amplitudy chwilowe;
2) im więcej zer znajduje się poza okręgiem jednostkowym tym więcej razy trajektoria sygnału zespolonego obiega początek układu współrzędnych w czasie jednego okresu (dwa razy dla sygnału minimalnofazowego, trzy razy dla sygnałów z jednym zerem poza okręgiem jednostkowym, cztery dla sygnałów z dwoma zerami poza okręgiem jednostkowym i pięć dla sygnału maksymalnofazowego); uwidocznione jest to również w większej liczbie przejść przez zero przebiegów części rzeczywistej tych

sygnałów oraz w ich widmach, w których prążkiem dominującym jest prążek pierwszy (2000 Hz) dla sygnału $u_{4,1}[n]$, drugi (3000 Hz) dla sygnałów $u_{4,2}[n]$, $u_{4,3}[n]$ i $u_{4,4}[n]$, trzeci (4000 Hz) dla sygnałów $u_{4,5}[n]$, $u_{4,6}[n]$ i $u_{4,7}[n]$ oraz czwarty (5000 Hz) dla sygnału $u_{4,8}[n]$;

3) PIFP obliczone dla sygnałów od $u_{4,2}[n]$ do $u_{4,8}[n]$ nie są czystymi sinusoidami zespolonymi (o niemodulowanej amplitudzie i częstotliwości); w ich przebiegach można zauważyć modulację częstotliwości;
4) prążki dominujące PIFP pokrywają się z prążkami dominującymi sygnałów zespolonych;
5) przebiegi IF fazorów dodatnioskrętnych nie są stałe, ich zmiany wskazują na modulację częstotliwości fazorów;
6) wartości średnie IF fazorów dodatnioskrętnych są równe wartościom średnim IF sygnałów zespolonych i pokrywają się z częstotliwościami dominującego prążka w widmach tych przebiegów (nie z częstotliwością środkową pasma sygnału zespolonego ani też z środkiem ciężkości widma) – są tym wyższe im więcej zer znajduje się poza okręgiem jednostkowym;
7) części rzeczywiste i urojone przebiegów $s_{4,i}[n]$ oraz $s_{mp4,i}[n]$ zachowują okresowość sygnałów $u_{4,i}[n]$; ta prawidłowość nie zawsze jest obserwowana dla przebiegów $s_{pif\,4,i}[n]$, bowiem $\omega_{pif\,4,1}[n]$ jest stała;
8) trajektorie sygnałów zespolonych, w których oba zera sprzężone (z_2 i z_3) znajdują się wewnątrz okręgu jednostkowego (sygnały $u_{4,1}[n]$ i $u_{4,2}[n]$) lub poza tym okręgiem (sygnały $u_{4,7}[n]$ i $u_{4,8}[n]$) wskazują, że faza początkowa tych sygnałów jest równa zeru (fazy początkowe wszystkich ich składowych są równe zeru); gdy jedno z zer sprzężonych znajduje się wewnątrz okręgu jednostkowego, a drugie poza nim faza początkowa sygnału wynosi $+\varphi_0$ (dla sygnałów $u_{4,3}[n]$ i $u_{4,5}[n]$ – dodatnie fazy początkowe drugiej i trzeciej składowej) lub $-\varphi_0$ (dla sygnałów $u_{4,4}[n]$ i $u_{4,6}[n]$ - ujemne fazy początkowe drugiej i trzeciej składowej); ponadto dla par sygnałów $u_{4,1}[n]$ i $u_{4,8}[n]$, $u_{4,2}[n]$ i $u_{4,7}[n]$, $u_{4,3}[n]$ i $u_{4,6}[n]$ oraz $u_{4,4}[n]$ i $u_{4,5}[n]$ amplitudy

składowych jednego sygnału z pary przyjmują te same wartości co w drugim sygnale z pary, ale w odwrotnej kolejności, natomiast dla par sygnałów $u_{4,3}[n]$ i $u_{4,4}[n]$ oraz $u_{4,5}[n]$ i $u_{4,6}[n]$ amplitudy kolejnych składowych są takie same, ale fazy składowych drugiej i trzeciej są przeciwne względem siebie. Wynikające z tego dalsze spostrzeżenia są następujące:

a) periodogramy sygnałów $u_{4,3}[n]$ i $u_{4,4}[n]$ oraz $u_{4,5}[n]$ i $u_{4,6}[n]$ są takie same, różnią się natomiast od periodogramów sygnałów odpowiednio $u_{4,2}[n]$ i $u_{4,7}[n]$;

b) periodogramy fazorów $\gamma_{pif\,4,2}[n]$, $\gamma_{pif\,4,3}[n]$ i $\gamma_{pif\,4,4}[n]$ są takie same; dotyczy to również periodogramów przebiegów $\gamma_{pif\,4,5}[n]$ i $\gamma_{pif\,4,6}[n]$;

c) przebiegi $\gamma_{pif\,4,2}[n]$, $\gamma_{pif\,4,3}[n]$ i $\gamma_{pif\,4,4}[n]$ różnią się od siebie wyłącznie przesunięciem fazowym; dotyczy to również przebiegów $\gamma_{pif\,4,5}[n]$ i $\gamma_{pif\,4,6}[n]$;

d) podobnie przebiegi $\omega_{pif\,4,2}[n]$, $\omega_{pif\,4,3}[n]$ i $\omega_{pif\,4,4}[n]$ różnią się od siebie wyłącznie przesunięciem fazowym; dotyczy to również przebiegów $\omega_{pif\,4,5}[n]$ i $\omega_{pif\,4,6}[n]$;

e) trajektorie przebiegów $s_{4,3}[n]$ i $s_{4,4}[n]$ oraz $s_{4,5}[n]$ i $s_{4,6}[n]$ są symetryczne względem osi urojonej (sygnały z tych par mają odwrotny układ zer z pary sprzężonej – mają dokładnie przeciwne fazy początkowe dwóch środkowych prążków, natomiast amplitudy poszczególnych składowych są takie same);

f) trajektorie przebiegów $s_{4,1}[n]$ i $s_{4,8}[n]$, $s_{4,3}[n]$ i $s_{4,6}[n]$; $s_{4,4}[n]$ i $s_{4,5}[n]$ oraz $s_{4,2}[n]$ i $s_{4,7}[n]$ są symetryczne względem linii imaginaris=0.4581, co odpowiada 3500 Hz – dokładnie środek pasma zajmowanego przez sygnały (sygnały z tych par mają wzajemnie odwrotny układ zer, przebiegi ich IF są odwrócone, jeśli usunąć z nich składową stałą, która jest oczywiście inna dla każdego przebiegu);

9) poszczególne sygnały z omawianej klasy różnią się brzmieniem:

a) im wyższą częstotliwość ma prążek dominujący w widmie, tym wyższa jest percypowana wysokość dźwięku (choć częstotliwość podstawowa wszystkich sygnałów jest taka sama);

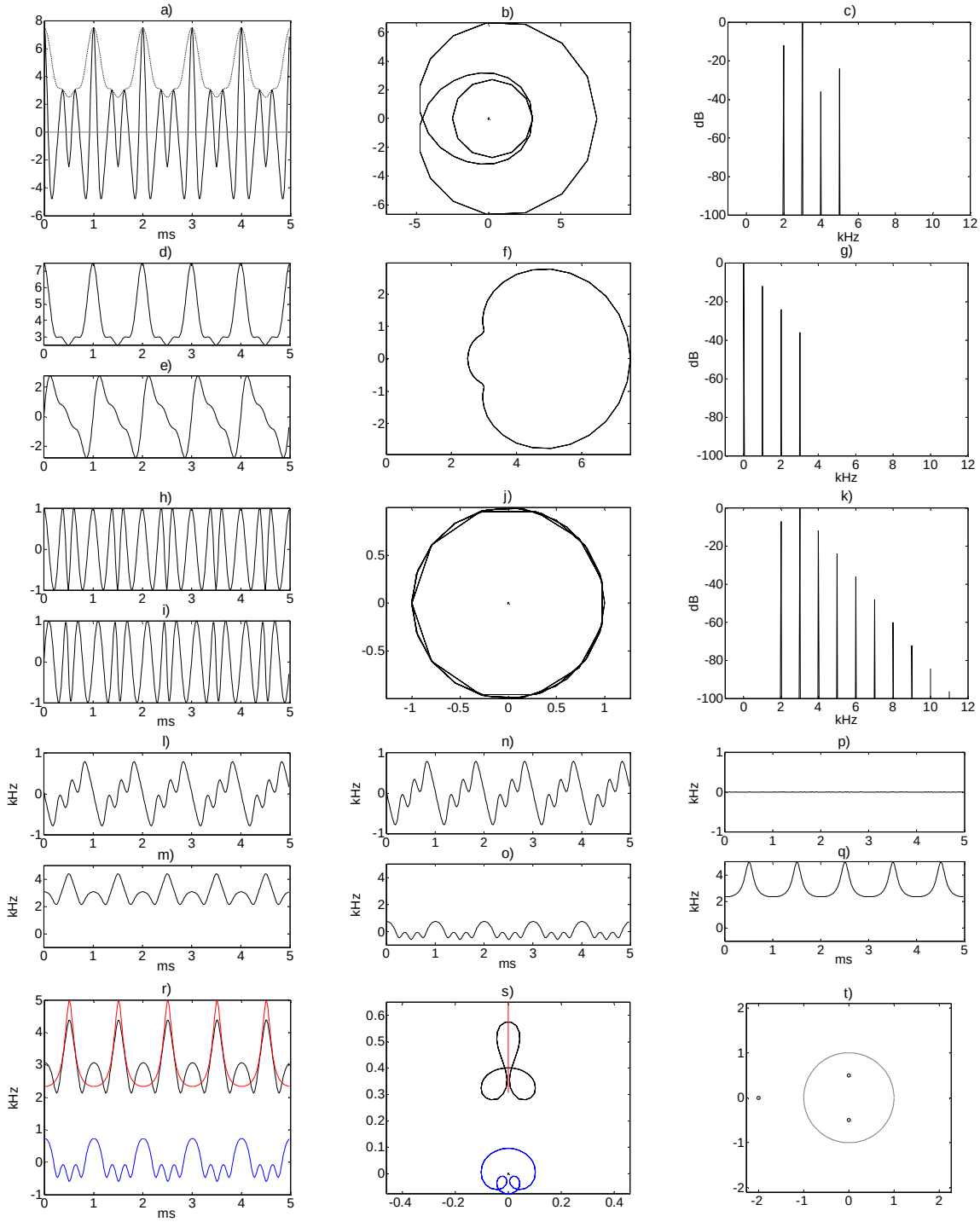
b) kolejne sygnały z omawianej klasy mają również coraz jaśniejszą barwę, co wynika z coraz większego udziału wyższych częstotliwości w widmie;

c) dźwięki $u_{4,3}[n]$ i $u_{4,4}[n]$ oraz $u_{4,5}[n]$ i $u_{4,6}[n]$ (które w parach różnią się wyłącznie fazami początkowymi drugiej i trzeciej składowej) w ogóle nie różnią się brzmieniem; jest to zgodne z przyjętym powszechnie założeniem, że różnice w fazach początkowych poszczególnych składowych sygnałów wielotonowych nie są percypowane (lub są bardzo słabo słyszalne) przez ucho ludzkie.

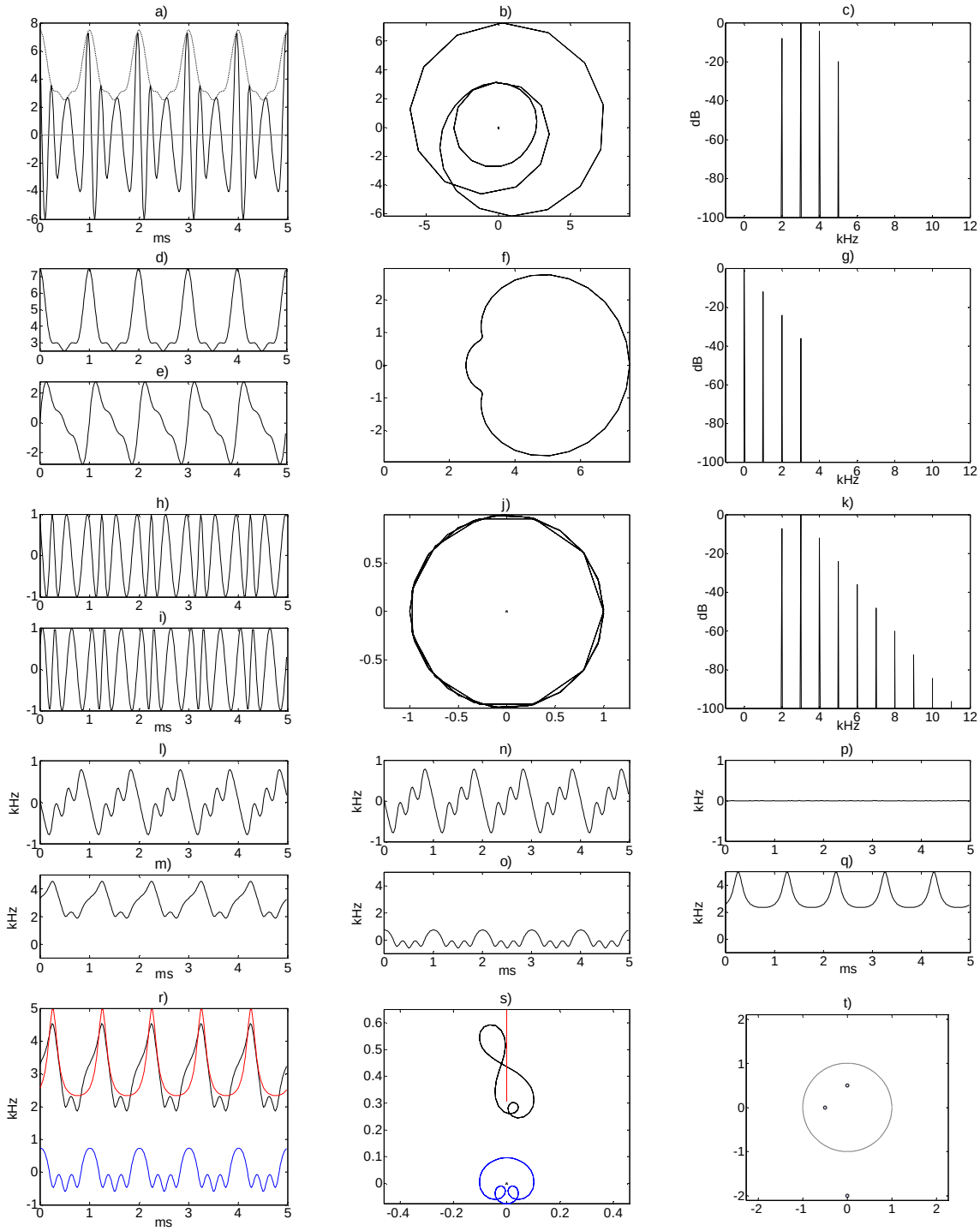
Warto w tym miejscu powrócić do definicji IF zaproponowanych przez Oliveirę i Barroso [OL00] oraz Loughlina i Tacera [LO96], o których pisaliśmy w podrozdz. 3.4. W obu wspomnianych definicjach sygnał analityczny zapisywany był w postaci (3.41), którą przytaczamy ponownie poniżej dla wygody Czytelnika

$$u(t) = a(t)e^{j\varphi_A(t)}e^{j\varphi_F(t)} \tag{4.38}$$

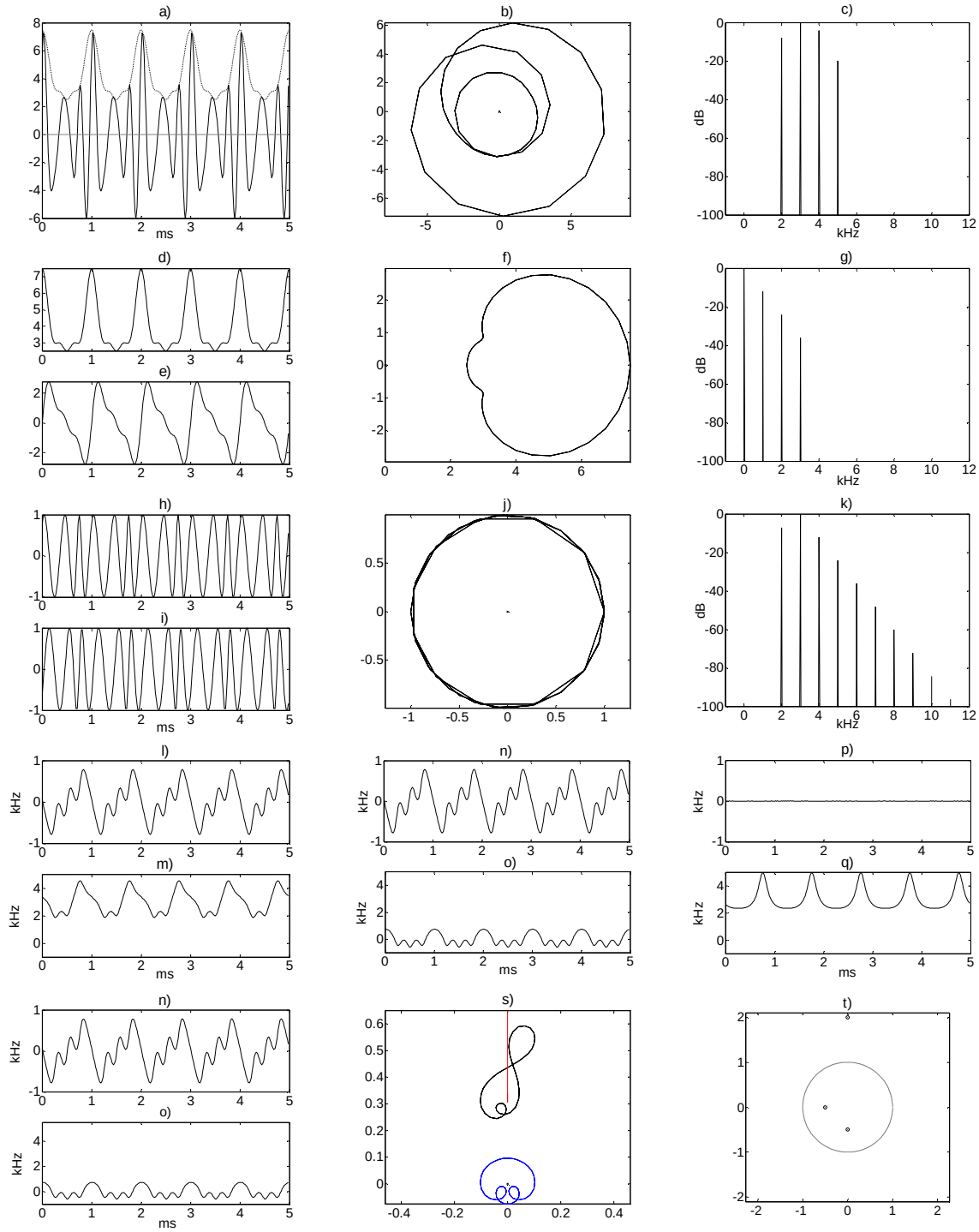
a IF było pochodną $\varphi_F(t)$. Według interpretacji Oliveiry i Barrosy [OL00], $u(t)$ jest wynikiem heterodynowania sygnału $a(t)e^{j\varphi_A(t)}$ do częstotliwości $\varphi'_F(t)$, a $\varphi_A(t)$ odpowiada za brak symetrii widma sygnału. Natomiast Loughlin i Tacer [LO96] argumentują, że $u(t)$ jest wynikiem modulacji częstotliwościowej sygnału $a(t)e^{j\varphi_A(t)}$ (gdzie $\varphi_A(t)$ może być interpretowane jako kwadraturowa modulacja amplitudy lub, alternatywnie, jako modulacja fazy), a IF jest średnią częstotliwością furierowską rozkładu czasowo-częstotliwościowego w każdej chwili czasu. Zauważmy, że w przypadku wykorzystania bifaktoryzacji V-K $u(t)$ zapisuje się jako $u(t) = a(t)e^{j\varphi_{mp}(t)}e^{j\varphi_{pif}(t)}$, a więc podobnie jak w (4.22), rozdziela się chwilową fazę sygnału na dwie składowe. W tym przypadku IF jest pochodną $\varphi_{pif}(t)$. Jak pokazały omawiane symulacje tak wyznaczona IF nie wskazuje na średnią częstotliwość rozkładu czasowo-częstotliwościowego (jak u Louglina i Tacera) ani na częsotliwość wokół której chwilowe widmo $u(t)$ wykazuje największą symetrię (jak u Oliveiry i Barosso), ale na prążek dominujący w widmie $u(t)$. Jak pokażemy dalej taka interpretacja ma uzasadnienie we właściwościach sygnału mowy, gdyż IF fazora dodatnioskrętnego wskazuje na częstotliwość środkową najwyższego formantu.



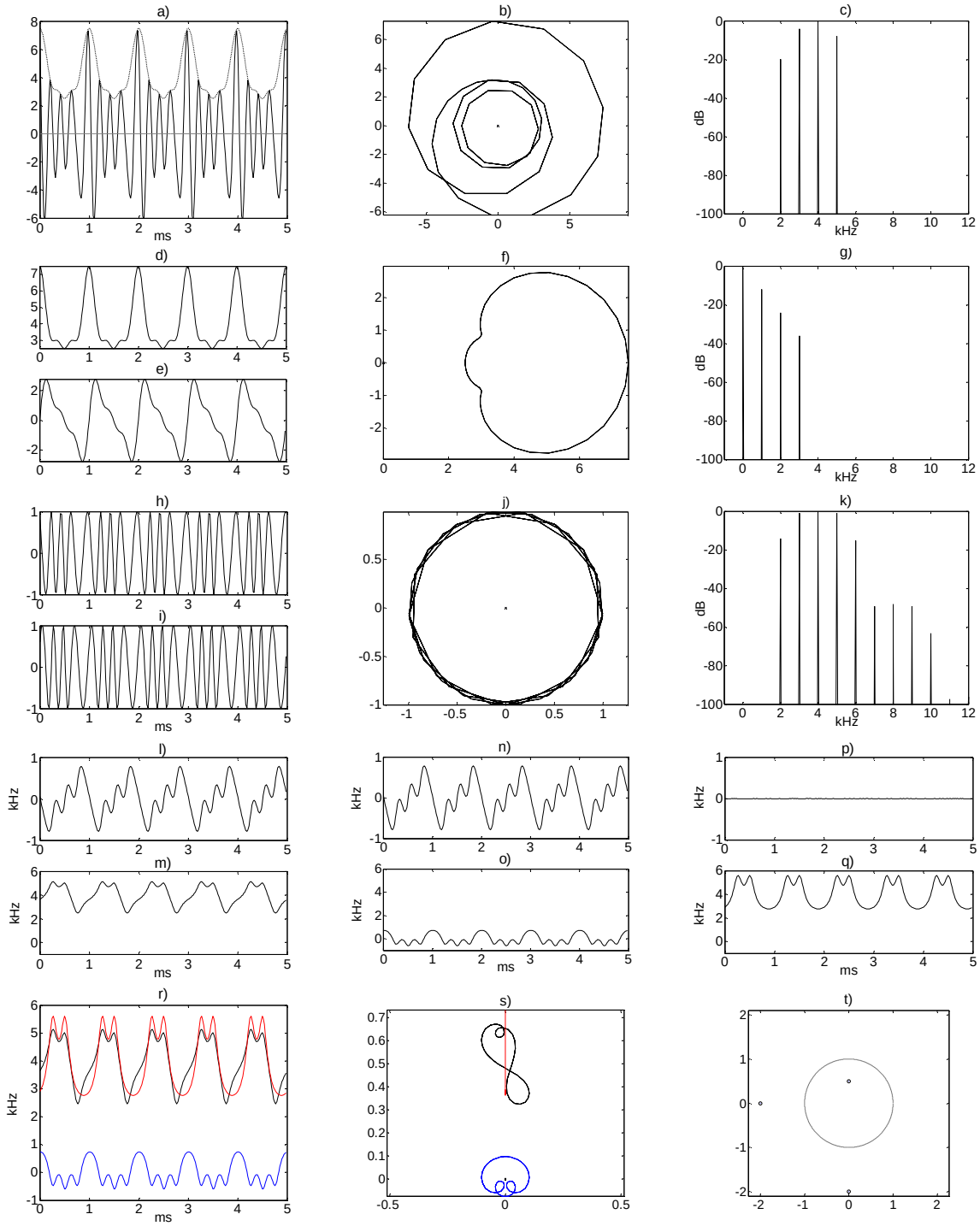
Rys. 4.5. Wyniki bifaktoryzacji sygnału 4-tonowego $u_{4,2}[n]$: część rzeczywista i obwiednia (linia przerywana) sygnału (a), trajektoria zespolona (b) i periodogram (c) sygnału, część rzeczywista (d) i urojona (e) MPE, trajektoria zespolona MPE (f) oraz periodogram MPE (g), część rzeczywista (h) i urojona (i) PIFP, trajektoria zespolona PIFP (j) oraz periodogram PIFP (k); przebiegi części rzeczywistych i urojonych ICF sygnału (l,m), jego obwiedni minimalnofazowej (n,o) oraz fazora dodatnioskrętnego (p,q); porównanie przebiegów IF (r) oraz trajektorii zespolonych pulsacji chwilowych (s) sygnału $u_{4,2}[n]$ (kolor czarny), jego obwiedni minimalnofazowej (niebieski) i fazora dodatnioskrętnego (czerwony); rozkład zer na płaszczyźnie zespolonej (t).



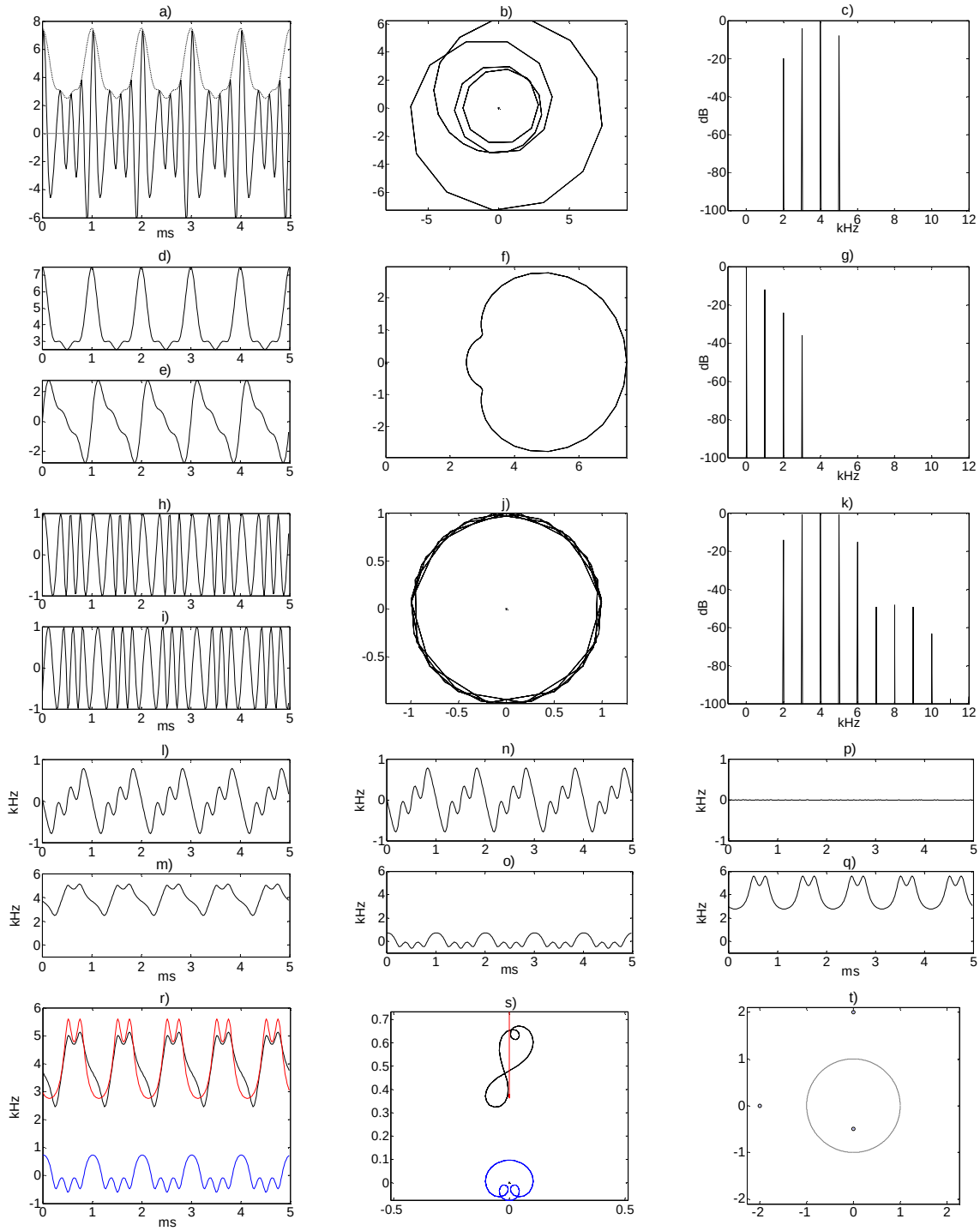
Rys. 4.6. Wyniki bifaktoryzacji sygnału 4-tonowego $u_{4,3}[n]$: część rzeczywista i obwiednia (linia przerywana) sygnału (a), trajektoria zespolona (b) i periodogram (c) sygnału, część rzeczywista (d) i urojona (e) MPE, trajektoria zespolona MPE (f) oraz periodogram MPE (g), część rzeczywista (h) i urojona (i) PIFP, trajektoria zespolona PIFP (j) oraz periodogram PIFP (k); przebiegi części rzeczywistych i urojonych ICF sygnału (l,m), jego obwiedni minimalnofazowej (n,o) oraz fazora dodatnioskrętnego (p,q); porównanie przebiegów IF (r) oraz trajektorii zespolonych pulsacji chwilowych (s) sygnału $u_{4,3}[n]$ (kolor czarny), jego obwiedni minimalnofazowej (niebieski) i fazora dodatnioskrętnego (czerwony); rozkład zer na płaszczyźnie zespolonej (t).



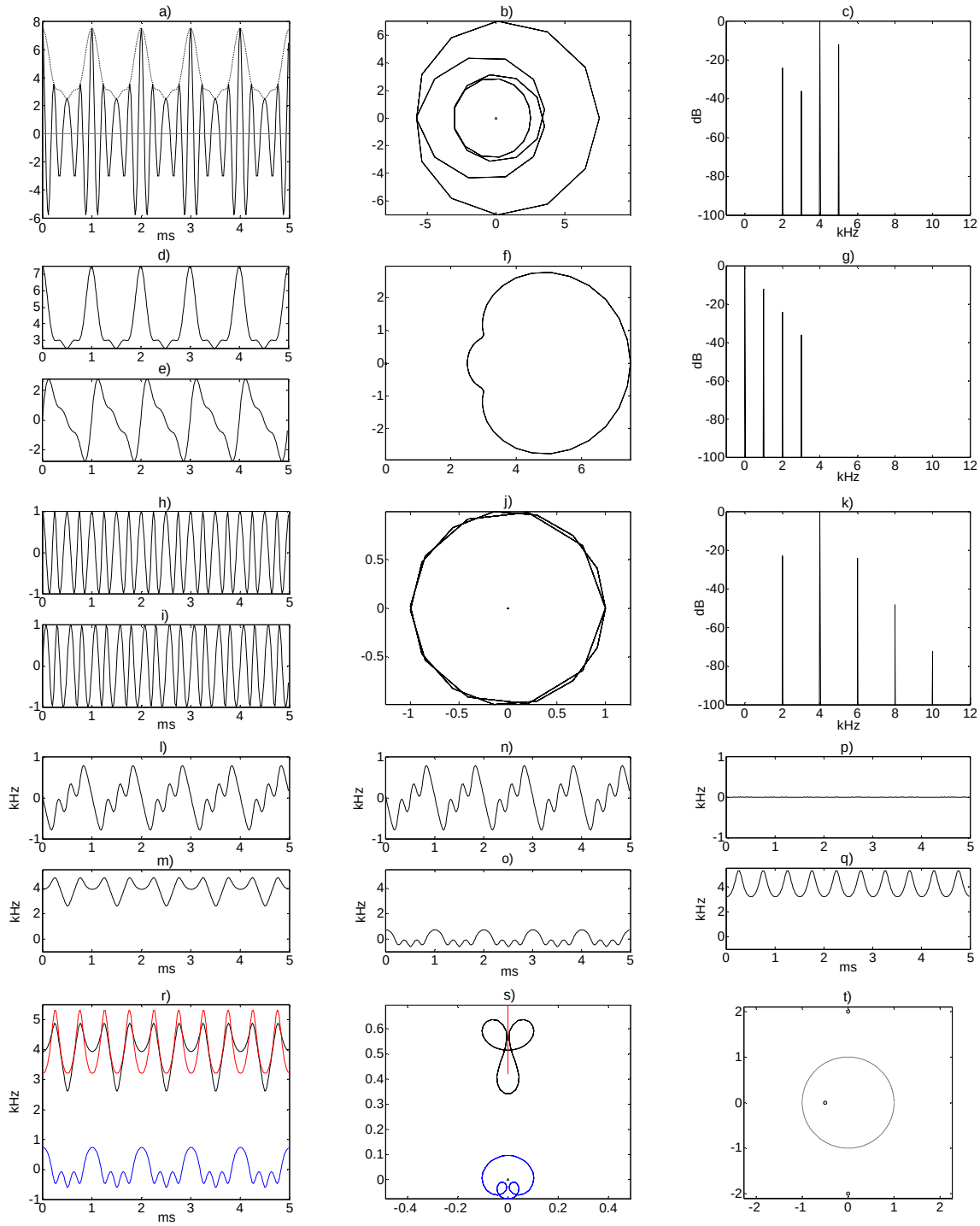
Rys. 4.7. Wyniki bifaktoryzacji sygnału 4-tonowego $u_{4,4}[n]$: część rzeczywista i obwiednia (linia przerywana) sygnału (a), trajektoria zespolona (b) i periodogram (c) sygnału, część rzeczywista (d) i urojona (e) MPE, trajektoria zespolona MPE (f) oraz periodogram MPE (g), część rzeczywista (h) i urojona (i) PIFP, trajektoria zespolona PIFP (j) oraz periodogram PIFP (k); przebiegi części rzeczywistych i urojonych ICF sygnału (l,m), jego obwiedni minimalnofazowej (n,o) oraz fazora dodatnioskrętnego (p,q); porównanie przebiegów IF (r) oraz trajektorii zespolonych pulsacji chwilowych (s) sygnału $u_{4,4}[n]$ (kolor czarny), jego obwiedni minimalnofazowej (niebieski) i fazora dodatnioskrętnego (czerwony); rozkład zer na płaszczyźnie zespolonej (t).



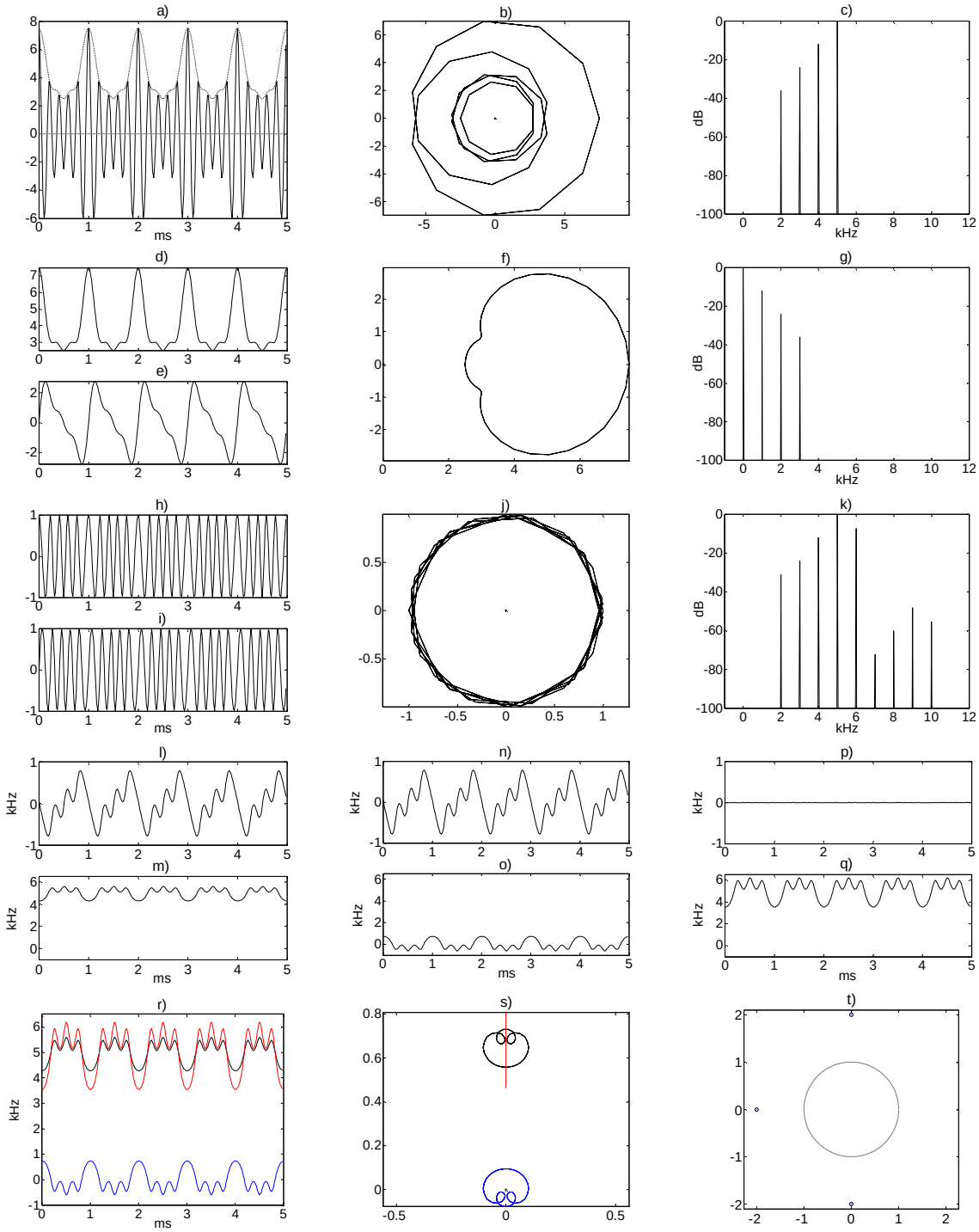
Rys. 4.8. Wyniki bifaktoryzacji sygnału 4-tonowego $u_{4,5}[n]$: część rzeczywista i obwiednia (linia przerywana) sygnału (a), trajektoria zespolona (b) i periodogram (c) sygnału, część rzeczywista (d) i urojona (e) MPE, trajektoria zespolona MPE (f) oraz periodogram MPE (g), część rzeczywista (h) i urojona (i) PIFP, trajektoria zespolona PIFP (j) oraz periodogram PIFP (k); przebiegi części rzeczywistych i urojonych ICF sygnału (l,m), jego obwiedni minimalnofazowej (n,o) oraz fazora dodatnioskrętnego (p,q); porównanie przebiegów IF (r) oraz trajektorii zespolonych pulsacji chwilowych (s) sygnału $u_{4,5}[n]$ (kolor czarny), jego obwiedni minimalnofazowej (niebieski) i fazora dodatnioskrętnego (czerwony); rozkład zer na płaszczyźnie zespolonej (t).



Rys. 4.9. Wyniki bifaktoryzacji sygnału 4-tonowego $u_{4,6}[n]$: część rzeczywista i obwiednia (linia przerywana) sygnału (a), trajektoria zespolona (b) i periodogram (c) sygnału, część rzeczywista (d) i urojona (e) MPE, trajektoria zespolona MPE (f) oraz periodogram MPE (g), część rzeczywista (h) i urojona (i) PIFP, trajektoria zespolona PIFP (j) oraz periodogram PIFP (k); przebiegi części rzeczywistych i urojonych ICF sygnału (l,m), jego obwiedni minimalnofazowej (n,o) oraz fazora dodatnioskrętnego (p,q); porównanie przebiegów IF (r) oraz trajektorii zespolonych pulsacji chwilowych (s) sygnału $u_{4,6}[n]$ (kolor czarny), jego obwiedni minimalnofazowej (niebieski) i fazora dodatnioskrętnego (czerwony); rozkład zer na płaszczyźnie zespolonej (t).



Rys. 4.10. Wyniki bifaktoryzacji sygnału 4-tonowego $u_{4,7}[n]$: część rzeczywista i obwiednia (linia przerywana) sygnału (a), trajektoria zespolona (b) i periodogram (c) sygnału, część rzeczywista (d) i urojona (e) MPE, trajektoria zespolona MPE (f) oraz periodogram MPE (g), część rzeczywista (h) i urojona (i) PIFP, trajektoria zespolona PIFP (j) oraz periodogram PIFP (k); przebiegi części rzeczywistych i urojonych ICF sygnału (l,m), jego obwiedni minimalnofazowej (n,o) oraz fazora dodatnioskrętnego (p,q); porównanie przebiegów IF (r) oraz trajektorii zespolonych pulsacji chwilowych (s) sygnału $u_{4,7}[n]$ (kolor czarny), jego obwiedni minimalnofazowej (niebieski) i fazora dodatnioskrętnego (czerwony); rozkład zer na płaszczyźnie zespolonej (t).



Rys. 4.11. Wyniki bifaktoryzacji sygnału 4-tonowego $u_{4,8}[n]$: część rzeczywista i obwiednia (linia przerywana) sygnału (a), trajektoria zespolona (b) i periodogram (c) sygnału, część rzeczywista (d) i urojona (e) MPE, trajektoria zespolona MPE (f) oraz periodogram MPE (g), część rzeczywista (h) i urojona (i) PIFP, trajektoria zespolona PIFP (j) oraz periodogram PIFP (k); przebiegi części rzeczywistych i urojonych ICF sygnału (l,m), jego obwiedni minimalnofazowej (n,o) oraz fazora dodatnioskrętnego (p,q); porównanie przebiegów IF (r) oraz trajektorii zespolonych pulsacji chwilowych (s) sygnału $u_{4,8}[n]$ (kolro czarny), jego obwiedni minimalnofazowej (niebieski) i fazora dodatnioskrętnego (czerwony); rozkład zer na płaszczyźnie zespolonej (t).

5. Zespolona pulsacja chwilowa w analizie głosu

W poprzednim rozdziale opisany został algorytm bifaktoryzacji V-K oraz właściwości ICF obu czynników tej faktoryzacji na przykładzie 4-tonowych sygnałów syntetycznych. W niniejszym rozdziale rozważymy możliwości zastosowania bifaktoryzacji V-K oraz zespolonej pulsacji chwilowej w analizie głosu.

5.1. Właściwości czynników V-KB oraz ich parametrów chwilowych na przykładzie polskich głosek

Zanim przejdziemy do omawiania metod analizy sygnału mowy za pomocą bifaktoryzacji V-K i zespolonej pulsacji chwilowej, przyjrzymy się właściwościom czynników faktoryzacji i ich parametrów chwilowych na przykładzie polskich głosek. Przedstawimy również sposób na wyznaczenie stopnia minimalnofazowości sygnału w oparciu o ICF. Opisane w tym podrozdziale eksperymenty przeprowadzone zostały w środowisku MATLAB. Wykorzystano w nich nagrania pojedynczych głosek (samogłosek i głosek sonarnych, które są kontynuantami, a więc ich wymowę można przedłużać) lub głoski wycięte z nagrań dłuższych wypowiedzi – fraz. Mowa nagrana była z 16-bitową rozdzielczością i szybkością próbkowania 48000 Sa/s, by spełnić warunek sporego nadpróbkowania dla poprawnej estymacji ICF, o czym pisaliśmy w rozdz. 4. Równoważnik hilbertowski rzeczywistego sygnału mowy otrzymaliśmy wykorzystując cyfrowy zespolony filtr Hilberta. Zastosowany filtr był pasmowoprzepustowy (200-7200 Hz), więc jednocześnie ograniczał pasmo sygnału mowy, zachowując jednak najważniejsze formanty (pierwsze cztery). Projektowanie tego filtru przeprowadzono w dwóch etapach: najpierw zaprojektowano filtr dolnoprzepustowy o częstotliwości odcięcia 3.5 kHz, a następnie jego odpowiedź impulsową przeskalowano tak, by uzyskać wzmocnienie 2 i przeheterodynowano na częstotliwość środkową 3.7 kHz.

5.1.1. Miara minimalnofazowości głosek

Mowa, jak większość naturalnie występujących sygnałów, jest sygnałem mieszanofazowym. Można jednak przypuszczać, że ze względu na kształt charakterystyki amplitudowej traktu głosowego i charakter widma tonu krtaniowego (o czym pisaliśmy w

podrozdz. 2.3), a także większe tłumienie wyższych częstotliwości w powietrzu, zarejestrowany sygnał mowy będzie prawie minimalnofazowy. Tezę tę należałoby sprawdzić stosując ilościową miarę udziału minimalnofazowości w sygnale mieszanofazowym. W celu określenia stopnia minimalnofazowości zespolonego sygnału $u[n]$ należy przeprowadzić dekompozycję jego zespolonej pulsacji chwilowej $s[n]$ na składową stałą $\bar{s}[n]$ oraz składowe dodatnio- i ujemnoczęstotliwościowe $s_+[n]$ oraz $s_-[n]$

$$s[n] = s_+[n] + s_-[n] + \bar{s}[n] \tag{5.1}$$

gdzie

$$s_+[n] = h_+[n] * s[n] \tag{5.2}$$

$$s_-[n] = h_-[n] * s[n] \tag{5.3}$$

a $h_+[n]$ i $h_-[n]$ są odpowiedziami impulsowymi filtrów o charakterystykach częstotliwościowych

$$H_+(\omega) = \begin{cases} 1; & 0 < \omega < \pi \\ 0; & -\pi < \omega < 0 \end{cases} \tag{5.4}$$

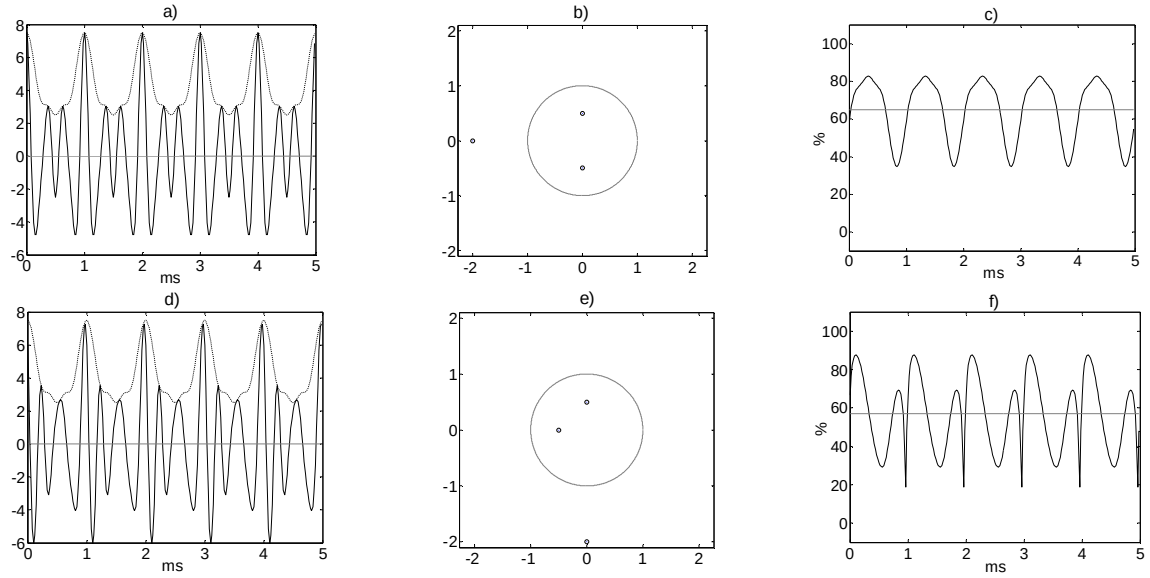
$$H_-(\omega) = \begin{cases} 1; & -\pi < \omega < 0 \\ 0; & 0 < \omega < \pi \end{cases} \tag{5.5}$$

Składowa stała $\bar{s}[n]$ jest w praktyce składową wolnozmienną. Dekompozycja ICF (5.1) odpowiada faktoryzacji sygnału analitycznego $u[n]$ na trzy czynniki: minimalnofazowy $u_+[n]$ (celowo oznaczony inaczej niż w rozdz. 4, by uzyskać zgodność z oznaczeniem filtru $h_+[n]$), maksymalnofazowy $u_-[n]$ i fazor $u_0[n]$, których zespolonymi pulsacjami chwilowymi są, odpowiednio, $s_+[n]$, $s_-[n]$ i $\bar{s}[n]$.

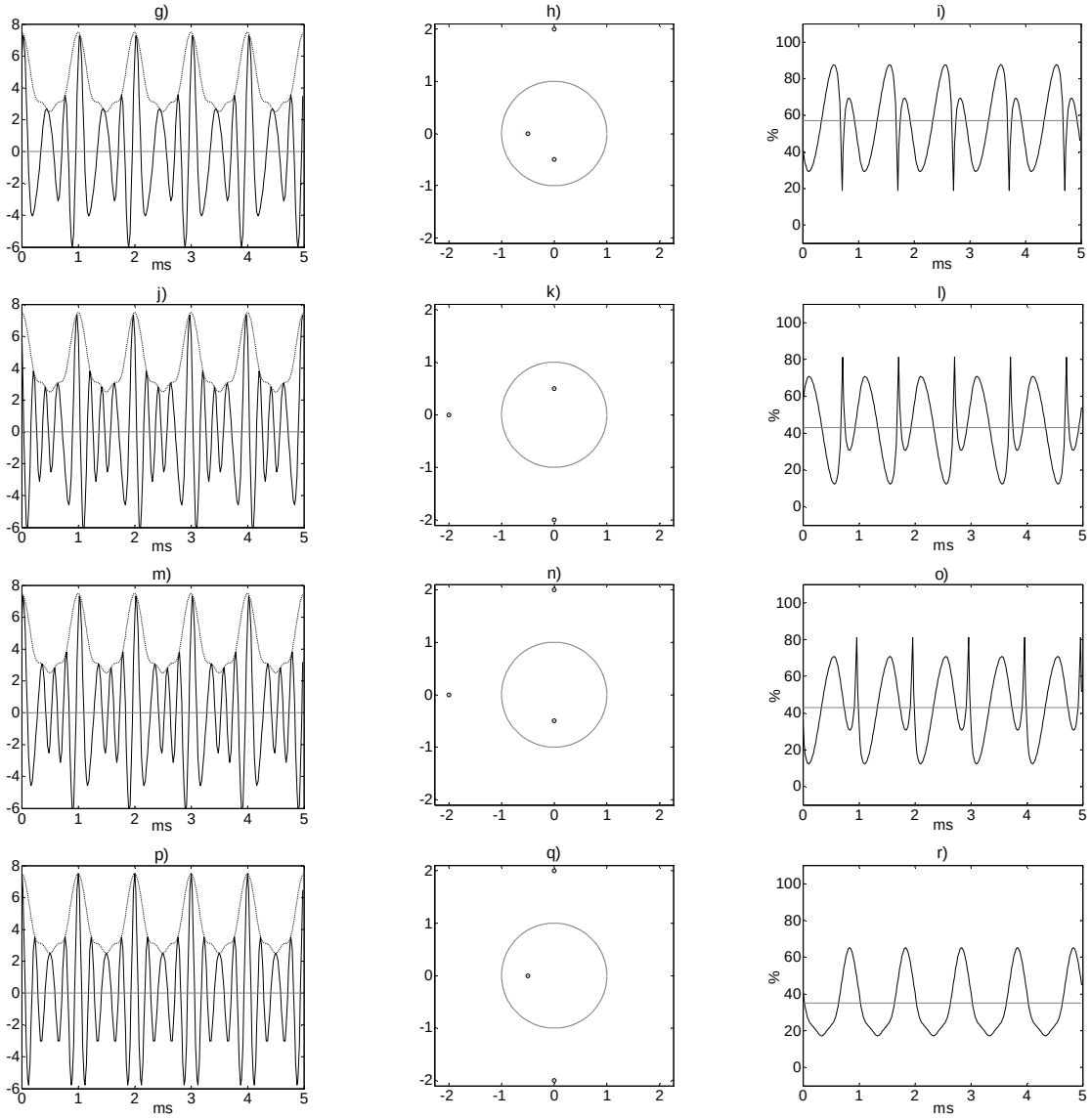
Zaproponowana poniżej miara $\eta[n]$ bieżącej chwilowej minimalnofazowości sygnałów

$$\eta[n] = \frac{|s_+[n]|}{|s_+[n]| + |s_-[n]|} \cdot 100\% \tag{5.6}$$

wykorzystuje przebiegi $s_+[n]$ oraz $s_-[n]$. Oczywiście spodziewamy się, że dla sygnału minimalnofazowego miara $\eta[n]$ będzie dla każdego n równa 100%, a dla sygnału maksymalnofazowego – 0%. Potwierdziły to symulacje przeprowadzone dla klasy ośmiu okresowych sygnałów 4-tonowych, których bifaktoryzację opisano w podrozdz. 4.6. Dla pozostałych sygnałów z tej klasy, będących sygnałami mieszanofazowymi, $\eta[n]$ jest przebiegiem zmiennym okresowym. Średnia $\eta[n]$ za okres wynosi: 65% dla sygnału $u_{4,2}[n]$, 57%, dla sygnałów $u_{4,3}[n]$ i $u_{4,4}[n]$, 43% dla sygnałów $u_{4,5}[n]$ i $u_{4,6}[n]$ oraz 35% dla sygnału $u_{4,7}[n]$. Przebiegi $\eta[n]$ dla tych sygnałów przedstawione zostały na rys. 5.1. Parametr $\eta[n]$ wykorzystano także do analizy stopnia minimalnofazowości polskich głosek. Zbadano 36 podstawowych głosek języka polskiego, a do eksperymentu użyto 10 realizacji każdej z nich wypowiadanych przez różnych mówców, kobiety i mężczyzn. Wyniki eksperymentu pokazały, że zakres parametru $\eta[n]$ dla każdej głoski może być bardzo szeroki. Jego wartość zmienia się nie tylko dla różnych mówców, ale również dla różnych realizacji jednej głoski wypowiadanej przez tą samą osobę.



Rys. 5.1. Oscylogramy (lewa kolumna), rozkład zer (środkowa kolumna) i przebieg parametru $\eta[n]$ (prawa kolumna) z zaznaczoną średnią wartością za okres dla sygnałów $u_{4,2}[n]$ (pierwszy wiersz) i $u_{4,3}[n]$ (drugi wiersz). Ciąg dalszy na następnej stronie.



Rys. 5.1. Oscylogramy (lewa kolumna), rozkład zer (środkowa kolumna) i przebieg parametru $\eta[n]$ (prawa kolumna) z zaznaczoną średnią wartością za okres dla sygnałów od $u_{4,4}[n]$ do $u_{4,7}[n]$ (wiersze od 1 do 4).

Otrzymane wyniki nie wskazują na bezpośrednią zależność między stopniem minimalnofazowości mowy a konkretnym mówcą, tzn. porównując dwóch mówców nie można stwierdzić, że stopień minimalnofazowści każdej wypowiadanej głoski jest niższy dla jednego z mówców. Ponadto, $\eta[n]$ może się znacznie zmieniać również w obrębie jednej wypowiadanej głoski, zwłaszcza dla głosek, które nie są kontynuantami.Ogólne wnioski, które można wyciągnąć na podstawie wartości średniej parametru $\eta[n]$, obliczonej dla każdej realizacji oraz dla poszczególnych głosek są następujące:

1) stopień minimalnofazowości głosek dźwięcznych jest wyższy niż dla głosek bezdźwięcznych;
2) głoski dźwięczne mają zawsze stopień minimalnofazowości wyższy niż 50% (dla ponad 90% głosek jest on wyższy niż 70%);
3) niższy stopień minimalnofazowości mają głoski dźwięczne, w których obok pobudzenia okresowego występuje również pobudzenie szumowe, są to głoski dentalizowane (np. /z/, /ź/, /ż/);
4) stopień minimalnofazowości głosek bezdźwięcznych nie przekracza 70%, a dla głosek bezdźwięcznych dentalizowanych jest on niższy niż 50%;
5) wszystkie samogłoski mają stopień minimalnofazowości wyższy niż 75%, najmniej minimalnofazowa jest głoska /a/, najbardziej głoska /i/; największe znaczenie dla stopnia minimalnofazowości samogłosek zdaje się mieć położenie języka w pionie podczas ich wypowiadania: głoski wysokie mają najwyższy stopień minimalnofazowości, a głoska niska /a/ - najniższy;
6) wśród spółgłosek najwyższy stopień minimalnofazowości mają głoski otwarte (półspółgłoski i półsamogłoski) oraz nosowe;
7) kolejność głosek bezdźwięcznych uszeregowanych względem rosnącego stopnia minimalnofazowości jest taka sama, jak kolejność odpowiadających im głosek dźwięcznych (odpowiadające sobie głoski bezdźwięczne i dźwięczne to np. /p/ i /b/, /t/ i /d/, /f/ i /w/);
8) zachowanie kolejności odpowiadających sobie głosek można również zauważyć wśród głosek dentalizowanych, tzn. głoski syczące mają kolejność: /c/, /s/, /z/, /dz/, głoski szumiące: /cz/, /sz/, /ż/, /dż/, głoski ciszące: /ć/, /ś/, /ź/, /dź/.

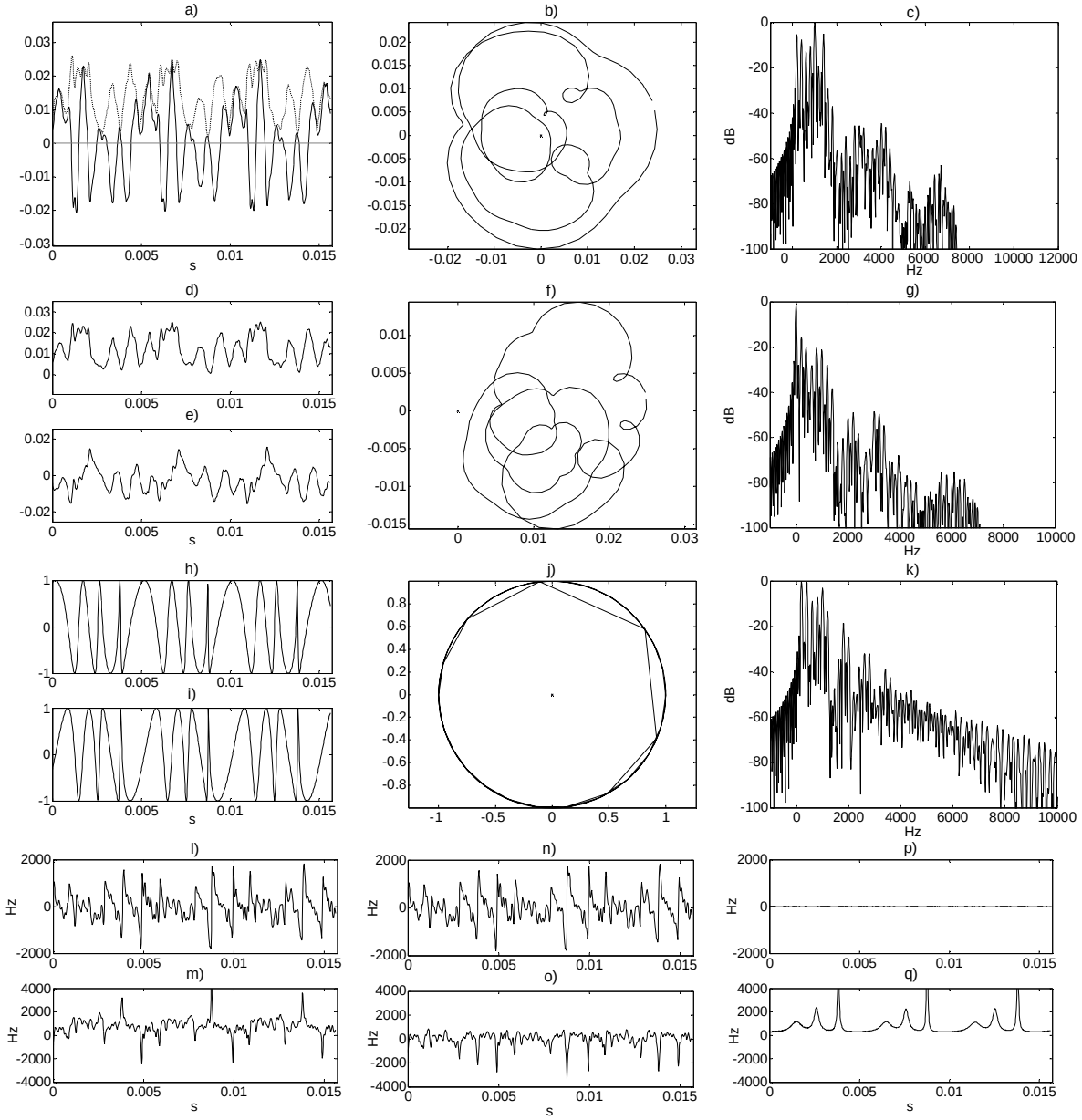
Ten eksperyment pokazał, że pomiar stopnia minimalnofazowości głosek może znaleźć zastosowanie w fonetyce i foniatrii, a prawdopodobnie również w badaniach nad mową zaburzoną. W kolejnych punktach pokażemy, że stopień minimalnofazowości głosek ma wpływ na kształt przebiegu ich częstotliwości chwilowej oraz ma duże znaczenie dla opisanych w tym rozdziale algorytmów estymacji częstotliwości podstawowej i ekstrakcji formantów.

5.1.2. Analiza polskich głosek

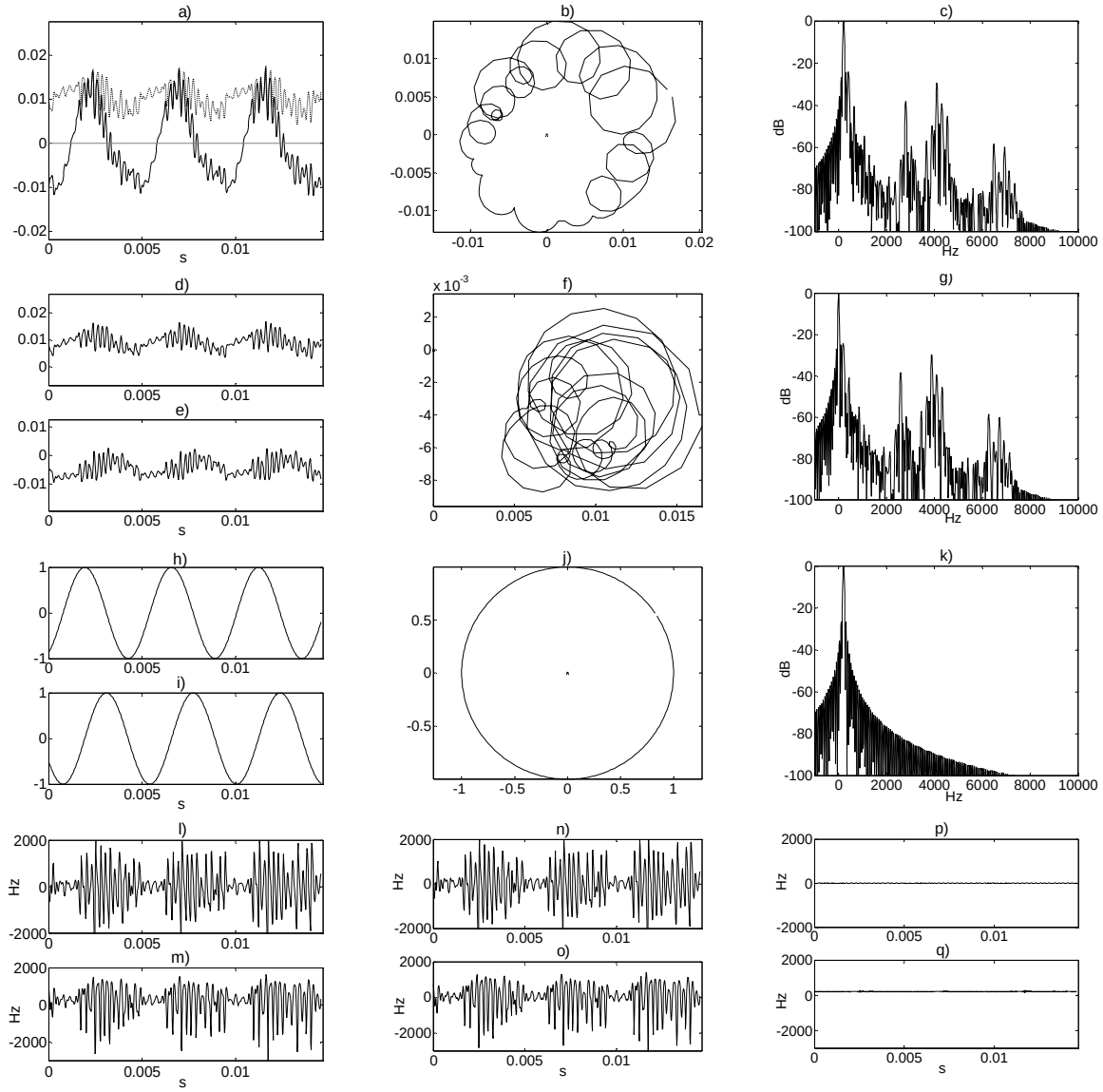
Wyniki bifaktoryzacji głosek zostaną przedstawione w sposób analogiczny jak w podrozdz. 4.6 dla sygnałów 4-tonowych. Pozwoli to obserwować i porównywać przebiegi otrzymane dla sygnału mowy oraz czynników jego faktoryzacji, a jednocześnie odnieść się do wniosków z poprzedniego rozdziału. Jako pierwsze przedstawione zostaną wyniki uzyskane dla samogłosek, gdyż są to głoski dźwięczne, mają wyraźną strukturę formantową, a ponadto mają wysoki stopień minimalnofazowości. Jako przykłady wybrane zostały wypowiadane przez tego samego mówcę (kobietę) głoski /a/ oraz /i/, odpowiednio o najniższym i najwyższym stopniu minimalnofazowości spośród wszystkich samogłosek. Wyniki bifaktoryzacji tych samogłosek przedstawiają rys. 5.2 oraz 5.3.

Obserwując widmo głoski /a/ (rys. 5.2c) widzimy formanty na częstotliwościach ok. 800, 1400, 3000 i 4000 Hz, przy czym pierwszy formant nie dominuje wyraźnie nad drugim. Środek ciężkości widma obliczony jako moment unormowany pierwszego rzędu wynosi tu 876 Hz. Formanty w widmie głoski /i/ (rys. 5.3c) znajdują się na częstotliwościach ok. 200, 2800, 4100 i 6900 Hz, a pierwszy formant wyraźnie dominuje nad pozostałymi. Środek ciężkości widma znajduje się na częstotliwości 472 Hz. W obu przypadkach widmo MPE (rys. 5.2g i 5.3g) jest oczywiście przesunięte na osi częstotliwości względem widma sygnału mowy (do zera), ale zachowuje jego strukturę formantową (widoczne są te same formanty na odpowiednio niższych częstotliwościach). Przy tym dla głoski /i/ kształt periodogramu MPE jest prawie niezmieniony w stosunku do periodogramu samej głoski. Dla głoski /a/ widoczne są zmiany w amplitudach poszczególnych formantów widocznych w periodogramie MPE w stosunku do periodogramu sygnału mowy. Jeśli natomiast porównamy przebiegi i periodogramy fazorów dodatnioskrętnych obu głosek, to zauważymy, że dla głoski /i/, która ma wyższy stopień minimalnofazowości (ponad 97%) fazor jest sinusoidą zespoloną o prawie niemodulowanej częstotliwości (rys. 5.3h,i). W jego periodogramie (rys. 5.3k) widoczny jest jeden prążek na częstotliwości 210 Hz (położenie pierwszego formantu w widmie głoski). Widzimy tu analogię do sygnału 4-tonowego minimalnofazowego omawianego w podrozdz. 4.6. W przebiegu PIFP głoski /a/ (rys. 5.2h,i) widać natomiast wyraźnie modulację częstotliwości, z bardzo szybkimi i dużymi zmianami fazy w chwilach, gdy amplituda sygnału mowy jest bliska zeru. Jego periodogram (rys. 5.2k) skupia się wokół częstotliwości 800 Hz (położenie pierwszego formantu w widmie głoski), ale zajmuje szersze pasmo niż

periodogram PIFP głoski /i/ (analogicznie do sygnałów 4-tonowych o jednym zerze poza okręgiem jednostkowym). Różnice PIFP głosek widoczne są również w przebiegach ich częstotliwości chwilowych. IF fazora głoski /i/ jest stała i ma wartość 210 Hz – wskazuje ona wyraźnie na pierwszy formant w widmie głoski.



Rys. 5.2. Wyniki bifaktoryzacji głoski /a/: przebieg i obwiednia (linia przerywana) głoski (a), trajektoria zespolona głoski (b), periodogram głoski (c), część rzeczywista (d) i urojona (e) MPE, trajektoria zespolona MPE (f) oraz periodogram MPE (g), część rzeczywista (h) i urojona (i) PIFP, trajektoria zespolona PIFP (j) oraz periodogram PIFP (k); przebiegi części rzeczywistych i urojonych ICF sygnału mowy (l,m), jego obwiedni minimalnofazowej (n,o) oraz fazora dodatnioskrętnego (p,q).



Rys. 5.3. Wyniki bifaktoryzacji głoski /i/: przebieg i obwiednia (linia przerywana) głoski (a), trajektoria zespolona głoski (b), periodogram głoski (c), część rzeczywista (d) i urojona (e) MPE, trajektoria zespolona MPE (f) oraz periodogram MPE (g), część rzeczywista (h) i urojona (i) PIFP, trajektoria zespolona PIFP (j) oraz periodogram PIFP (k); przebiegi części rzeczywistych i urojonych ICF sygnału mowy (l,m), jego obwiedni minimalnofazowej (n,o) oraz fazora dodatnioskrętnego (p,q).

Dla głoski /a/ przebieg IF fazora jest zmienny i widoczne są w nim szpilki w chwilach, gdy amplituda sygnału mowy jest bliska zeru (wysokość tych szpilek nie ma znaczenia). Jak zauważono analizując również wyniki bifaktoryzacji dla innych głosek, IF fazora dodatnioskrętnego ma tym gładszy przebieg im wyższy jest stopień minimalnofazowości analizowanej głoski. Średnia IF fazora głoski /a/ wynosi 805 Hz, a więc również wskazuje ona na pierwszy formant w widmie głoski. Jeśli przyjrzymy się częściom rzeczywistym

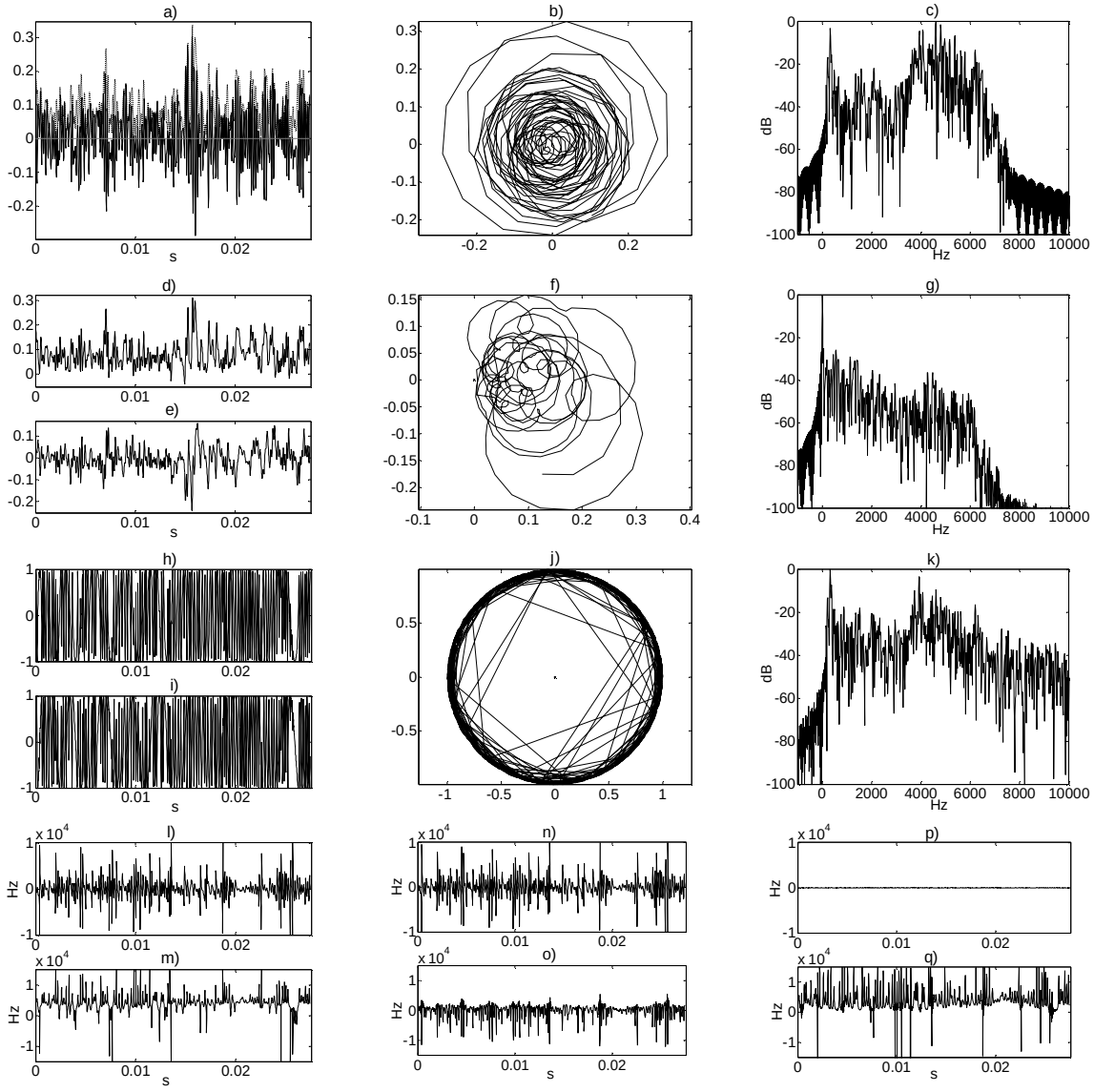
przebiegów ICF obu głosek, czyli $\sigma[n]$, zauważymy, że wartość średnia $|\sigma[n]|$ wyrażona w hercach odpowiada szerokości pasma pierwszego formantu. Para $(\omega_{pif}[n], |\sigma[n]|)$ jest więc bardzo przydatnym chwilowym deskryptorem głównego formantu w widmie sygnału mowy.

Warto w tym miejscu jeszcze raz podkreślić, że wartość żadnej z częstotliwości chwilowych (IF sygnału mowy, obwiedni minimalnofazowej czy fazora dodatnioskrętnego) nie wskazuje na częstotliwość podstawową mowy, środek ciężkości widma czy też częstotliwość środkową pasma zajmowanego przez sygnał mowy. Natomiast to, że $\omega_{pif}[n]$ i $|\sigma[n]|$ wskazują na położenie i szerokość pasma głównego formantu, a głos ma charakter prawie minimalnofazowy, pozwala wykorzystać bifaktoryzację V-K oraz ICF do ekstrakcji formantów, co zostanie opisane w podrozdz. 5.3.

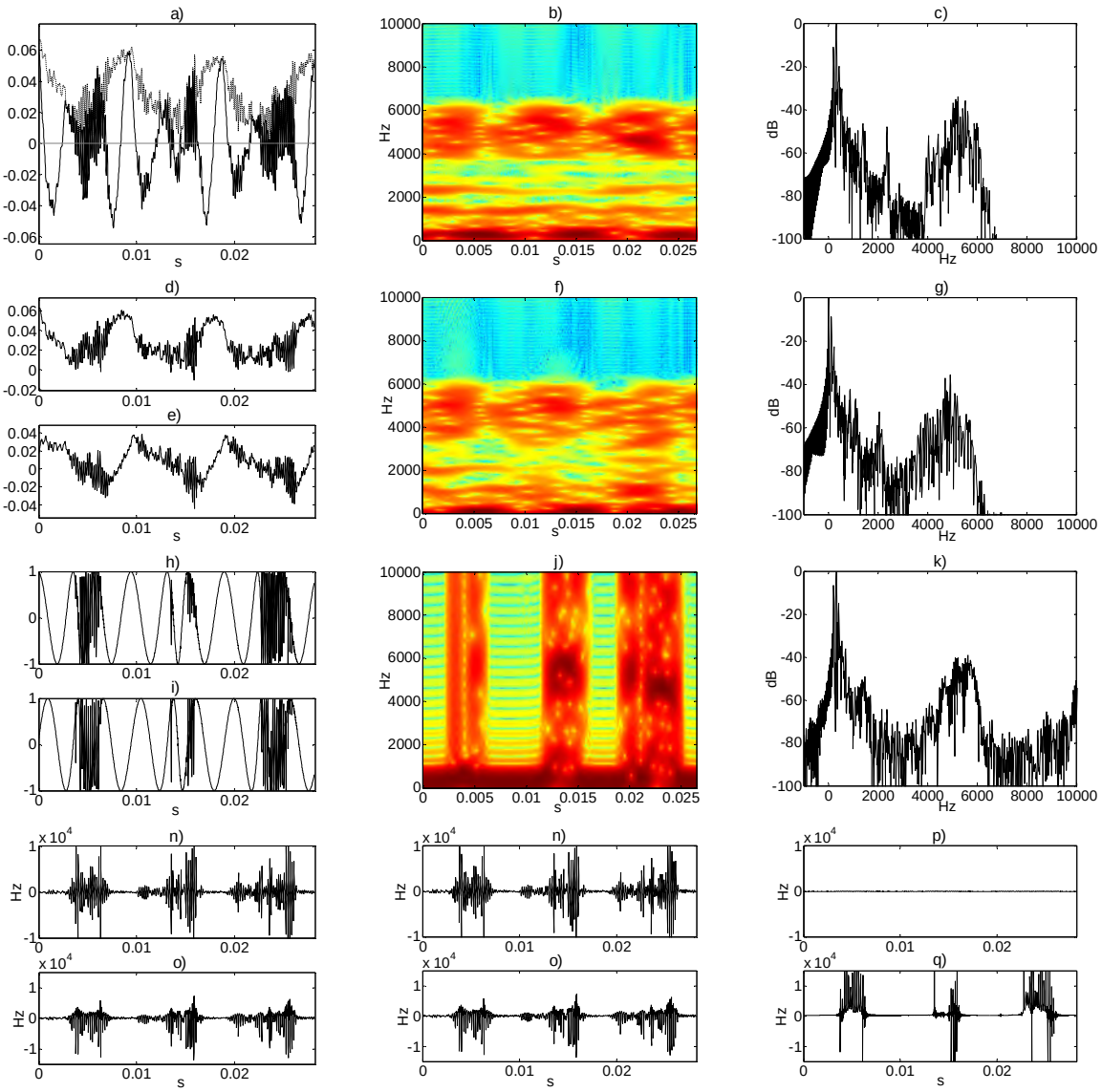
Analogiczne wyniki, jak dla głosek /a/ oraz /i/, uzyskano dla innych głosek dźwięcznych. Różnice między przebiegami ich częstotliwości chwilowych wynikają przede wszystkim ze stopnia ich minimalnofazowości. Warto jednak przyjrzeć się również wynikom bifaktoryzacji głosek bezdźwięcznych, oraz dźwięcznych, w których obok pobudzenia okresowego występuje również pobudzenie szumowe. Dla zilustrowania ich właściwości wybrano głoski /s/ oraz /z/. Wykresy otrzymane dla tych głosek zaprezentowano na rysunkach, odpowiednio 5.4 oraz 5.5. Z rys. 5.4 widać, że głoska /s/ ma wyraźnie pobudzenie szumowe, brak jest struktury harmonicznej w widmie. Ponadto widmo jest dość wyrównane, ale można w nim wyodrębnić poszczególne formanty. Ze względu na duży udział wysokich częstotliwości w widmie, sygnał tej głoski ma bardzo niski stopień minimalnofazowości (zawsze poniżej 50%), a środek ciężkości widma znajduje się w okolicach 4000 Hz. Widmo MPE nie zachowuje zupełnie struktury formantowej głoski. Warto też zwrócić uwagę, że widmo PIFP ma nieograniczone pasmo. Z tego względu w przebiegu IF fazora można zauważyć bardzo szybkie zmiany o dużej amplitudzie. IF oscyluje wokół środka ciężkości widma głoski, ponieważ nie ma w nim żadnego wyraźnie dominującego formantu.

Analizując oscylogram głoski /z/ można zauważyć wyraźną segmentację, tj. przedziały czasowe, w których charakter sygnału jest różny (z tego względu oprócz periodogramów, które przedstawiają uśrednione widmo mocy sygnału, na rys. 5.5 umieszczono również spektrogramy, które pozwalają zaobserwować, jak widmo zmienia się w czasie). W pierwszym przedziale (od ok. 7 do 12 ms) występuje wyłącznie pobudzenie quasi-okresowe,

przebieg sygnału jest gładki, a na spektrogramie można zauważyć cztery formanty, z których pierwszy jest wyraźnie dominujący. W drugim przedziale (od ok. 12 do 16 ms) obserwujemy, że oprócz pobudzenia quasi-okresowego występuje również pobudzenie szumowe, które powoduje znaczne podwyższenie amplitud składowych widma wokół 5000 Hz, przez co pierwszy formant przestaje być dominujący. Taki podział powtarza się w każdym cyklu sygnału widocznym na rys. 5.5a.



Rys. 5.4. Wyniki bifaktoryzacji głoski /s/: przebieg i obwiednia (linia przerywana) głoski (a), trajektoria zespolona głoski (b), periodogram głoski (c), część rzeczywista (d) i urojona (e) MPE, trajektoria zespolona MPE (f) oraz periodogram MPE (g), część rzeczywista (h) i urojona (i) PIFP, trajektoria zespolona PIFP (j) oraz periodogram PIFP (k); przebiegi części rzeczywistych i urojonych ICF sygnału mowy (l,m), jego obwiedni minimalnofazowej (n,o) oraz fazora dodatnioskrętnego (p,q).



Rys. 5.5. Wyniki bifaktoryzacji głoski /z/: przebieg i obwiednia (linia przerywana) głoski (a), spektrogram (b) oraz periodogram (c) głoski, część rzeczywista (d) i urojona (e) MPE, spektrogram MPE (f) oraz periodogram MPE (g), część rzeczywista (h) i urojona (i) PIFP, spektrogram PIFP (j) oraz periodogram PIFP (k); przebiegi części rzeczywistych i urojonych ICF sygnału mowy (l,m), jego obwiedni minimalnofazowej (n,o) oraz fazora dodatnioskrętnego (p,q).

Dodajmy też, że w pierwszym przedziale stopień minimalnofazowości głoski /z/ jest bardzo wysoki, ok. 90%, a w drugim nie wykracza ponad 20%.

Najbardziej interesujące nas różnice można zauważyć w przebiegach PIFP i jego częstotliwości chwilowej. W pierwszym przedziale PIFP i jego IF zachowują się analogicznie, jak w omawianej wcześniej głosce /i/ – widmo fazora jest ograniczone do ok. 1000 Hz, a przebieg IF jest prawie stały o wartości ok. 250 Hz (położenie pierwszego, dominującego

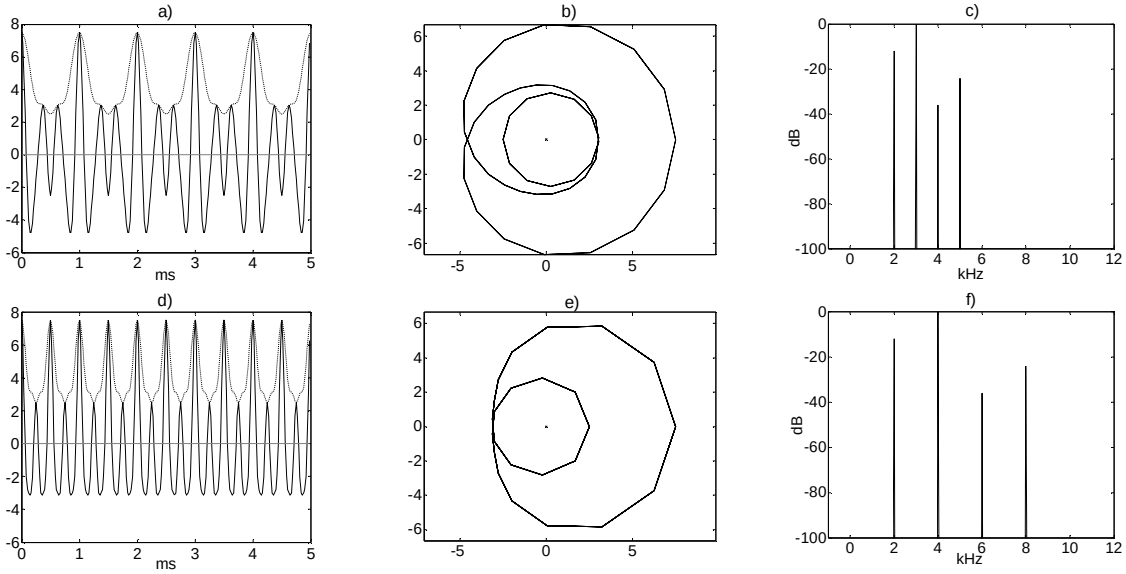
formantu w widmie). W drugim przedziale fazor głoski /z/ wykazuje większe podobieństwo do bezdźwięcznej głoski /s/ – jego widmo ma nieograniczone pasmo, a zmiany jego IF są bardzo szybkie, o dużej amplitudzie. Ponadto w drugim przedziale wartość uśrednionej IF podnosi się nawet do 3000 Hz (w zależności od tego jak duży jest udział pobudzenia szumowego i jak duże jest w związku z tym podniesienie amplitud składowych widma wokół częstotliwości 5000 Hz). Efekt ten można częściowo wyeliminować wymuszając większy stopień minimalnofazowości sygnału mowy poprzez zastosowanie deemfazy, co omówimy w dalszej części rozdziału.

Analizując mowę warto również posłuchać jak brzmią czynniki bifaktoryzacji V-K. Jak można było oczekiwać, PIFP nie przenosi treści wypowiedzi. Im wyższy jest stopień minimalnofozowości analizowanej głoski, tym bardziej brzmienie fazora przypomina sygnał sinusoidalny. Przy odsłuchu MPE, mimo pewnych zniekształceń, możliwe jest całkowite zrozumienie treści wypowiadanej frazy (co jest zgodne z badaniami, które pokazują, że najważniejsze dla zrozumiałości mowy są wolne zmiany w widmie sygnału, które są wynikiem niskoczęstotliwościowej modulacji amplitudy [KI98]). Zachowany jest również przebieg częstotliwości podstawowej.

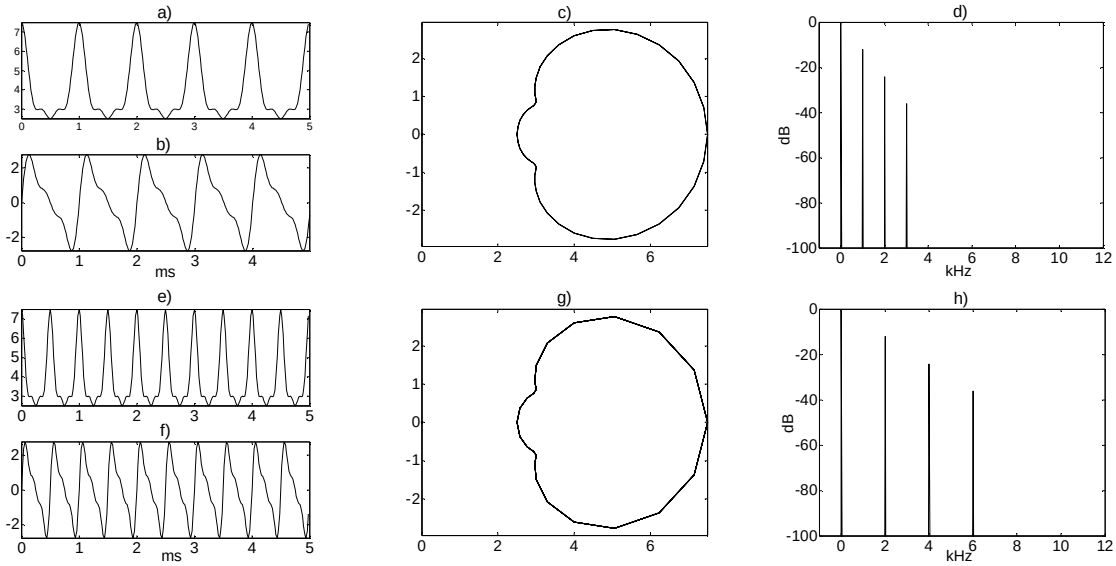
5.1.3. IF a częstotliwość podstawowa

Jak pokazały wyniki symulacji przeprowadzonych dla sygnałów syntetycznych (podrozdz. 4.6) oraz analiza sygnału mowy (p. 5.1.2), wartości przebiegów IF sygnału analitycznego ani czynników faktoryzacji V-K nie wskazują na częstotliwość podstawową analizowanego sygnału. Można jednak zauważyć, że jeżeli analizowany sygnał jest okresowy, to zarówno $s[n]$, jak i $s_{mp}[n]$ (ich części rzeczywiste i urojone) są również okresowe, o tym samym okresie. Dla częstotliwości chwilowej PIFP prawidłowość ta nie zawsze występuje, gdyż np. $\omega_{pif}[n]$ sygnału minimalnofazowego jest przebiegiem stałym. Aby sprawdzić, jak zmiana częstotliwości podstawowej analizowanego sygnału wpływa na przebiegi $s[n]$, $s_{mp}[n]$ oraz $\omega_{pif}[n]$, wykonano ponownie symulacje dla klasy ośmiu sygnałów syntetycznych z podrozdz. 4.6, przy czym zmniejszono dwukrotnie ich okres, pozostawiając niezmienione amplitudy poszczególnych prążków oraz częstotliwość prążka o najniższej częstotliwości.

Częstotliwości kolejnych składowych wynosiły więc 2000 Hz, 4000 Hz, 6000 Hz i 8000 Hz. Kształt obwiedni sygnałów pozostał niezmieniony, tylko jej okres uległ dwukrotnemu skróceniu. Wyniki przeprowadzonego eksperymentu przedstawiono na przykładzie sygnału $u_{4,2}[n]$ na rys. 5.6 – 5.9. Dla pozostałych siedmiu sygnałów z badanej klasy rezultaty były analogiczne.



Rys. 5.6. Oscylogram części rzeczywistej sygnału $u_{4,2}[n]$ wraz z obwiednią (lewa kolumna), jego trajektoria zespolona (środkowa kolumna) oraz periodogram (prawa kolumna) przed (górny wiersz) i po (dolny wiersz) podwyższeniu częstotliwości podstawowej.

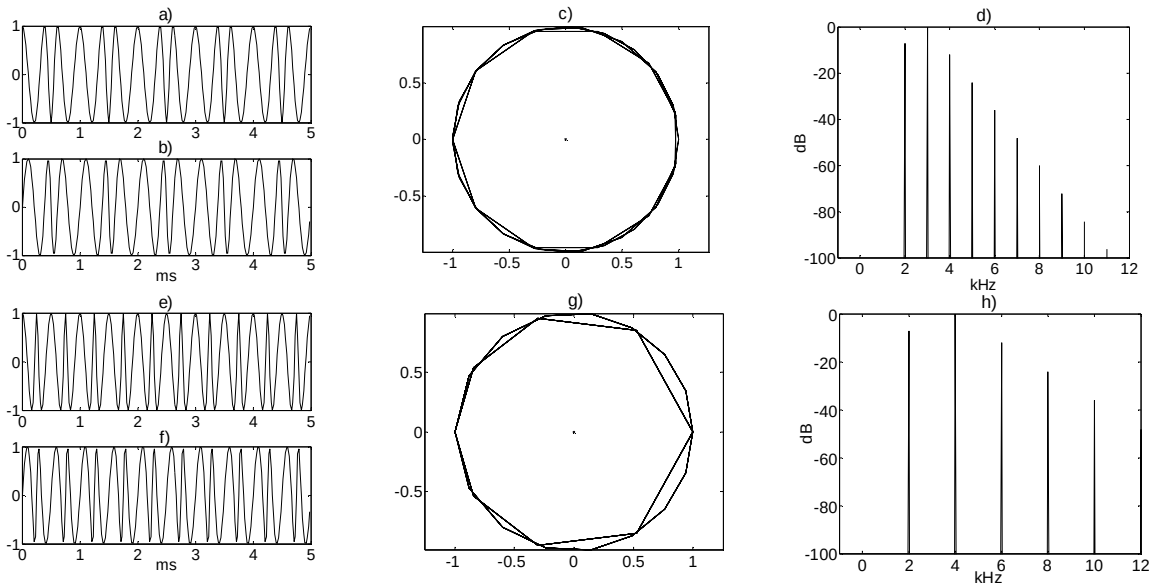


Rys. 5.7. Oscylogramy części rzeczywistej i urojonej MPE (lewa kolumna), jej trajektoria zespolona (środkowa kolumna) oraz periodogram (prawa kolumna) przed (górny wiersz) i po (dolny wiersz) podwyższeniu częstotliwości podstawowej.

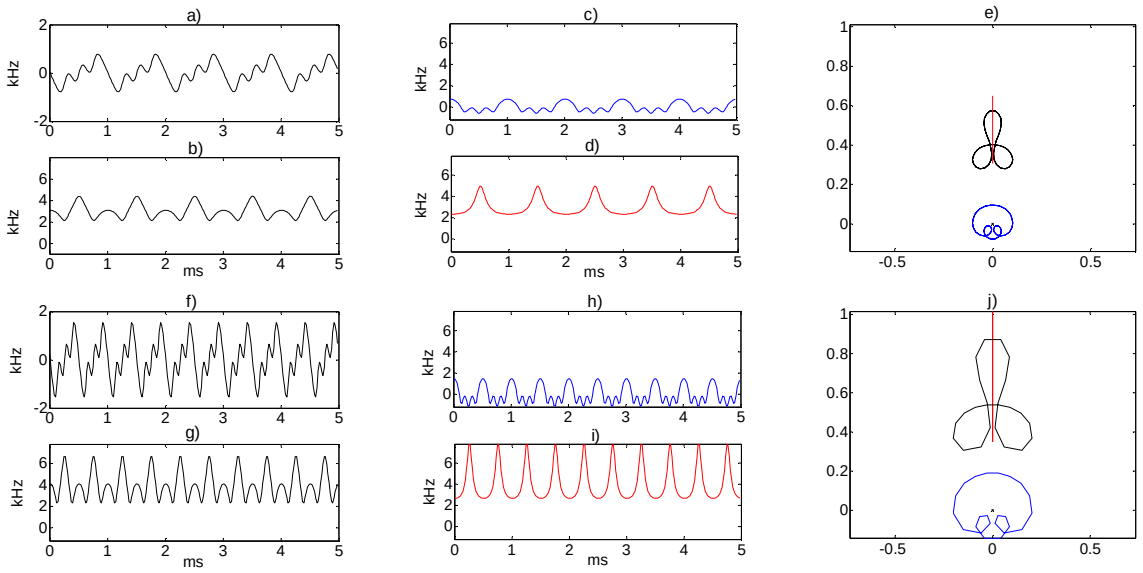
Z analizy rysunków można wyciągnąć następujące wnioski:

1) w przebiegach części rzeczywistych i urojonych MPE oraz PIFP można zauważyć wyłącznie zmianę długości okresu (jest on tu, rzecz jasna, dwa razy krótszy);
2) kolejne prążki MPE oraz PIFP mają niezmienioną amplitudę, zwiększył się tylko dwukrotnie odstęp pomiędzy nimi;
3) kształt trajektorii ICF na płaszczyźnie zespolonej nie zmienił się;
4) przebiegi części rzeczywistych i urojonych $s[n]$ oraz $s_{mp}[n]$ mają dwukrotnie krótszy okres oraz amplituda ich zmian zwiększyła się ok. dwa razy (wynika to z poszerzenia pasma sygnału); taką samą prawidłowość można zauważyć dla przebiegu $\omega_{pif}[n]$ z wyjątkiem sygnału minimalnofazowego, gdy $\omega_{pif}[n]$ jest stała;
5) średnie wartości IF sygnału oraz fazora dodatnioskrętnego wynoszą 2000 Hz (sygnał $u_{4,1}[n]$), 4000 Hz ($u_{4,2}[n], u_{4,3}[n], u_{4,4}[n]$), 6000 Hz ($u_{4,5}[n], u_{4,6}[n], u_{4,7}[n]$) oraz 8000 Hz ($u_{4,8}[n]$), natomiast przed zmniejszeniem okresu wynosiły odpowiednio 2000 Hz, 3000 Hz, 4000 Hz i 5000 Hz.

Zmiana wartości średniej IF wynika ze zmiany położenia dominujących prążków w widmach poszczególnych sygnałów i nie jest proporcjonalna do zmiany częstotliwości podstawowej.

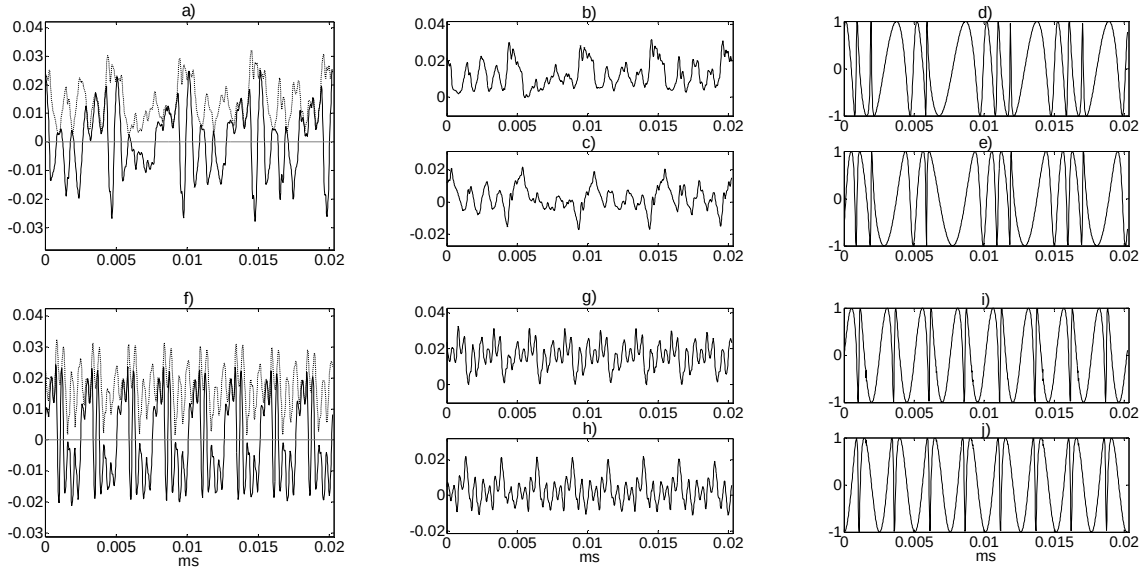


Rys. 5.8. Oscylogramy części rzeczywistej i urojonej PIFP (lewa kolumna), jego trajektoria zespolona (środkowa kolumna) oraz periodogram (prawa kolumna) przed (górny wiersz) i po (dolny wiersz) podwyższeniu częstotliwości podstawowej.

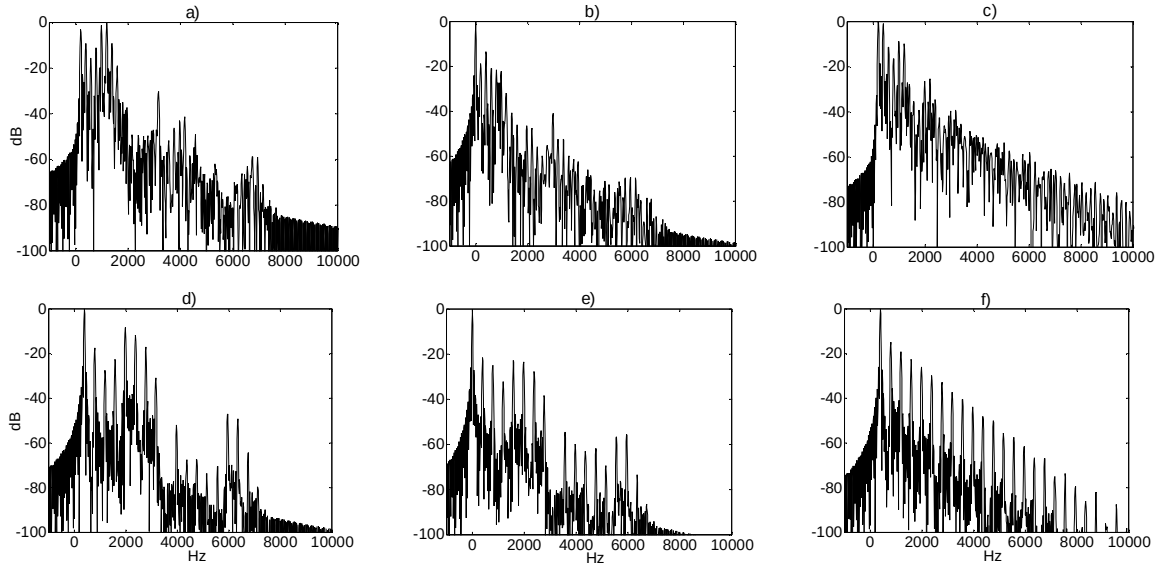


Rys. 5.9. Oscylogramy części rzeczywistej i urojonej ICF sygnału $u_{4,2}[n]$ (lewa kolumna), przebiegi IF obwiedni minimalnofazowej i fazora dodatnioskrętnego (środkowa kolumna) oraz trajektorie zespolone ICF sygnału $u_{4,2}[n]$ (kolor czarny), obwiedni minimalnofazowej (niebieski) i fazora dodatnioskrętnego (czerwony) (prawa kolumna) przed (górny wiersz) i po (dolny wiersz) podwyższeniu częstotliwości podstawowej.

Analogiczne wyniki można uzyskać analizując sygnał mowy o zmieniającej się częstotliwości podstawowej. W tym celu dokonano porównania czynników bifaktoryzacji V-K oraz ich ICF-ów głoski /a/ po zmodyfikowaniu jej częstotliwości podstawowej w programie Adobe Audition (dwukrotne zwiększenie częstotliwości podstawowej). Oscylogramy i periodogramy poszczególnych przebiegów przed i po zmianie częstotliwości podstawowej przedstawiają rys. 5.10, 5.11 i 5.12. Rys. 5.10 pokazuje, że po dwukrotnym zwiększeniu częstotliwości podstawowej okres wszystkich przebiegów (sygnału mowy, obwiedni minimalnofazowej i fazora dodatnioskrętnego) uległ dwukrotnemu skróceniu. Ponadto, w niewielkim stopniu zmienił się kształt przebiegu sygnału mowy (rys. 5.10a i 5.10f), co wynika z zastosowanego algorytmu modyfikacji. Widać to również w widmie sygnału mowy (rys. 5.11a i 5.11d), w którym poza rozsunięciem prążków można zauważyć zmiany amplitud poszczególnych prążków (np. znaczne wzmocnienie prążka o częstotliwości podstawowej). Mimo tych zniekształceń można jednak na podstawie tego eksperymentu wyciągnąć pewne wnioski dotyczące wpływu zmiany częstotliwości podstawowej sygnału mowy na zespolone pulsacje chwilowe. Analogicznie jak dla sygnałów syntetycznych, ICF sygnału mowy i MPE zachowują quasi-okresowość sygnału mowy, a po dwukrotnym podwyższeniu częstotliwości podstawowej okres skraca się dwukrotnie.



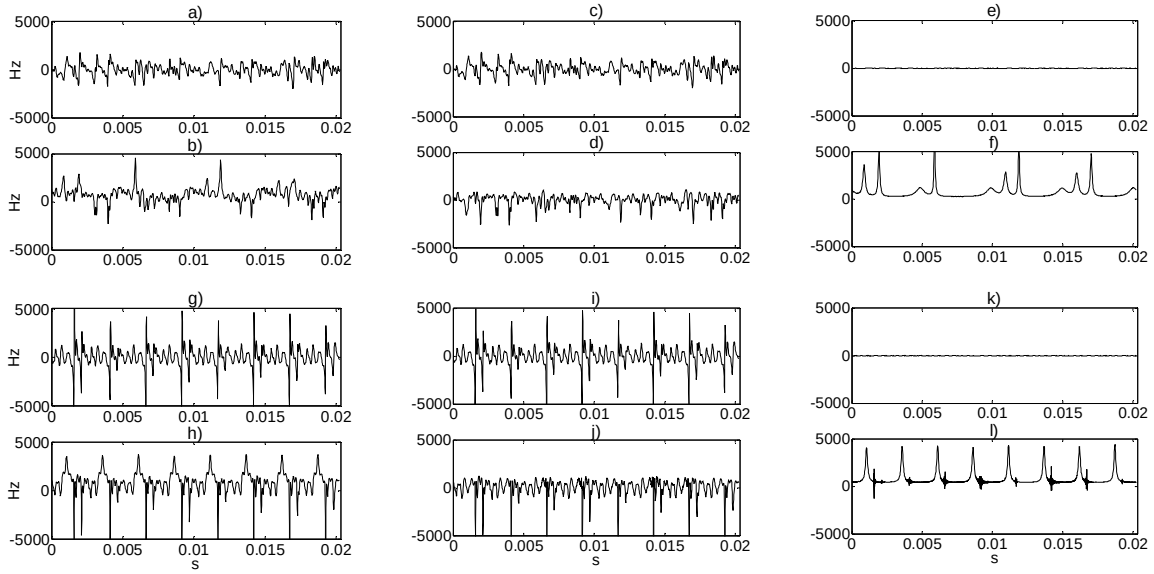
Rys. 5.10. Oscylogramy głoski /a/ wraz z obwiednią chwilową (lewa kolumna), części rzeczywistej i urojonej MPE (środkowa kolumna) oraz części rzeczywistej i urojonej PIFP (prawa kolumna) przed (górny wiersz) i po (dolny wiersz) modyfikacji częstotliwości podstawowej.



Rys. 5.11. Periodogramy sygnału mowy (lewa kolumna), MPE (środkowa kolumna) oraz PIFP (prawa kolumna) przed (górny wiersz) i po (dolny wiersz) modyfikacji częstotliwości podstawowej.

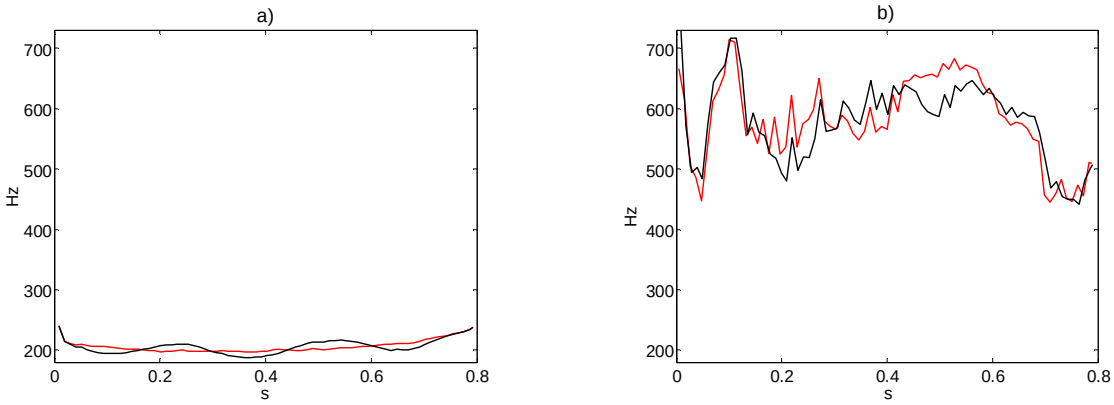
Nie można natomiast zauważyć zwiększenia amplitudy zmian tych przebiegów, gdyż szerokości pasm poszczególnych formantów ani sygnału mowy nie zmieniły się. Szpilki zauważalne w przebiegach $s[n]$ i $s_{mp}[n]$ są, jak już wspominaliśmy, wynikiem zbliżania się wartości amplitudy chwilowej sygnału mowy do zera, a ich wielkość nie ma znaczenia. Przebieg IF fazora dodatnioskrętnego również zachowuje quasi-okresowość sygnału mowy,

która zmienia się po modyfikacji częstotliwości podstawowej.Bardziej istotna jest jednak zmiana wartości IF. Przed modyfikacją F_0 średnia wartość IF za okres wynosiła 742 Hz, po modyfikacji wartość ta zmieniła się na 793 Hz, co było wynikiem przesunięcia częstotliwości środkowych formantów. Jednakże żadna bezpośrednia zależność między czynnikiem skalowania F_0 a zmianą wartości średniej IF tu nie występuje.



Rys. 5.12. Części rzeczywiste i urojone ICF sygnału mowy (lewa kolumna), obwiedni minimalnofazowej (środkowa kolumna) oraz fazora dodatnioskrętnego (prawa kolumna) przed (górny wiersz) i po (dolny wiersz) modyfikacji częstotliwości podstawowej.

Potwierdza to kolejny eksperyment, w którym częstotliwość podstawowa wypowiadanej głoski /a/ zmieniana była płynnie w zakresie 180-240 Hz. Wyniki tego eksperymentu przedstawia rys. 5.13, na którym zamieszczono przebiegi F_0 oraz wygładzone przebiegi IF przed i po modyfikacji częstotliwości podstawowej (F_0 estymowana była za pomocą algorytmu YIN [CH02][KA06][KA07], natomiast wygładzony przebieg IF uzyskano stosując okno Hamminga o długości 512 próbek). Analiza rys. 5.13 wyraźnie pokazuje, że nie ma proporcjonalności pomiędzy zmianami F_0 a zmianami IF. Przykładowo, dla chwili *t*=0.1s, F_0 po modyfikacji jest niższa niż w sygnale oryginalnym, natomiast IF prawie się nie zmieniła. Dla *t*=0.5s IF po modyfikacji ma niższą wartość, mimo że F_0 została podwyższona. Zmiany IF fazora dodatnioskrętnego nie są więc bezpośrednim wynikiem modyfikacji F_0, ale skutkiem zmian w strukturze formantowej sygnału mowy.



Rys. 5.13. Przebiegi F_0 (a) oraz IF (b) przed (kolor czerwony) i po (czarny) modyfikacji częstotliwości podstawowej.

Warto również zaznaczyć, że modyfikacja częstotliwości podstawowej wpływa na brzmienie przebiegów $a_{mp}[n]$ oraz $\lambda_{pif}[n]$; podczas ich odsłuchiwania wyraźnie słychać zmiany wysokości dźwięku odpowiadające zmianom F_0.

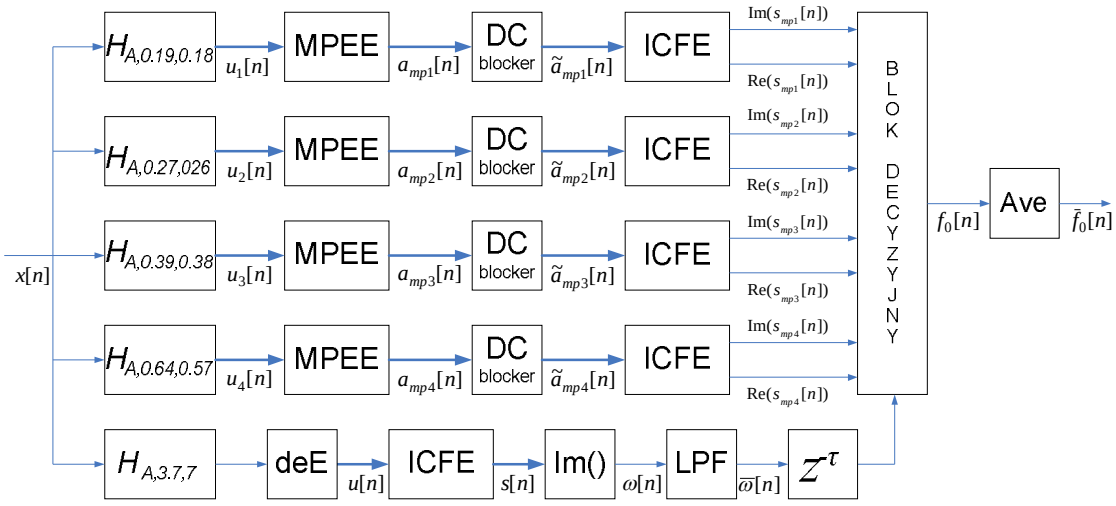
5.2. Estymacja częstotliwości podstawowej

Częstotliwość podstawowa sygnału mowy jest parametrem ściśle związanym, choć nie identycznym z percypowaną wysokością głosu, co szerzej wyjaśnimy w rozdz. 6. Odpowiada ona częstotliwości powtarzania impulsów krtaniowych, generowanych przez fałdy głosowe. Detekcja i pomiar częstotliwości podstawowej F_0 jest wciąż jednym z najważniejszych zadań analizy sygnału mowy. Znajduje ona zastosowanie w wielu aplikacjach: rozpoznawaniu mowy, identyfikacji mówców, syntezie i rekonstrukcji głosu, przesyłaniu mowy w systemach telekomunikacyjnych, poszerzaniu wiedzy z zakresu językoznawstwa i fonetyki, a także w diagnostyce i terapii zaburzeń mowy [DE93] [GE03][JO03][JA07]. Przeprowadzone badania wykazały, że dokładność estymacji częstotliwości podstawowej ma istotne znaczenie dla jakości przetwarzania sygnału mowy (syntezy, kodowania, transmisji, kompresji itp.). W najczęściej stosowanych algorytmach estymacji częstotliwości podstawowej, bazujących np. na predykcji liniowej, autokorelacji, czy cepstrum [DE93][GE03][KA06][RA07], mowa przetwarzana jest w ramkach. Takie podejście wymaga założenia, że sygnał mowy jest lokalnie stacjonarny, tzn. jego parametry, w tym częstotliwość podstawowa, nie ulegają zmianie w obrębie ramki. W rzeczywistości jednak chwilowa częstotliwość podstawowa

mowy może zmieniać się znacząco nawet w obrębie jednego okresu [RA76], podczas gdy ramka w algorytmach analizy mowy powinna obejmować zazwyczaj co najmniej dwa okresy sygnału.

Od pewnego czasu badana jest możliwość wykorzystania IF do estymacji częstotliwości podstawowej [GI05][MA95][RA00]. Zaletą takich metod jest uwzględnienie niestacjonarnego charakteru sygnału mowy i możliwość estymowania szybkich zmian częstotliwości podstawowej. W niniejszym podrozdziale zaproponujemy potokowy algorytm estymacji częstotliwości podstawowej, wykorzystujący zespoloną pulsację chwilową ICF. Jego schemat blokowy przedstawia rys. 5.14. W proponowanym algorytmie sygnał mowy przetwarzany jest równolegle w pięciu gałęziach. Pierwsze cztery gałęzie estymują prawdopodobne wartości częstotliwości podstawowej. Estymacja ta oparta jest na wyznaczeniu IF (części urojonej ICF) przebiegu w każdej z czterech gałęzi. Wybór jednego z kandydatów dokonywany jest natomiast na podstawie IB (części rzeczywistej ICF). W ostatniej z pięciu gałęzi widocznych na rys. 5.14 przeprowadzana jest klasyfikacja mowy na dźwięczną i bezdźwięczną. Algorytm ten jest algorytmem potokowym – estymacja częstotliwości podstawowej jak również klasyfikacja mowy na dźwięczną i bezdźwięczną, przeprowadzane są w takt kolejnych próbek wejściowego sygnału mowy. Na rys. 5.14 zastosowano oznaczenia:

- $H_{A,F_{Ci}B_i}$, i=1,2,3,4 – bank zespolonych filtrów Hilberta (F_{Ci} określa częstotliwość środkową i-tego filtru, a B_i szerokość jego pasma w kHz);
- $H_{A,3.7,7}$ – zespolony filtr Hilberta o częstotliwości środkowej 3.7 kHz i szerokości pasma 7 kHz;
- MPEE – estymator obwiedni minimalnofazowej;
- DC blocker – filtr wycinający składową stałą;
- ICFE – estymator zespolonej pulsacji chwilowej;
- Im( ) – część urojona;
- Re( ) – część rzeczywista;
- deE – filtr deemfazy IIR;
- LPF – filtr dolnoprzepustowy o częstotliwości odcięcia 50 Hz;



Rys. 5.14. Schemat blokowy potokowego algorytmu estymacji częstotliwości podstawowej.

- $z^{-\tau}$ – element opóźniający, wyrównujący opóźnienie w pierwszej gałęzi względem pozostałych gałęzi (opóźnienie w gałęziach 2, 3, 4 i 5 jest takie samo);
- Ave – filtr uśredniający zrealizowany za pomocą okna Hamminga o długości 512 próbek.

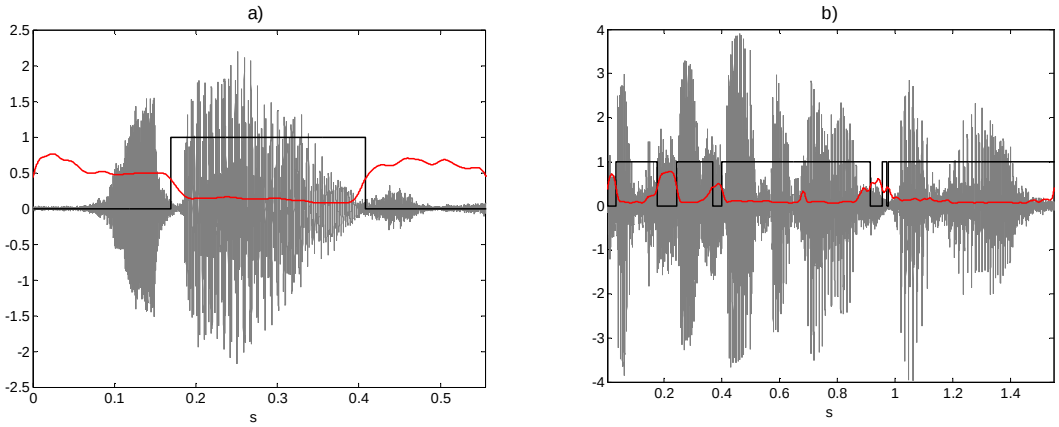
W kolejnych podrozdziałach omówione zostaną dokładnie wszystkie gałęzie przedstawionego algorytmu.

5.2.1. Klasyfikacja mowy na dźwięczną i bezdźwięczną

Klasyfikacja mowy na dźwięczną i bezdźwięczną jest istotnym elementem wszystkich algorytmów estymacji częstotliwości podstawowej. Ponieważ pobudzenie w bezdźwięcznych fragmentach mowy ma charakter szumowy, wyznaczanie okresu podstawowego byłoby dla nich bezcelowe i błędne. Konieczne jest w związku z tym dokładne wyznaczenie granic fragmentów dźwięcznych i bezdźwięcznych. Większość metod takiej klasyfikacji bazuje na przetwarzaniu w ramkach [DE93][KA06][RA07]. Ponieważ jednak niektóre ramki mogą częściowo zawierać mowę dźwięczną, a częściowo bezdźwięczną, może to prowadzić do błędów w estymacji częstotliwości podstawowej (zwłaszcza, gdy taka ramka zostanie zaklasyfikowana jako dźwięczna).

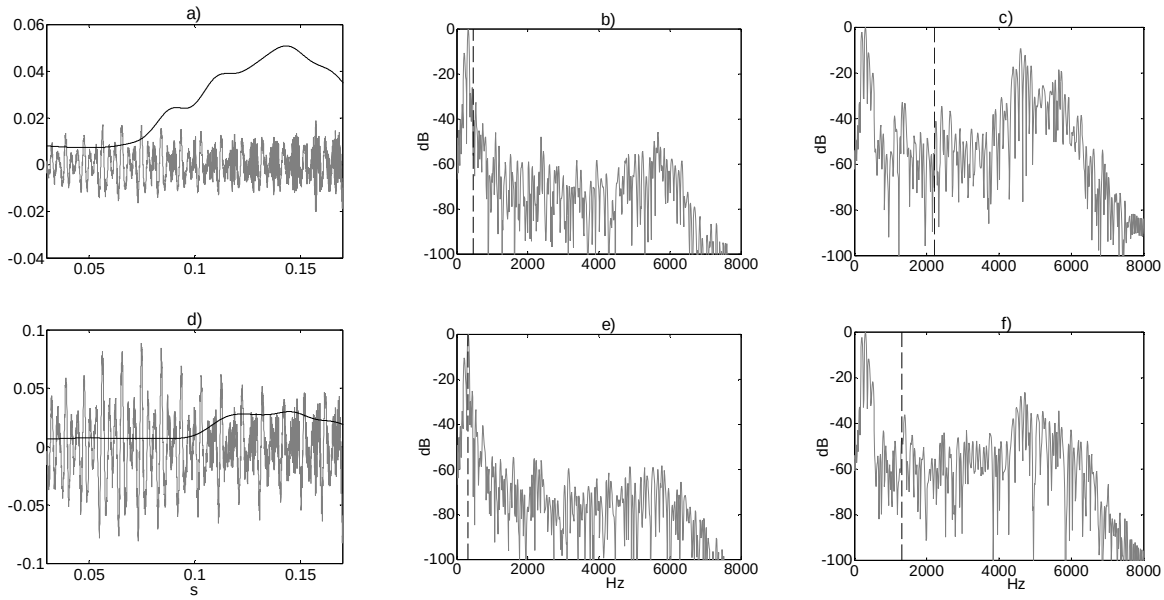
Zaproponowana metoda klasyfikacji mowy na dźwięczną i bezdźwięczną, bazująca na zespolonej pulsacji chwilowej, eliminuje ten problem, gdyż przeprowadza klasyfikację online

tj. z próbki na próbkę. Wejściowy sygnał mowy jest filtrowany za pomocą zespolonego filtru Hilberta o szerokości pasma przepustowego równej 7 kHz i częstotliwości środkowej równej 3.7 kHz, a następnie poddawany jest deemfazie. Dla tak otrzymanego przebiegu $u[n]$ wyznaczana część urojona ICF. Następnie filtrowana jest ona dolnoprzepustowo (filtr o częstotliwości odcięcia 50 Hz) w celu uzyskania jej składowej wolnozmiennej. Przyjmuje ona wartości wyższe dla głosek bezdźwięcznych i niższe dla głosek dźwięcznych. Stosując odpowiedni próg można więc, na podstawie wartości przefiltrowanej IF, dokonać klasyfikacji mowy na dźwięczną i bezdźwięczną. Wartość tego progu została dobrana w trakcie eksperymentów. Rys. 5.15. przedstawia wynik klasyfikacji mowy na dźwięczną i bezdźwięczną dla słowa „*sześć*” oraz zdania „*Czy Krzysio idzie na stadion?*”.



Rys. 5.15. Klasyfikacja mowy na dźwięczną i bezdźwięczną dla słowa „*sześć*” (a) oraz zdania „*Czy Krzysio idzie na stadion?*” (b): przebieg sygnału mowy (kolor szary), przebieg przefiltrowanej dolnopasmowo IF (czerwony), wynik klasyfikacji (czarny) na fragmenty dźwięczne (wartość 1) i bezdźwięczne (wartość 0).

Po omówieniu algorytm klasyfikacji mowy na dźwięczną i bezdźwięczną, wrócimy do wyjaśnienia powodów umieszczenia filtru deemfazy deE na początku dolnej gałęzi algorytmu z rys. 5.14. Konieczność zastosowania deemfazy wynika z obecności w mowie głosek dźwięcznych, w których wraz z pobudzeniem quasi-okresowym występuje również pobudzenie szumowe. Przykłady takich głosek omawialiśmy w podrozdz. 2.3 oraz 5.1. Pokazaliśmy również, że występowanie pobudzenia szumowego objawia się dużym udziałem składowych wysokoczęstotliwościowych w widmie głoski. Można to zaobserwować na rys. 5.16, na którym pokazano przebiegi i periodogramy głoski /z/, wyciętej ze słowa „zero”. W pierwszym wierszu zaprezentowano wykresy uzyskane bez zastosowania deemfazy, w drugim – z zastosowaniem deemfazy.



Rys. 5.16. Klasyfikacja mowy na dźwięczną i bezdźwięczną bez zastosowania deemfazy (pierwszy wiersz) oraz z jej zastosowaniem (drugi wiersz). Lewa kolumna: przebieg głoski /z/ (kolor szary) oraz jej IF (kolor czarny); środkowa kolumna: periodogram fragmentu głoski, w którym udział pobudzenia szumowego jest niewielki; prawa kolumna: periodogram fragmentu głoski /z/, w którym udział pobudzenia szumowego jest znaczący.

Jak można zauważyć, na wykresie 5.16a, głoskę /z/ można podzielić na dwa fragmenty – w pierwszym z nich (trwającym od ok. 0.03 s do 0.08 s) udział pobudzenia szumowego jest niewielki, w drugim jest znaczący. Zaobserwować to można również na periodogramach obu fragmentów, pokazanych, odpowiednio, na rys. 5.16b oraz 5.16c. Na wykresie 5.16c widać bardzo duży udział składowych wysokoczęstotliwościowych (na częstotliwościach powyżej 4000 Hz) w widmie sygnału.Takie wzmocnienie wysokich częstotliwości powoduje również podniesienie wartości IF (średnia wartość IF dla omawianych fragmentów głoski /z/ została zaznaczona na periodogramach pionową linią przerywaną), przez co dźwięczna głoska /z/ klasyfikowana jest jako bezdźwięczna. Po zastosowaniu deemfazy wyższe częstotliwości zostają stłumione, dzięki czemu IF jest przesuwana w dół (rys. 5.16a i 5.16d oraz 5.16c i 5.16f) i głoska jest klasyfikowana poprawnie.

5.2.2. Estymacja prawdopodobnych częstotliwości podstawowych

Częstotliwość podstawowa estymowana jest w pierwszych czterech gałęziach algorytmu z rys. 5.14. Na wyjściu każdej z gałęzi otrzymujemy kandydata na F_0. Wszystkie

te gałęzie zawierają takie same bloki przetwarzania, a różnią się tylko zastosowanym na początku zespolonym filtrem Hilberta $H_{A,F_{Ci}B_i}$ dla i=1,2,3,4. Bank filtrów ma tu dwa zastosowania. Po pierwsze, na wyjściu zespolonego filtru Hilberta otrzymujemy sygnał zespolony

$$u_i[n] = x[n] * h_{Ai}[n] = a_i[n]\exp(j\varphi_i[n]) \tag{5.7}$$

który umożliwia wyznaczenie zespolonej pulsacji chwilowej. W (5.7) $u_i[n]$ to przebieg zespolony na wyjściu i-tego filtru Hilberta, $h_{Ai}[n]$ to odpowiedź impulsowa i-tego filtru Hilberta, a $a_i[n]$ i $\varphi_i[n]$ to odpowiednio: amplituda chwilowa i faza chwilowa przebiegu $u_i[n]$. Drugim powodem zastosowania banku filtrów jest uzyskanie na ich wyjściu przebiegów o odpowiednim paśmie. Jak pokażemy dalej, poprawne wyniki estymacji uzyskuje się, gdy widmo sygnału na wyjściu przynajmniej jednego filtru zawiera dokładnie dwie kolejne harmoniczne. Zasadę tę zastosowano podczas projektowania banku filtrów. Celem było takie wybranie szerokości pasm i częstotliwości środkowych filtrów, by dla częstotliwości podstawowych od 90 Hz do 500 Hz wejściowego sygnału mowy (zakres typowych częstotliwości podstawowych mowy) widmo przebiegu na wyjściu jednego z filtrów zawierało dokładnie dwie harmoniczne (pierwszą i drugą bądź drugą i trzecią). Częstotliwości środkowe i szerokości pasm zaprojektowanych filtrów wyszczególniono w tab. 5.1.

TAB. 5.1. SZEROKOŚCI PASM I CZĘSTOTLIWOŚCI ŚRODKOWE ZASTOSOWANYCH FILTRÓW HILBERTA.

NUMER GAŁĘZI ALGORYTMU	1	2	3	4
CZĘSTOTLIWOŚĆ ŚRODKOWA [HZ]	190	270	390	635
SZEROKOŚĆ PASMA [HZ]	180	260	380	570

Dla każdego przebiegu uzyskanego na wyjściu czterech filtrów wyznaczana jest obwiednia minimalnofazowa $a_{mpi}[n]$ zgodnie z definicją (por. (4.18), (4.35) i (4.36))

$$a_{mpi}[n] = a_i[n]\exp(j\varphi_{mpi}[n]) \tag{5.8}$$

$$a_i[n] = \left|u_i[n]\right| \tag{5.9}$$

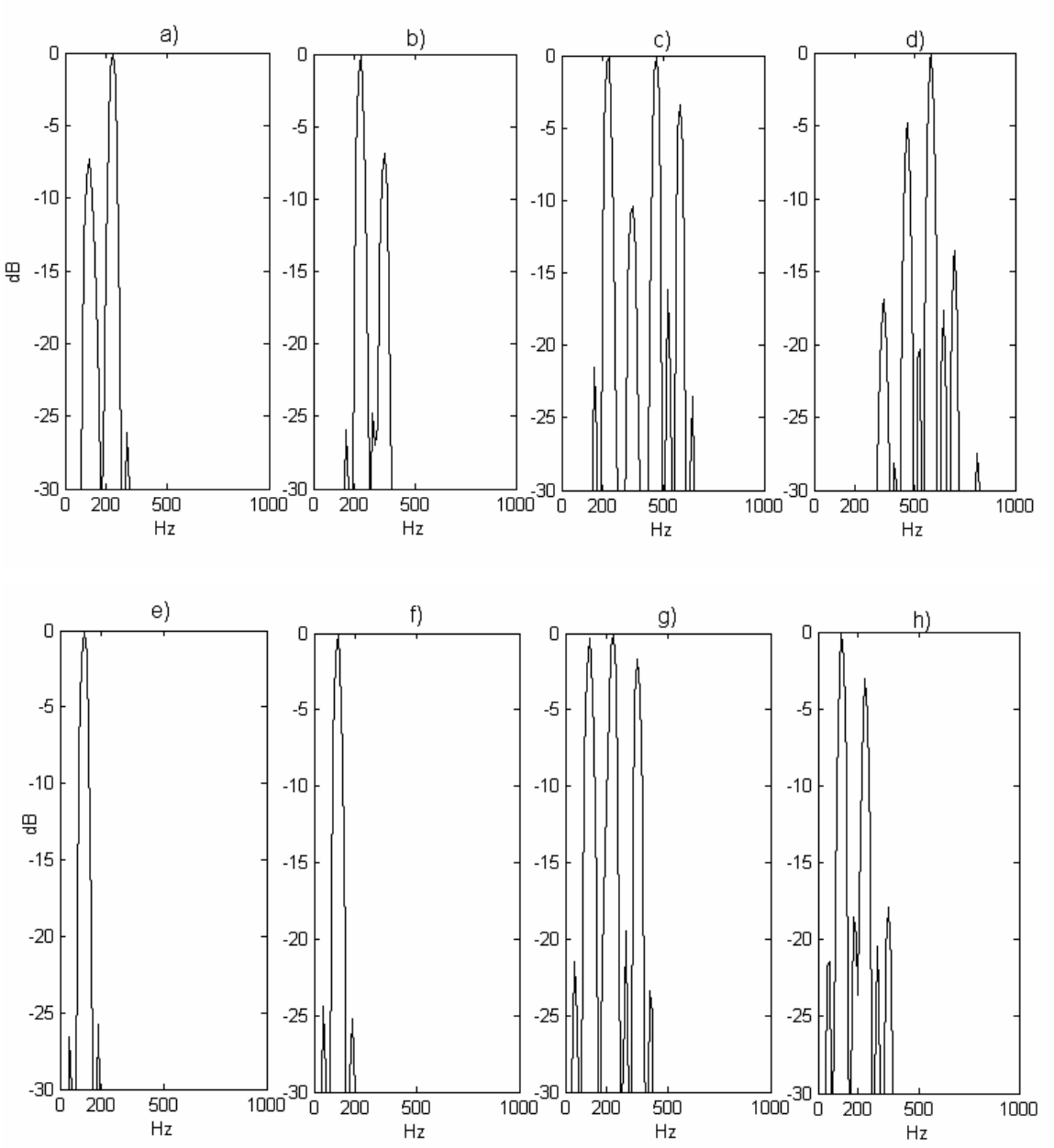
$$\varphi_{mp}[n] = \ln\left(a[n]\right) * h_T[n] \tag{5.10}$$

gdzie $h_T[n]$ jest odpowiedzią impulsową przyczynowego filtru FIR aproksymującego idealny transformator Hilberta, jak poprzednio. Liczba harmonicznych w widmie obwiedni minimalnofazowej $a_{mpi}[n]$ oraz odstępy między nimi są takie same jak dla przebiegów $u_i[n]$ na wyjściach filtrów, z tym że widmo jest przesunięte na osi częstotliwości tak, że pierwszy prążek znajduje się w zerze (DC). Po zastosowaniu filtru wycinającego składową stałą liczba prążków zmniejszy się o jeden, a więc dla przebiegów, których widma zawierały dwie kolejne harmoniczne, uzyskamy widmo jednoprążkowe, o prążku znajdującym się dokładnie na częstotliwości podstawowej sygnału mowy. Na rys. 5.17 pokazano periodogramy przebiegów na wyjściu każdego filtru i periodogramy składowych zmiennych ich obwiedni minimalnofazowych (z usuniętą składową stałą). Wybrany do prezentacji fragment sygnału (głoska /o/) ma częstotliwość podstawową 120 Hz, więc widma przebiegów na wyjściach filtrów H_{A1} i H_{A2} powinny zawierać po dwie harmoniczne na częstotliwościach 120 Hz i 240 Hz (kanał 1) oraz 240 Hz i 360 Hz (kanał 2), natomiast widma ich obwiedni minimalnofazowych powinny zawierać tylko jeden prążek na częstotliwości 120 Hz. Potwierdzają to rysunki 5.17 a, b, e i f.

W kolejnym kroku dla każdego przebiegu $\tilde{a}_{mpi}[n]$ estymowana jest zespolona pulsacja chwilowa, zgodnie ze wzorem [RO94] (por. (4.33))

$$s_{mpi}[n] = \mathrm{Ln}\frac{\tilde{a}_{mpi}[n]}{\tilde{a}_{mpi}[n-1]} = \sigma_i[n] + j\omega_{mpi}[n] \tag{5.11}$$

Części urojone zespolonych pulsacji chwilowych, będące częstotliwościami chwilowymi przebiegów $\tilde{a}_{mpi}[n]$ są kandydatami na estymatę częstotliwości podstawowej. Jak pokazano wcześniej, ich IF oscyluje wokół częstotliwości, na której skumulowana jest moc widma danego przebiegu. Jeżeli widmo przebiegu $\tilde{a}_{mpi}[n]$ zawiera jedną harmoniczną zlokalizowaną na częstotliwości podstawowej, to wyznaczona częstotliwość chwilowa będzie stanowić poprawną estymatę F_0.



Rys. 5.17. Periodogramy przebiegów na wyjściu filtrów Hilberta w kanale 1 (a), 2 (b), 3 (c) i 4 (d) oraz periodogramy składowych zmiennych ich obwiedni minimalnofazowych (e-h).

Błędną estymatę otrzymamy, gdy widmo przebiegu $\tilde{a}_{mpi}[n]$ nie zawiera żadnych harmonicznych, natomiast gdy jest ich więcej niż jedna, to estymata może być nadal poprawna, gdy $\tilde{a}_{mpi}[n]$ jest przebiegiem minimalnofazowym (o dominującym pierwszym prążku w widmie).

Wykorzystanie banku zespolonych filtrów Hilberta w algorytmie estymacji częstotliwości podstawowej zaproponowali w swojej pracy Blok i in. [BL04]. Koncepcję tę stosował również Zawidzki [ZA10] uzupełniając algorytm o blok wyznaczania obwiedni minimalnofazowej. Jednakże bank filtrów zastosowany w obu powyższych rozwiązaniach był inaczej zaprojektowany. W pracy Bloka i in. było to pięć filtrów półoktawowych pokrywających zakres od 50 Hz do $200\sqrt{2}$ Hz. Dla tak dobranego banku filtrów przebieg na wyjściu co najmniej jednego z nich zawiera tylko jedną składową częstotliwościową analizowanego sygnału mowy. Zawidzki skorzystał natomiast z czterech filtrów 5/6-oktawowych, pokrywających pasmo telefoniczne. W tym rozwiązaniu jednak nie zawsze jest spełnione, że przebieg na wejściu estymatora ICF w co najmniej jednej gałęzi algorytmu zawiera tylko jedną harmoniczną sygnału mowy. Natomiast zaproponowany tu bank filtrów jest najbliższy optymalnemu, tzn. przy użyciu najmniejszej liczby filtrów, a więc najmniejszej liczby gałęzi analizy, pokrywa cały zakres częstotliwości podstawowych mowy i gwarantuje, że przebieg na wejściu estymatora IF w co najmniej jednej gałęzi będzie zawierał tylko jedną harmoniczną, dokładnie na częstotliwości podstawowej analizowanego sygnału mowy.

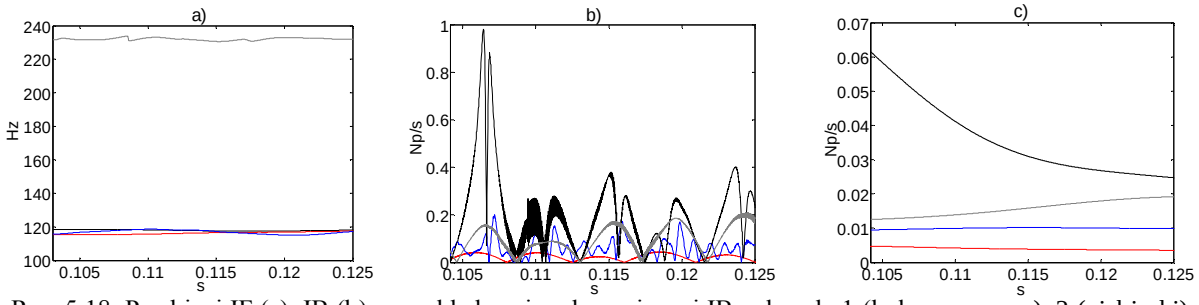
5.2.3. Wybór poprawnej estymaty częstotliwości podstawowej

Przetwarzanie końcowe zaproponowanego algorytmu estymacji częstotliwości podstawowej obejmuje blok decyzyjny oraz filtr uśredniający (rys. 5.14). W bloku decyzyjnym spośród czterech prawdopodobnych estymat częstotliwości podstawowej wybierana jest jedna. Decyzja podejmowana jest dla każdej próbki sygnału na podstawie wartości części rzeczywistych ICF, wyznaczonych w każdej z czterech gałęzi, które są chwilowymi szerokościami pasma IB [CO95]. Jak pokazały eksperymenty, prawdopodobieństwo poprawnej estymacji częstotliwości podstawowej jest większe dla gałęzi, w których IB ma niższą wartość. W [RO10] M. Rojewski zaproponował, by decyzję o wyborze poprawnej estymaty częstotliwości podstawowej podejmować na podstawie znormalizowanej wartości IB

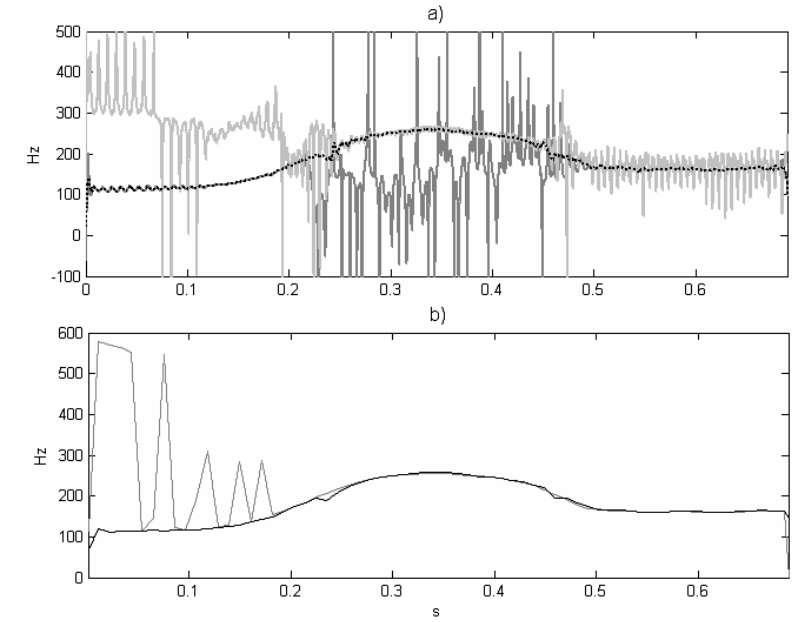
$$\mu_i[n] = \frac{|\sigma_i[n]|F_S}{2\pi B_i} \tag{5.12}$$

gdzie F_S jest częstotliwością próbkowania, $|\sigma_i[n]|$ to IB przebiegu $\tilde{a}_{mpi}[n]$, a B_i jest szerokością pasma filtru Hilberta w i-tej gałęzi algorytmu. Za poprawną uznawana jest estymata otrzymana w gałęzi, w której składowa wolnozmienna $\mu_i[n]$ (po przejściu przez filtr dolnoprzepustowy o częstotliwości odcięcia 50 Hz) ma najmniejszą wartość. Rys. 5.18 przedstawia przebiegi IF oraz znormalizowanych IB wyznaczonych dla fragmentu głoski /o/. Przebiegi z rys. 5.18 wskazują, że poprawną estymatę częstotliwości podstawowej zapewniają kanały 1, 2 i 4. Porównując te wyniki z periodogramami przedstawionymi na rys. 5.17 można zauważyć, że w tych kanałach widma obwiedni minimalnofazowych po usunięciu składowej stałej zawierają tylko jedną harmoniczną (kanał 1 i 2) lub mają dominujący pierwszy prążek (kanał 4). Przebieg w kanale 3 nie spełnia żadnego z tych warunków i uzyskana w nim estymata F_0 jest błędna. Z przebiegów części wolnozmiennych IB (rys. 5.18c) wynika, że dla opisywanego przykładu w bloku decyzyjnym jako poprawna wybrana zostanie estymata z kanału 1.

Otrzymany w ostatnim kroku przebieg estymaty częstotliwości podstawowej jest wygładzany za pomocą filtru uśredniającego FIR o długości 256 próbek i mnożony przez czynnik $F_S / 2\pi$, by uzyskać wartość w hercach. Rys. 5.19a przedstawia przebieg F_0 dla głoski /o/ wypowiadanej ze zmieniającą się (rosnącą, a następnie znów opadającą) częstotliwością podstawową. F_0 jest estymowana poprawnie w kanale pierwszym i trzecim przy odpowiednio niższej i wyższej częstotliwości podstawowej mowy. Na rys. 5.19a zamieszczono również przebiegi IF estymowanych w tych dwóch kanałach. Na rys. 5.19b umieszczono dla porównania przebieg F_0 estymowanej za pomocą algorytmu YIN [CH02][KA06][KA07], w którym wystąpiły typowe dla algorytmów estymacji częstotliwości podstawowej błędy oktawowe.



Rys. 5.18. Przebiegi IF (a), IB (b) oraz składowej wolnozmiennej IB w kanale 1 (kolor czerwony), 2 (niebieski), 3 (szary) i 4 (czarny).



Rys. 5.19. Wynik estymacji częstotliwości podstawowej dla głoski /o/: a) przebieg estymaty F_0 z kanału pierwszego (kolor ciemny szary) i trzeciego (jasny szary) oraz na wyjściu algorytmu (czarny, linia przerywana); b) przebieg F_0 estymowanej za pomocą algorytmu YIN (kolor szary) i zaproponowanego algorytmu potokowego (czarny)

Cząstkowe wyniki estymacji F_0 za pomocą zaproponowanego algorytmu opisane zostały w artykułach konferencyjnych [KA10c][KA11].

5.2.4. Eksperymenty

W eksperymentach, których wyniki przedstawiamy poniżej, zbadano dokładność estymacji częstotliwości podstawowej oraz poprawność klasyfikacji mowy na dźwięczną i bezdźwięczną. W testach wykorzystano dwie bazy nagrań dostępne w Internecie pod adresem ftp://ftp.cs.keele.ac.uk/pub/pitch/Speech oraz http://data.cstr.ed.ac.uk/mocha/. Pierwsza z nich (DB1) składa się z dziesięciu nagrań tych samych wypowiedzi w języku angielskim, nagranych przez pięć kobiet i pięciu mężczyzn w języku angielskim. Wypowiedzi mają średnią długość 30 s i nagrane są w formacie WAV z szybkością próbkowania 48000 Sa/s. Druga baza (DB2) obejmuje nagrania głosów trzech kobiet i trzech mężczyzn, wypowiadających po 460 zdań (wszystkie osoby wypowiadają te same zdania, również w języku angielskim). Szybkość próbkowania i format nagrań są takie same, jak w bazie DB1.

Dodatkowo w obu bazach znajdują się zarejestrowane jednocześnie z sygnałem mowy zapisy z laryngografu (aparat mierzący rezystancję elektryczną między elektrodami umieszczonymi po obu stronach krtani). Nagrania te, zapisane w tym samym formacie jak sygnał mowy, stanowią dodatkowe przebiegi referencyjne dla testów. Dokładne opisy zawartości i sposobu przygotowania obu baz można znaleźć w literaturze [PL95] [WR00].

5.2.4.1. Ocena poprawności klasyfikacji mowy na dźwięczną i bezdźwięczną

Ocenę algorytmu opisanego w p. 5.2.3 rozpoczęto od sprawdzenia poprawności klasyfikacji mowy na dźwięczną i bezdźwięczną. Każdy wynik klasyfikacji porównano z odpowiednim przebiegiem referencyjnym, w którym, w przypadku idealnym, fragmenty dźwięczne i bezdźwięczne mowy byłyby oznaczone bezbłędnie. Ponieważ klasyczne metody klasyfikacji przeprowadzają analizę mowy w ramkach, podczas gdy proponowany w pracy algorytm działa potokowo, próbka po próbce, zdecydowano, że przebieg referencyjny zostanie wyznaczony ręcznie poprzez oznaczenie początków i końców fragmentów dźwięcznych i bezdźwięcznych na podstawie oscylogramu sygnału mowy. Dodatkowo, takie podejście zapewnia, że przebieg referencyjny zostanie oznaczony poprawnie, podczas gdy wykorzystanie automatycznego algorytmu klasyfikacji zawsze wprowadza pewne błędy. Ponieważ jednak nie można wyznaczyć konkretnej próbki, w której pobudzenie zmienia się z okresowego na szumowe, a przejście z głoski dźwięcznej na bezdźwięczną lub odwrotnie jest płynne, próbki wokół postawionych znaczników zostały wyłączone z testów, tzn. odrzucono je przy porównywaniu wyniku klasyfikacji zastosowanego tu algorytmu z przebiegiem referencyjnym. Liczbę odrzuconych próbek ustalono na 256 wokół każdego znacznika – jest to połowa szerokości ramki, którą standardowo stosuje się w analizie sygnału mowy próbkowanego z szybkością 48000 Sa/s w klasycznych metodach klasyfikacji [RA07]. Do oceny poprawności klasyfikacji wykorzystano tylko bazę DB1. Wyniki eksperymentów zamieszczono w tab. 5.2. W kolejnych wierszach tab. 5.2 zamieszczono wyniki dla nagrań kobiet (k1-k5) i mężczyzn (m1-m5) oraz zbiorcze wyniki dla wszystkich kobiet (K), wszystkich mężczyzn (M) i wszystkich nagrań (K+M). W kolejnych kolumnach znajdują się: czas wypowiedzi w próbkach, błędy zaklasyfikowania mowy dźwięcznej jako bezdźwięcznej (fUV – ang. *false unvoiced*) wyrażone w liczbie próbek i w procentach, błędy

zaklasyfikowania mowy bezdźwięcznej jako dźwięcznej (fV – ang. *false voiced*) wyrażone w liczbie próbek i w procentach, wszystkie błędy klasyfikacji (fUV+fV) wyrażone w procentach oraz poprawność klasyfikacji wyrażona w procentach.

TAB. 5.2. WYNIKI OCENY POPRAWNOŚCI KLASYFIKACJI MOWY NA DŹWIĘCZNĄ I BEZDŹWIĘCZNĄ

	liczba próbek	błędy fUV	błędy fUV [%]	błędy fV	błędy fV [%]	fUV+fV [%]	poprawność klasyfikacji [%]
k1	1352678	16584	*1,2*	118642	*8,8*	***10,0***	***90,0***
k2	1422669	45357	*3,2*	68827	*4,8*	***8,0***	***92,0***
k3	1291029	36347	*2,8*	74447	*5,8*	***8,6***	***91,4***
k4	1336872	16063	*1,2*	35237	*2,6*	***3,8***	***96,2***
k5	1629092	23068	*1,4*	40298	*2,5*	***3,9***	***96,1***
K	**7032340**	**137419**	***2,0***	**337451**	***4,8***	***6,8***	***93,2***
m1	1584205	124805	*7,9*	48317	*3,0*	***10,9***	***89,1***
m2	1341165	52638	*3,9*	93742	*7,0*	***10,9***	***89,1***
m3	1145243	88621	*7,7*	10973	*1,0*	***8,7***	***91,3***
m4	1428813	139998	*9,8*	16671	*1,2*	***11,0***	***89,0***
m5	1716160	137753	*8,0*	7650	*0,4*	***8,5***	***91,5***
M	**7215586**	**543815**	***7,5***	**177353**	***2,5***	***10,0***	***90,0***
K+M	**14247926**	**681234**	***4,8***	**514804**	***3,6***	***8,4***	***91,6***

Jak wynika z tab. 5.2 poprawność klasyfikacji mowy dźwięcznej i bezdźwięcznej proponowanego algorytmu przekracza 91% i jest wyższa dla nagrań głosów żeńskich (92.7%) niż męskich (90.0%). Warto również zauważyć, że dla wszystkich głosów żeńskich wyższy jest procent błędów zaklasyfikowania głosek bezdźwięcznych jako dźwięcznych, podczas gdy dla głosów męskich jest odwrotnie (z wyjątkiem nagrania m2). Przyczyny uzyskania takich wyników wyjaśnimy przy omawianiu konkretnych rodzajów błędów klasyfikacji. Pierwszy z nich występuje przy przejściu z głoski bezdźwięcznej na dźwięczną i odwrotnie. Jak już zaznaczaliśmy wcześniej, przejście takie jest płynne. W stanach przejściowych przebieg IF, na którym opiera się działanie klasyfikatora, powoli narasta lub opada, a zmiana wyniku klasyfikacji następuje po osiągnięciu przez IF wartości zadanego progu. Podczas eksperymentów założono, że 256 próbek wokół znacznika zmiany rodzaju pobudzenia odpowiada stanowi przejściowemu i, jak wcześniej zaznaczono, jest wyłączonych z porównania. Jednak często czas narastania lub opadania wartości IF do zadanego progu jest

dłuższy, co powoduje błędną klasyfikację. Taki rodzaj błędów jest bardziej znaczący w przypadku głosów żeńskich.

Drugi rodzaj błędów zaklasyfikowania mowy bezdźwięcznej jako dźwięcznej występuje, gdy w widmie sygnału nieokresowego składowe niskoczęstotliwościowe (do 1000 Hz) mają wyraźnie wyższą amplitudę, co powoduje przesunięcie wartości IF w dół, poniżej zadanego progu klasyfikacji. Tego rodzaju błędy pojawiały się w równym stopniu dla głosów męskich i żeńskich.

Pozostałe dwa błędy to błędy zaklasyfikowania mowy dźwięcznej jako bezdźwięcznej. Pierwszy z nich występuje wyłącznie dla głosów, które charakteryzują się niższą częstotliwością podstawową i wynika głównie z charakterystyki zastosowanego filtru Hilberta. Mianowicie, gdy F_0 jest niższa niż 200 Hz, to po zastosowania filtru Hilberta (o dolnej częstotliwości odcięcia 200 Hz), prążek o częstotliwości podstawowej zostaje prawie całkowicie wycięty, a kolejny prążek zostaje znacznie stłumiony. Z tego powodu różnica amplitud składowych nisko- i wysokoczęstotliwościowych w widmie znacznie się zmniejsza. Powoduje to przesunięcie wartości IF w górę ponad próg klasyfikacji i błędne zaklasyfikowanie mowy jako bezdźwięcznej.

Ostatnim rodzajem błędów klasyfikacji, jaki zaobserwowano podczas analizy wyników, jest zaklasyfikowanie jako bezdźwięcznej głoski dźwięcznej, w której wraz z pobudzeniem okresowym występuje pobudzenie szumowe. Problem klasyfikacji takich głosek omawiany był już w p. 5.2.1. Błędy te zostały w dużym stopniu wyeliminowane przez zastosowanie deemfazy w algorytmie klasyfikacji. Jak pokazała analiza wyników, wystąpiły one tylko kilkakrotnie dla całej bazy DB1, wyłącznie dla głosów męskich, co można wytłumaczyć, analogicznie do wcześniej omawianego rodzaju błędów, niską częstotliwością podstawową mowy dla głosów męskich oraz charakterystyką zastosowanego filtru Hilberta.

Analiza błędów klasyfikacji pokazała, że dla głosów kobiecych (o wyższej częstotliwości podstawowej) występuje więcej błędów zaklasyfikowania mowy bezdźwięcznej jako dźwięcznej, wynikających z powolnego narastania wartości IF w stanach przejściowych, niż dla głosów męskich. Jednocześnie dla tych nagrań nie występują dwa ostatnie z opisywanych rodzaje błędów zaklasyfikowania mowy dźwięcznej jako bezdźwięcznej. Stąd dla głosów żeńskich (oraz dla nagrania m2, w którym częstotliwość podstawowa jest wyższa niż dla pozostałych głosów męskich) większa jest liczba błędów fUV

niż fV, a dla głosów męskich jest odwrotnie. Jednocześnie z analizy błędów można wywnioskować, że zastosowana deemfaza przyczynia się do powstawania niektórych błędów klasyfikacji. Należy jednak zaznaczyć, że została ona wprowadzone by wyeliminować inne błędy klasyfikacji i, jak pokazały eksperymenty, jej pominięcie prowadziłoby do globalnego pogorszenia poprawności klasyfikacji.

5.2.4.2. Ocena wyników estymacji częstotliwości podstawowej

Dla sprawdzenia poprawności estymacji częstotliwości podstawowej wykorzystano dwie wspomniane wcześniej bazy, DB1 i DB2, które zawierają nie tylko nagrania mowy, ale również przebieg zarejestrowany przez laryngograf, przy czym sygnały zarejestrowane przez mikrofon i laryngograf są ze sobą zsynchronizowane. Referencyjne przebiegi częstotliwości podstawowej do testów uzyskano estymując F_0 zarówno nagrań mowy, jak i sygnałów z laryngografu za pomocą algorytmu YIN [CH02][KA06][KA07] w ramkach o szerokości 512 próbek. W testach wykorzystano wyłącznie ramki, dla których estymaty dla obu sygnałów różnią się od siebie nie więcej niż o 20% wartości F_0 estymowanej dla sygnału z laryngografu. Dla tych ramek przyjęto, że częstotliwość podstawowa została wyznaczona poprawnie i może być uznana za wartość prawdziwą. Pozostałe ramki zostały wyłączone z testów. Z tak uzyskanym przebiegiem porównywano estymatę częstotliwości podstawowej obliczoną za pomocą proponowanego algorytmu potokowego. Ponieważ algorytm YIN przetwarza mowę w ramkach, przebieg częstotliwości podstawowej z wyjścia algorytmu potokowego, w którym F_0 obliczane jest dla każdej próbki, musiał być przetworzony tak, by porównanie było możliwe. W tym celu przebieg na wyjściu algorytmu potokowego podzielono na ramki o szerokości 512 próbek, a następnie dla każdej ramki obliczono wartość średnią F_0.

Zgodnie z przyjętą powszechnie konwencją błędy estymacji podzielono na dwie kategorie:

- błędy grube, gdy wartość estymowana różni się od wartości rzeczywistej o więcej niż 20% (wartość 20% przyjmowana jest przez dużą część autorów prac dotyczących estymacji częstotliwości podstawowej [CH02][SH09][MA05] [YE09]).

- błędy drobne, które określają dokładność estymacji częstotliwości podstawowej w ramkach, w których wartość estymowana różni się od wartości rzeczywistej nie więcej niż o 20%.

Błędy grube (GE od ang. *Gross Error*) obliczono jako stosunek liczby ramek dźwięcznych, w których estymata różni się od wartości rzeczywistej o więcej niż 20%, do liczby wszystkich ramek dźwięcznych i wyrażono w procentach. Dodatkowo policzono liczbę ramek, w których estymata różni się od wartości rzeczywistej o więcej niż 5% i 1%. Dla oszacowania błędów drobnych wykorzystano definicję błędu średniokwadratowego – RMS wyrażonego w hercach oraz w procentach

$$\mathrm{RMS[Hz]} = \sqrt{\frac{\sum_{m=1}^{M}\left(F_0^r(m) - F_0^{est}(m)\right)^2}{M}} \tag{5.13}$$

$$\mathrm{RMS[\%]} = \sqrt{\frac{\sum_{m=1}^{M}\left(\frac{F_0^r(m) - F_0^{est}(m)}{F_0^r(m)}\right)^2}{M}} \cdot 100\% \tag{5.14}$$

We wzorach (5.13) oraz (5.14) M jest liczbą ramek dźwięcznych, w których różnica między wartością estymaty a faktyczną wartością F_0 nie przekroczyła 20% (nie wystąpiły błędy grube), m to numer kolejnej ramki, F_0^r jest wartością prawdziwą F_0, a F_0^{est} – wartością estymowaną. Wyniki uzyskane dla baz DB1 i DB2 przedstawiono w tab. 5.3 oraz tab. 5.4. W kolejnych wierszach tabel zamieszczono wyniki dla nagrań kobiet (k1-k5 dla bazy DB1 i k6-k8 dla bazy DB2) i mężczyzn (m1-m5 i m6-m8) oraz zbiorcze wyniki dla wszystkich kobiet (K), wszystkich mężczyzn (M) i wszystkich nagrań (K+M). W kolejnych kolumnach znajdują się: średnia częstotliwość podstawowa obliczona dla poszczególnych mówców ($\overline{F_0}$), liczba ramek dźwięcznych (VFN od ang. *Voiced Frames Number*), błąd gruby (GE) wyrażony w procentach, odsetek ramek, w których estymata różni się od wartości rzeczywistej o więcej niż 5% oraz więcej niż 1%, a także błąd RMS wyrażony w hercach i w procentach. W powyższych wynikach pominięto ramki dźwięczne zaklasyfikowane jako bezdźwięczne, gdyż ocena poprawności klasyfikacji przeprowadzona została oddzielnie, a na tym etapie istotne było sprawdzenie poprawności samej estymacji częstotliwości podstawowej.

TAB. 5.3. WYNIKI OCENY POPRAWNOŚCI ESTYMACJI CZĘSTOTLIWOŚCI PODSTAWOWEJ DLA DB1

	$\overline{F_0}$ [Hz]	VFN	GE [%]	E5[%]	E1 [%]	RMS [Hz]	RMS [%]
k1	217	1099	*7,8*	*24,3*	*71,8*	10,9	*4,9*
k2	253	1510	*7,4*	*22,8*	*71,5*	11,9	*4,6*
k3	215	1148	*9,8*	*27,3*	*73,7*	11	*5,1*
k4	263	1174	*12,2*	*28,4*	*75,1*	12,2	*4,8*
k5	254	1427	*5,5*	*19,3*	*73,3*	11,5	*4,4*
K		**6358**	***8,3***	***24,1***	***73,0***	**11,5**	***4,7***
m1	142	539	*8,3*	*43,2*	*82,9*	9,1	*6,9*
m2	160	943	*10,5*	*39,7*	*84,1*	9,3	*6*
m3	156	950	*9,2*	*33,8*	*78,3*	8,6	*5,4*
m4	137	369	*36,6*	*73,2*	*91,1*	11,6	*8,9*
m5	142	571	*9,3*	*46,1*	*83,9*	10,4	*7,6*
M		**3372**	***12,4***	***43,3***	***83,0***	**9,5**	***6,6***
K+M		**9730**	***9,8***	***30,8***	***76,5***	**10,9**	***5,4***

TAB. 5.4. WYNIKI OCENY POPRAWNOŚCI ESTYMACJI CZĘSTOTLIWOŚCI PODSTAWOWEJ DLA DB2

	$\overline{F_0}$ [Hz]	VFN	GE [%]	E5[%]	E1 [%]	RMS [Hz]	RMS [%]
k6	196	129528	*3,3*	*8,2*	*19,9*	10,9	*6*
k7	207	159247	*5,8*	*11,2*	*26,2*	10,3	*5,3*
k8	189	133668	*4,1*	*9,8*	*32,3*	9,7	*5,5*
K		**422443**	***4,5***	***9,8***	***26,2***	**10,3**	***5,6***
m6	122	110033	*4,8*	*28,4*	*38,7*	11,6	*9,6*
m7	154	174406	*2,7*	*18,8*	*34,9*	9,5	*6,9*
m8	147	145327	*1,7*	*16,2*	*24,8*	10,8	*8,3*
M		**429766**	***2,9***	***20,4***	***32,4***	**10,6**	***8,3***
K+M		**852209**	***3,7***	***15,2***	***29,4***	**10,4**	***7,2***

Jak pokazują powyższe tabele, błąd gruby estymacji wynosi 9.8% dla DB1 i 3.7% dla DB2. Analizując wyniki estymacji zauważono, że część błędów grubych wynika z występowania stanów przejściowych między głoskami dźwięcznymi i bezdźwięcznymi. Dla tych fragmentów sygnału częstotliwość F_0 estymowana potokowo narasta lub maleje stopniowo, próbka po próbce. Brana do porównania średnia częstotliwośći F_0 w ramce, w której występuje stan przejściowy, nie jest więc równa ani zeru, ani wartości, która wyznaczona by była dla samej głoski dźwięcznej, a znajduje się pomiędzy nimi. Natomiast algorytm YIN klasyfikuje ramkę ze stanem przejściowym w całości jako dźwięczną lub bezdźwięczną, stąd rozbieżności pomiędzy wartością estymowaną i referencyjną F_0. Jeśli z porównania wyeliminujemy ramki zawierające stany przejściowe, to błąd gruby estymacji zmniejszy się do 2.5% dla DB1 oraz 1.6% dla DB2. Większość z tych błędów wynika z

wyboru złego kanału w bloku decyzyjnym algorytmu potokowego, gdyż nie zawsze spełnione jest, że prawdopodobieństwo poprawnej estymacji częstotliwości podstawowej jest większe dla gałęzi, w których IB ma mniejszą wartość. Niewielki procent błędów jest efektem przyjętego założenia, że widmo przebiegu na wyjściu jednego z filtrów w gałęziach 1-4 zawiera dokładnie dwie harmoniczne oryginalnego sygnału mowy (pierwszą i drugą bądź drugą i trzecią). Jak pokazały przeprowadzone testy, w widmie sygnału mowy mogą nie występować niektóre harmoniczne (lub są silnie stłumione), przez co F_0 jest estymowana błędnie w kanale, w którym estymata powinna być poprawna. Nie występują tu natomiast błędy oktawowe typowe dla algorytmów estymacji częstotliwości podstawowej. Ponadto zaletą proponowanego algorytmu jest brak ograniczenia zakresu estymowanych częstotliwości podstawowych. W rozwiązaniach tradycyjnych, bazujących na przetwarzaniu ramek, zakres częstotliwości, które mogą być estymowane poprawnie, jest ograniczony przez szerokość ramki (poza tym zakresem poprawność estymacji znacząco się obniża). W proponowanym algorytmie wystarczy dołożyć kolejną gałąź analizy, by rozszerzyć zakres estymowanych częstotliwości.

5.3. Ekstrakcja formantów mowy

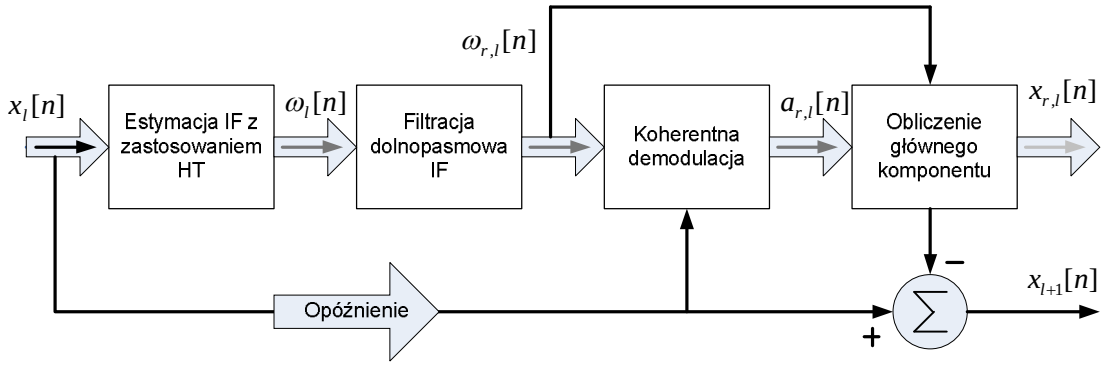
W niniejszym podrozdziale zostanie przedstawiona metoda ekstrakcji formantów mowy, bazująca na algorytmie dekompozycji sygnałów wielokomponentowych zaproponowanym przez Feldmana [FE06][FE08][FE11][BR11]. Estymacja częstotliwości środkowych, jak również szerokości pasm formantów znajduje zastosowanie w wielu aplikacjach takich jak kodowanie czy rozpoznawanie mowy [DE93][RA07]. Najczęściej dla znajdowania formantów mowy stosuje się metody oparte na LP i cepstum, o których pisaliśmy w podrozdz. 2.4. Algorytmy te wykorzystują model „źródło-filtr” i przetwarzają sygnał mowy w ramkach. Opracowana metoda ekstrakcji formantów jest algorytmem potokowym (działającym próbka po próbce) i pozwala na wyodrębnienie z sygnału mowy pojedynczych formantów jako oddzielnych przebiegów, z wykorzystaniem faktoryzacji V-K oraz zespolonej pulsacji chwilowej i wyznaczenie ich częstotliwości środkowych oraz szerokości pasm na podstawie ICF. Zaproponowany dalej algorytm jest modyfikacją opracowanej przez Feldmana

metody dekompozycji sygnałów, będących superpozycją quasi-harmonicznych funkcji, na pojedyncze składowe.

W podrozdz. 3.3 przedstawiono zagadnienia dotyczące sygnałów wielokomponentowych i ich częstotliwości chwilowej. Sygnały takie są sumą dwóch lub więcej sygnałów monokomponentowych. W modelu mowy zaproponowanym przez Maragosa i in. [HA94] [MA95] (o którym pisaliśmy w podrozdz. 2.4) sygnał mowy modeluje się właśnie jako sygnał wielokomponentowy, którego składowymi są poszczególne formanty. Każdy formant traktowany jest jako wąskopasmowy, jednokomponentowy sygnał zmodulowany amplitudowo i częstotliwościowo. Szerokości pasm formantów są na ogół mniejsze niż odstępy między nimi na osi częstotliwości – spełniają warunek zdefiniowany przez Cohena [CO95] dla sygnału wielokomponentowego. Wybrane metody, które umożliwiają dekompozycję sygnałów wielokomponentowych (wyodrębnienie poszczególnych komponentów), w tym mowy, przedstawiliśmy w p. 3.3.1. Metodę Feldmana HVD (ang. *Hilbert Vibration Decomposition*) wyróżnia spośród nich to, że wykorzystuje jedynie proste algorytmy przetwarzania sygnałów, zapewniające małą złożoność obliczeniową, skąd wynika możliwość jej stosowania w systemach czasu rzeczywistego. Oryginalny algorytm HVD nie może być wykorzystany wprost do dekompozycji sygnału mowy, ale w kolejnym podrozdziale pokażemy sposób jego modyfikacji tak, by możliwe było jego zastosowanie dla ekstrakcji formantów.

5.3.1. Metoda Feldmana dekompozycji sygnałów wielokomponentowych

HVD jest algorytmem iteracyjnym – liczba kroków jest równa liczbie komponentów w analizowanym sygnale. Rys. 5.20 przedstawia poglądowo schemat blokowy *l*-tej iteracji algorytmu dekompozycji sygnału wielokomponentowego zaproponowanego przez Feldmana [BR11]. Każda iteracja składa się z dwóch zasadniczych etapów: estymacji IF komponentu o największej mocy, a następnie wyznaczenia obwiedni tego komponentu. Wyodrębniony w ten sposób komponent jest następnie odejmowany od sygnału podanego na wejście algorytmu. Tak uzyskany przebieg stanowi wejście algorytmu w następnym kroku iteracyjnym.



Rys. 5.20. Schemat blokowy *l*-tej iteracji algorytmu HVD [BR11].

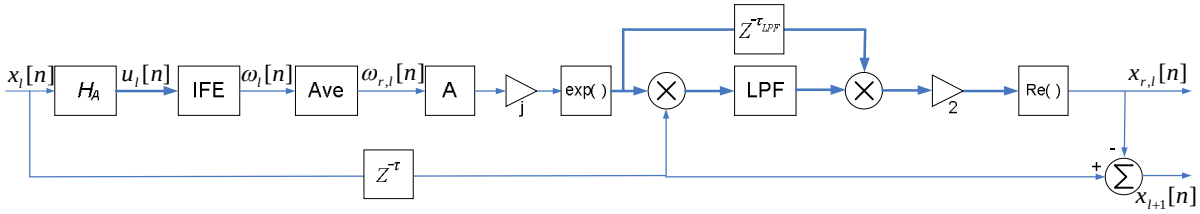
Estymacja IF komponentu o największej mocy opiera się na założeniu, że IF sygnału wielokomponentowego składa się z części wolno- i szybkozmiennej (powolnej i szybkiej), i że część powolna odpowiada częstotliwości chwilowej szukanego komponentu [FE06] [BR11]. Aby wyodrębnić część wolnozmienną należy przefiltrować IF za pomocą filtru uśredniającego, wygładzającego lub dolnoprzepustowego. Uzyskana w ten sposób wolnozmienna IF jest częstotliwością chwilową $\omega_{r,l}(t)$ głównego komponentu w sygnale w *l*-tej iteracji. Na jej podstawie wyznaczana jest następnie obwiednia odpowiadająca temu komponentowi. W tym celu stosowana jest demodulacja koherentna znana z techniki radiowej, tzn. sygnał wejściowy mnożony jest przez dwa sygnały różniące się w fazie o 90°: $\cos(\int_t \omega_{r,l}(t)dt)$ oraz $\sin(\int_t \omega_{r,l}(t)dt)$. Każdy z uzyskanych w ten sposób przebiegów jest następnie filtrowany dolnoprzepustowo, by wyodrębnić ich części wolnozmienne. Amplituda chwilowa $a_{r,l}(t)$ poszukiwanego komponentu $x_{r,l}(t) = a_{r,l}(t)\cos\left(\int_t \omega_{r,l}(t)dt\right)$ wyznaczana jest jako pierwiastek z sumy kwadratów amplitud dwóch sygnałów na wyjściu demodulatora.

5.3.2. Adaptacja metody HVD dla analizy mowy

Założeniem dekompozycji metodą HVD jest by analizowany sygnał był superpozycją quasi-harmonicznych funkcji, by obwiednia każdej składowej różniła się od pozostałych oraz by czas trwania każdej składowej był co najmniej tak długi, jak kilka okresów najwolniejszego komponentu [FE06][BR11]. Zauważmy, że sygnał mowy zamodelowany jako superpozycja sygnałów jednokomponentowych, spełnia te założenia dla mowy dźwięcznej.

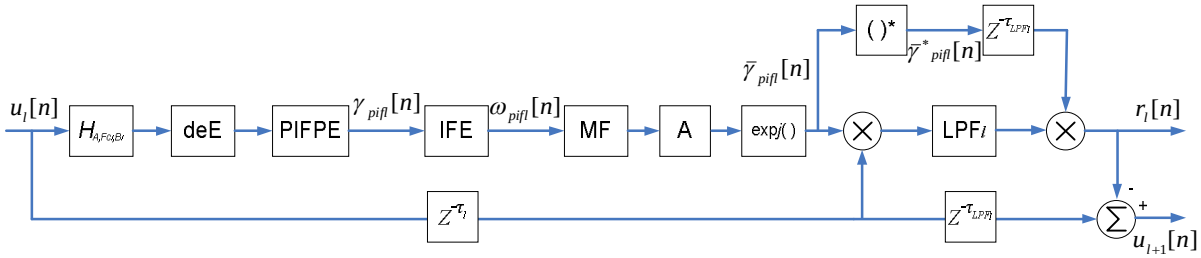
Aby zmodyfikować algorytm HVD należy przyjrzeć się dokładnie zasadzie jego działania. W pierwszym kroku estymowana jest częstotliwość chwilowa sygnału, a następnie filtrowana dolnoprzepustowo w celu wyodrębnienia części wolnozmiennej. Uzyskany w ten sposób przebieg $\omega_{r,l}(t)$ jest częstotliwością chwilową głównego komponentu w sygnale. Następnie komponent ten jest wyodrębniany z sygnału poprzez znalezienie jego amplitudy i fazy chwilowej. Amplituda wyznaczana jest jako pierwiastek z sumy kwadratów amplitud dwóch sygnałów uzyskanych przez wymnożenie sygnału $x_l(t)$ na wejściu algorytmu przez $\cos(\int_t \omega_{r,l}(t)dt)$ i $\sin(\int_t \omega_{r,l}(t)dt)$, a faza jako arcustangens stosunku tych dwóch sygnałów. Zauważmy, że oznacza to tyle samo, co wyznaczenie amplitudy i fazy wskazu $x_l(t)\left(\cos(\int_t \omega_{r,l}(t)dt) + j\sin(\int_t \omega_{r,l}(t)dt)\right) = x_l(t)\exp(j\int_t \omega_{r,l}(t)dt)$. To wymnożenie $x_l(t)$ przez zespoloną sinusoidę oznacza przesunięcie widma sygnału $x_l(t)$ na osi częstotliwości w prawo o wartość $\omega_{r,l}(t)$. Przesunięcie to powoduje, że komponent o częstotliwości $(-\omega_{r,l}(t))$ znajdzie się na częstotliwości zerowej. Ponieważ $x_l(t)$ jest sygnałem rzeczywistym o widmie symetrycznym względem osi rzędnych, komponenty na częstotliwościach $(-\omega_{r,l}(t))$ i $\omega_{rl}(t)$ są takie same. Zastosowanie filtru dolnoprzepustowego o bardzo niskiej częstotliwości odcięcia spowoduje wycięcie z sygnału komponentu o częstotliwości $(-\omega_{rl}(t))$. Następnie należy komponent ten przesunąć na częstotliwość $\omega_{rl}(t)$ poprzez ponowne wymnożenie przez $\exp(j\int_t \omega_{r,l}(t)dt)$ i po wymnożeniu przez dwa wyciągnąć z niego część rzeczywistą. Taką interpretację algorytmu HVD przedstawia rys. 5.21. Na rys. zastosowano oznaczenia:

- H_A – zespolony filtr Hilberta;
- IFE – estymator IF;
- Ave – filtr uśredniający;
- A – akumulator;
- LPF – filtr dolnoprzepustowy, wycinający znaleziony komponent;
- $z^{-\tau}$ – element opóźniający, wyrównujący opóźnienie filtrów H_A i Ave;
- $z^{-\tau_{LPF}}$ – element opóźniający, wyrównujący opóźnienie filtru LPF;
- Re( ) – część rzeczywista.



Rys. 5.21. Schemat blokowy algorytmu HVD.

Opisany powyżej algorytm można zmodyfikować, wykorzystując po pierwsze faktoryzację V-K. Jak pokazaliśmy w podrozdz. 5.1, IF fazora dodatnioskrętnego dla głosek dźwięcznych wskazuje na częstotliwość środkową najwyższego formantu, czyli wartość oznaczoną w algorytmie Feldmana przez $\omega_{r,l}(t)$, a sam fazor to $\exp(j\int_t \omega_{r,l}(t)dt)$. IF fazora można więc wykorzystać do przesunięcia widma sygnału na osi częstotliwości. Drugą modyfikacją jaką warto zaproponować jest zastosowanie szerszego filtru dolnoprzepustowego dla wyodrębnienia znalezionego komponentu. Ponieważ dla mowy szukanymi komponentami są formanty, a nie pojedyncze harmoniczne, filtr dolnoprzepustowy musi być dobrany tak, by pokrywał się z pasmem przesuniętego formantu. Wykorzystamy filtry o różnych szerokościach pasm dla kolejnych formantów. Jak pokazały eksperymenty, dobór tych filtrów ma duże znaczenie dla poprawności ekstrakcji formantów. Zaprojektowane filtry mają zbocza o łagodnym nachyleniu, a 6-decybelowe szerokości ich pasm wynoszą: 400 Hz, 500 Hz i 600 Hz, odpowiednio dla formantów F1, F2 i F3. Ostatnią wprowadzoną dalej modyfikacją jest zrealizowanie całego algorytmu na sygnałach zespolonych, tzn. reprezentacja zespolona rzeczywistego sygnału mowy wyznaczana jest przed pierwszą iteracją, wynikiem każdej iteracji jest zespolony przebieg reprezentujący pojedynczy formant, a dopiero po zakończeniu wszystkich iteracji obliczana jest część rzeczywista każdego wyodrębnionego komponentu. Wykonywane są trzy iteracje dla wyodrębnienia trzech najważniejszych formantów. Dodatkowo przeprowadzana jest estymacja częstotliwości podstawowej (za pomocą algorytmu opisanego w poprzednim podrozdziale). Schemat blokowy *l*-tej iteracji zmodyfikowanego algorytmu HVD dla analizy mowy przedstawia rys. 5.22. Na rysunku zastosowano oznaczenia:



Rys. 5.22. Schemat blokowy l-tej iteracji zaproponowanego algorytmu ekstrakcji formantów mowy.

- l – numer iteracji, odpowiadający numerowi poszukiwanego formantu: l=1,2,3;
- H_{A,F_{Cl},B_l} – pasmowy zespolony filtr Hilberta o częstotliwości środkowej F_{Cl} (800 Hz, 2000 Hz i 3000 Hz, odpowiednio, dla l=1,2,3) oraz szerokości pasma B_l (1200 Hz, 3000 Hz i 4000 Hz);
- deE – filtr deemfazy
- PIFPE – estymator fazora dodatnioskrętnego (bifaktoryzacja V-K);
- IFE – estymator IF;
- MF – filtr medianowy;
- A – akumulator;
- LPF$_l$ – filtr dolnoprzepustowy, wycinający znaleziony formant;
- ( )* – operacja sprzężenia;
- $z^{-\tau_l}$ – element opóźniający, wyrównujący opóźnienie filtru zastosowanego w bloku PIFPE oraz medianowego;
- $z^{-\tau_{LPF\,l}}$ – element opóźniający, wyrównujący opóźnienie filtru LPF$_l$;
- $u_l[n]$ - zespolona reprezentacja rzeczywistego sygnału $x_l[n]$ w l-lej iteracji;
- $r_l[n]$ - formant wyodrębniony w l-tej iteracji.

Przed pierwszą iteracją algorytmu sygnał rzeczywisty zamieniany jest na reprezentację zespoloną za pomocą zespolonego filtru Hilberta o paśmie 200 Hz – 5 kHz (trzeci formant nie występuje powyżej częstotliwości 5 kHz). Drugi filtr Hilberta (tj. ten uwidoczniony na rys. 5.14) ogranicza zakres częstotliwości, w którym poszukujemy formantu (pierwszy formant występuje do częstotliwości ok. 1200 Hz, drugi formant występuje w zakresie 500 – 3500 Hz, a trzeci w zakresie 1000 – 5000 Hz). Stosujemy również filtr deemfazy, który ma za zadanie zwiększyć stopień minimalnofazowości sygnału. Następnie w każdej iteracji obliczany jest

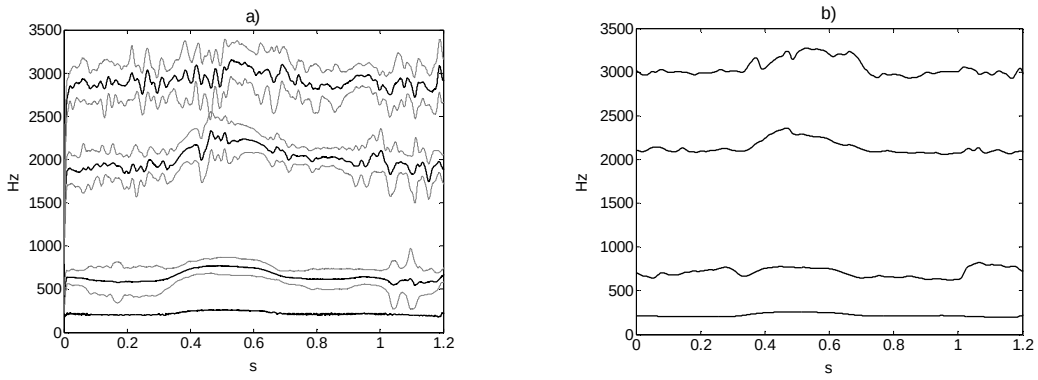
fazor dodatnioskrętny i estymowana jest jego częstotliwość chwilowa. Filtr medianowy jest stosowany w celu usunięcia niepożądanych szpilek w przebiegu IF. Aby przesunąć komponent o wyznaczonej IF na częstotliwość zerową należy sygnał zespolony pomnożyć przez $\exp(-j\int_t \omega_{r,l}(t)dt)$, czyli przez sprzężenie fazora dodatnioskrętnego, obliczone na podstawie wygładzonej IF. W tym przypadku mnożymy przez $\exp(-j\int_t \omega_{r,l}(t)dt)$ a nie przez $\exp(j\int_t \omega_{r,l}(t)dt)$ ponieważ działania przeprowadzane są na sygnałach zespolonych o niesymetrycznym widmie. Widmo należy więc przesunąć na osi częstotliwości w lewo, by szukany komponent znalazł się na częstotliwości zerowej. Następnie komponent jest wycinany za pomocą filtru dolnoprzepustowego i przesuwany z powrotem na częstotliwość $\omega_{r,l}(t)$ poprzez wymnożenie przez $\exp(j\int_t \omega_{r,l}(t)dt)$. Tak uzyskany formant odejmowany jest od sygnału mowy. Jego częstotliwość środkową i szerokość pasma obliczamy estymując ICF, a następnie wygładzając przebiegi IF oraz IB i mnożąc przez $F_S / 2\pi$ (dla uzyskania wartości w hercach).

Zaproponowany algorytm ekstrakcji formantów, tak jak opisywane w p. 3.3.1 metody Maragosa i in. [MA93] [HA94] oraz Kumaresana i Rao [KU99] [RA00] oparty jest na prostym podejściu, polegającym na wyodrębnianiu składowych sygnału wielokomponentowego poprzez ich odfiltrowanie, z tym, że zamiast stosować bank równoległych filtrów, komponenty wyodrębniane są po kolei. Co prawda, iteracje powodują zwiększenie opóźnienia całego algorytmu, ale przy odpowiednio zaprojektowanych filtrach (o krótkich odpowiedziach impulsowych) opóźnienie to nie jest na tyle duże, by nie można było wykorzystywać zaproponowanej metody w systemach czasu rzeczywistego. Ponadto to nie częstotliwości środkowe filtrów adaptują się do estymowanej częstotliwości chwilowej (jak w [MA93][HA94] i [KU99][RA00]), ale znaleziony komponent heterodynowany jest na częstotliwość zerową i wycinany filtrem dolnoprzepustowym. Proponowane podejście wykorzystuje właściwość IF fazora dodatnioskrętnego, której wartość wskazuje na częstotliwość środkową najwyższego formantu w widmie. Ponieważ sygnał mowy jest sygnałem prawie minimalnofazowym, to amplitudy formantów generalnie maleją wraz ze wzrostem ich częstotliwości środkowych (efekt ten można pogłębić stosując filtr deemfazy przed estymacją IF). W algorytmie nie ma konieczności ustalania a priori początkowych

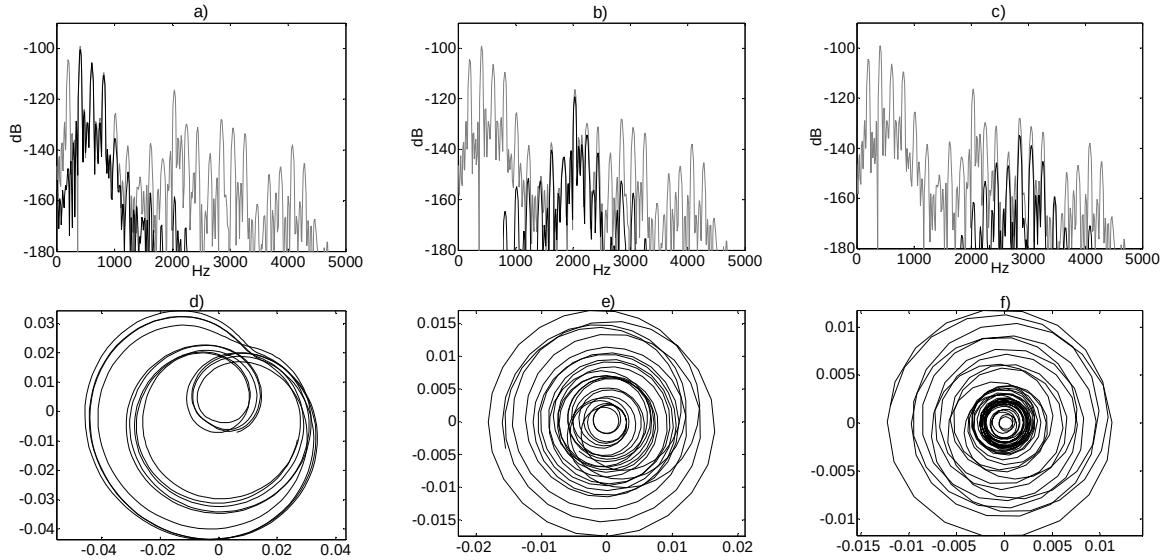
częstotliwości środkowych zastosowanych filtrów (problem ten sygnalizowali autorzy [MA93] [HA94] oraz [KU99] [RA00], na co zwróciliśmy uwagę w p. 3.3.1). Problemem, który może wystąpić w algorytmach stosujących filtrację adaptacyjną, może być opóźnienie adaptacji częstotliwości środkowej filtru względem zmian częstotliwości środkowej formantu. Estymacja częstotliwości środkowej formantu jest poprawna tylko wtedy, gdy częstotliwość środkowa formantu i częstotliwość środkowa filtru różnią się o mniej niż ok. 500 Hz [HA94]. Jeśli częstotliwość środkowa formantu zmieni się znacząco w krótkim czasie, adaptacja filtru może nie nadążyć za tą zmianą, przez co formant znajdzie się poza zakresem pasma filtru. W proponowanym algorytmie problem ten nie występuje.

5.3.3. Eksperymenty

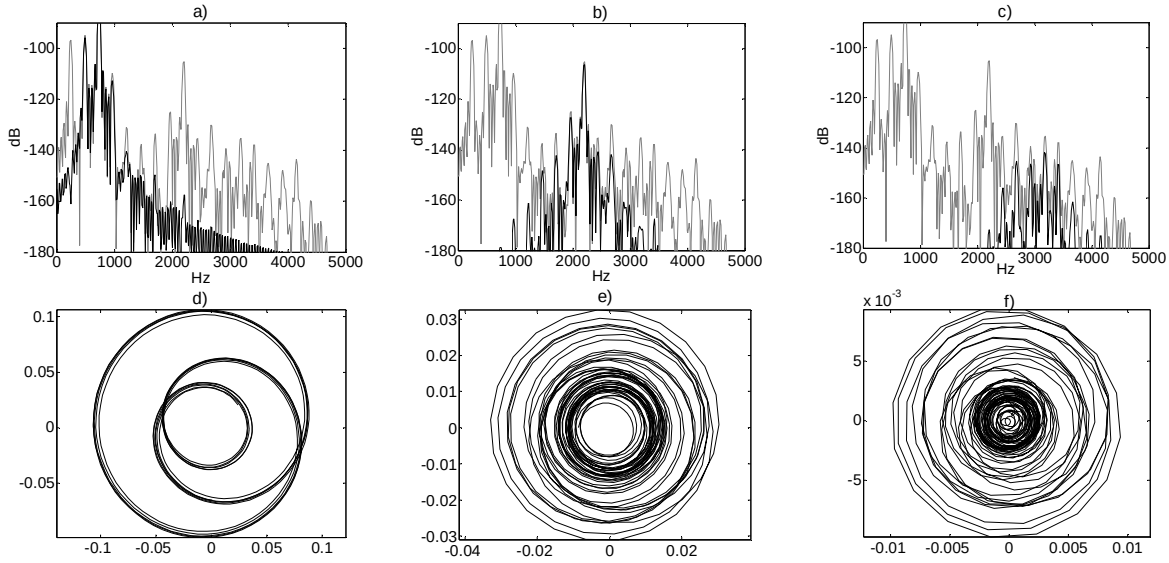
Zaproponowany algorytm ekstrakcji formantów (rys. 5.22) przetestowano najpierw na 24 nagraniach samogłosek, a następnie na nagraniach 200 logatomów, z których każdy składał się z trzech głosek dźwięcznych (każda głoska wystąpiła w nagraniach przynajmniej pięć razy). Dla testowanych sygnałów wyodrębniono trzy pierwsze formanty oraz estymowano ich częstotliwości środkowe. Przykładowe wyniki, uzyskane dla głoski /e/ wypowiadanej ze zmieniającą się częstotliwością podstawową (w zakresie 200 – 260 Hz), przedstawiają rys. 5.23-5.25. Na rys. 5.23a pokazano przebiegi częstotliwości podstawowej oraz częstotliwości środkowych formantów F1, F2 i F3, estymowane za pomocą zaproponowanego algorytmu. Dodatkowo, na rysunku zaznaczono szarym kolorem granice pasm wyodrębnionych formantów, wyznaczone na podstawie części rzeczywistej ICF.



Rys. 5.23. Wynik estymacji częstotliwości środkowych (czarny) i szerokości pasm (szary) formantów F0, F1, F2 i F3 za pomocą zaproponowanego algorytmu (a) oraz algorytmu wykorzystującego LP (b).



Rys. 5.24. Wynik ekstrakcji formantów dla fragmentu głoski /e/ wypowiadanej z częstotliwością podstawową 200 Hz: pierwszy wiersz – periodogramy głoski (kolor szary) oraz wyodrębnionych formantów F1 (a), F2 (b) i F3 (c) (czarny); drugi wiersz – trajektorie zespolone formantów F1 (d), F2 (e) i F3 (f).



Rys. 5.25. Wynik ekstrakcji formantów dla fragmentu głoski /e/ wypowiadanej z częstotliwością podstawową 260 Hz: pierwszy wiersz – periodogramy głoski (kolor szary) oraz wyodrębnionych formantów F1 (a), F2 (b) i F3 (c) (czarny); drugi wiersz – trajektorie zespolone formantów F1 (d), F2 (e) i F3 (f).

Rys. 5.23b przedstawia przebiegi częstotliwości podstawowej oraz częstotliwości rezonansowych formantów, estymowane za pomocą algorytmu opartego na LP (algorytm ten został opracowany i udostępniony przez Morrisona i Nearey’a [WWW3]). Estymaty otrzymane za pomocą tych dwóch algorytmów są zbliżone. Kolejne dwa rysunki

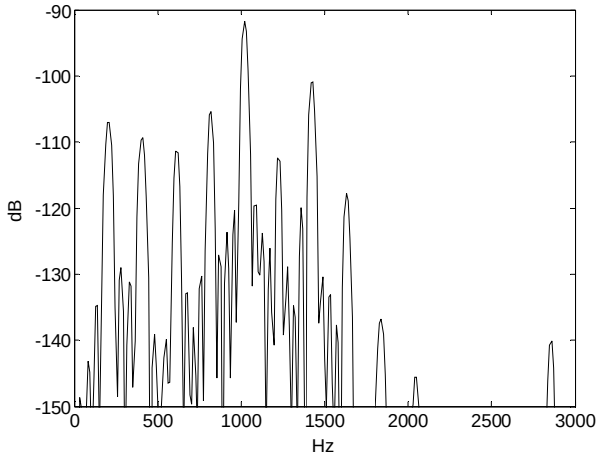
przedstawiają periodogramy oraz trajektorie wyodrębnionych formantów F1, F2 i F3 dla fragmentów głoski /e/ wypowiadanych z częstotliwością podstawową 200 Hz (rys. 5.24) i 260 Hz (rys. 5.25). Widać na nich, że zaproponowany algorytm poprawnie znajduje i wycina kolejne formanty. Warto również zwrócić uwagę na fakt, że wyodrębnione formanty są sygnałami monokomponentowymi (reprezentujące je wskazy obracają się wyłącznie w jednym kierunku – dodatnim).

5.3.3.1. Ocena poprawności estymacji częstotliwości środkowych formantów

Wyniki estymacji częstotliwości środkowych formantów, otrzymanych za pomocą zaproponowanego algorytmu, porównano z przebiegami częstotliwości rezonansowych obliczonych za pomocą algorytmu bazującego na metodzie LP, opracowanego przez Morrisona i Nearey'a. Autorzy udostępnili swój algorytm w Internecie [WWW3] w postaci gotowych skryptów MATLABa. Jest to algorytm nadzorowany przez użytkownika. Dla każdego z trzech pierwszych formantów obliczanych jest osiem estymat częstotliwości rezonansowych. Estymaty obliczane są dla ośmiu różnych częstotliwości, wyznaczających górną granicę wystąpienia trzeciego formantu. Estymaty częstotliwości F_1, F_2 i F_3 dla wszystkich częstotliwości granicznych prezentowane są na wykresach naniesionych na spektrogramy analizowanych dźwięków mowy. Użytkownik może wybrać te z nich, które najlepiej pokrywają się z widocznymi na spektrogramie maksimami widma. Istnieje również możliwość ręcznego skorygowania przebiegów częstotliwości F_1, F_2 i F_3. Dodatkowo, poprawność estymacji można zweryfikować odsłuchując dźwięk zsyntetyzowany na podstawie wyznaczonych estymat częstotliwości rezonansowych.

W eksperymencie wykorzystano 24 samogłoski oraz 200 logatomów, nagranych z szybkością próbkowania 48000 Sa/s. Ponieważ algorytm Morrisona i Nearey'a przetwarza sygnał mowy w ramkach, estymaty otrzymane za pomocą zaproponowanego, potokowego algorytmu podzielono na ramki i dla każdej z nich obliczono wartość średnią, by możliwe było porównanie. Poprawność estymacji określono obliczając błąd gruby (GE) jako stosunek liczby ramek, w których estymata była błędna do liczby wszystkich ramek, wyznaczony w procentach. Estymata była uznawana za poprawną, jeżeli referencyjna częstotliwość

rezonansowa zawierała się paśmie formantu wyodrębnionego za pomocą zaproponowanego algorytmu. Przyjęto takie założenie, ponieważ metoda bazująca na LP znajduje częstotliwości rezonansowe formantów, natomiast zaproponowana metoda – ich częstotliwości środkowe, a wartości te nie muszą się pokrywać (natomiast częstotliwość rezonansowa zawsze zawiera się w granicach pasma formantu). Jako przykład na rys. 5.26 przedstawiono fragment widma amplitudowego głoski /a/ zawierający formanty F1 i F2.



Rys. 5.26. Fragment widma amplitudowego głoski /a/ zawierający formanty F1 i F2.

Zauważmy, że na rys. 5.26 formant F1 obejmuje częstotliwości od ok. 400 Hz do ok. 1200 Hz, a F2 od 1200 Hz do 1800 Hz. Częstotliwości środkowe tych formantów wynoszą więc, odpowiednio, 800 Hz i 1500 Hz. Natomiast ich częstotliwości rezonansowe wynoszą 1000 Hz i 1400 Hz.

Dla oszacowania, jak bardzo estymata obliczona za pomocą proponowanego algorytmu różni się od częstotliwości rezonansowej estymowanej za pomocą metody LP wyznaczono RMS w hercach oraz w procentach (por. (5.13) i (5.14)):

$$\mathrm{RMS}[\mathrm{Hz}] = \sqrt{\frac{\sum_{m=1}^{M}\left(F_l^r(m) - F_l^c(m)\right)^2}{M}} \tag{5.15}$$

$$\mathrm{RMS}[\%] = \sqrt{\frac{\sum_{m=1}^{M}\left(\frac{F_l^r(m) - F_l^c(m)}{F_l^r(m)}\right)^2}{M}} \cdot 100\% \tag{5.16}$$

We wzorach (5.15) oraz (5.16) F_l^r jest częstotliwością rezonansową *l*-tego formantu (*l*=1,2,3), wyznaczoną metodą LP, a F_l^c – częstotliwością środkową, estymowaną za pomocą zaproponowanego algorytmu. *M* jest liczbą ramek, w których nie wystąpiły błędy grube, a *m* to numer kolejnej ramki. Wyniki zestawiono w tab. 5.5.

TAB. 5.5. WYNIKI OCENY POPRAWNOŚCI ESTYMACJI CZĘSTOTLIWOŚCI ŚRODKOWYCH FORMANTÓW.

	Wszystkie nagrania			*Tylko samogłoski*		
	GE [%]	RMS [Hz]	RMS [%]	GE [%]	RMS [Hz]	RMS [%]
F1	0.2	97.7	30.7	0.1	117.4	24.3
F2	9.5	158.7	12.5	16.6	160.7	14.5
F3	18	269.8	10.3	5	248.9	9.1
F1+F2+F3	**9.2**	**183.3**	**20.7**	**7.3**	**183.9**	**17.4**

Błędy oszacowano oddzielnie dla każdego z trzech formantów oraz łącznie. Wyodrębniono również wyniki otrzymane dla nagrań samych samogłosek. Całkowita poprawność estymacji jest na poziomie 90.8 % (co jest wynikiem wysokim dla algorytmu działającego nie tylko w czasie rzeczywistym, ale również potokowo). Najlepiej znajdowany jest formant F1, gdyż w głoskach dźwięcznych dominuje on wyraźnie nad pozostałymi formantami. Wartość RMS w hercach pokazuje, że estymowane wartości częstotliwości środkowych niewiele różnią się od przebiegów referencyjnych (97.7 Hz dla formantu F1 to tylko jeden prążek widma przy niskiej częstotliwości podstawowej – ok. 100 Hz, dla formantu F3 są to już 2-3 prążki, ale jego pasmo jest znacznie szersze niż formantu F1). Eksperymenty pokazały również, że częściej częstotliwość środkowa jest estymowana zbyt nisko niż zbyt wysoko. Dla formantu F1 jest to 73% wszystkich błędnie estymowanych ramek, dla F2 – 91%, a dla F3 – 87%. Występujące błędy estymacji mają trzy główne przyczyny:

1. Formant F1 zachodzi na F2 – ich częstotliwości środkowe znajdują się blisko siebie, a dodatkowo formant F1 ma znacznie większą amplitudę. Przykładem jest samogłoska /u/ (tylna), w której różnica pomiędzy częstotliwościami środkowymi formantów F1 i F2 może wynosić nawet 200 Hz, a F2 ma amplitudę niższą od F1 nawet o 50 dB. Wycinając formant F1 filtrem o szerokości pasma 800 Hz tłumimy znacznie formant F2 przez co niemożliwa staje się jego poprawna detekcja.

2. Poszukiwany formant ma amplitudę niższą lub zbliżoną do kolejnego formantu. Przykładem może być głoska /z/ w której widmie, ze względu na występowanie pobudzenia szumowego wraz z pobudzeniem quasi-okresowym, podwyższone są znacznie amplitudy składowych na częstotliwościach ok. 4000 Hz (czwarty formant). Formant F3 ma amplitudę mniejszą niż F4, czego nie koryguje nawet zastosowana deemfaza. Z tego względu estymowana częstotliwość środkowa F3 jest zbyt wysoka.
3. Poszukiwany formant znajduje się w dużym odstępie od poprzedniego formantu, a jednocześnie zbyt słabo dominuje nad resztą widma. Przykładem jest samogłoska /i/, dla której odstęp między F1 i F2 może wynosić nawet 2000 Hz. Wycięcie formantu F1 powoduje wytłumienie tylko części składowych (o niższych częstotliwościach) znajdujących się pomiędzy F1 i F2. Pozostałe składowe mają na tyle dużą amplitudę, by przesuwać wartość estymowanej częstotliwości chwilowej poniżej rzeczywistej wartości częstotliwości środkowej F2.

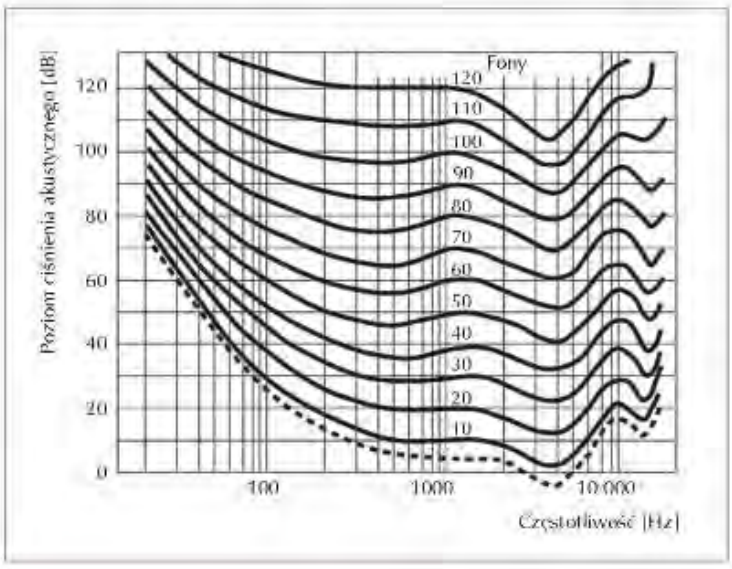
Jak pokazały eksperymenty, próby wyeliminowania jednego z wymienionych błędów powodują zwiększenie częstości występowania innych błędów. Zaproponowany algorytm zaprojektowano tak, by minimalizował całkowitą liczbę błędów.

6. Konwersja głosu w oparciu o czynniki V-KB i ich parametry chwilowe

Dźwięki, w tym mowa, tradycyjnie charakteryzowane są przez trzy atrybuty, które odnoszą się do sposobu ich percypowania przez człowieka. Są to głośność, wysokość oraz barwa. W kolejnym podrozdziale opiszemy te atrybuty oraz związane z nimi parametry sygnału mowy. Następnie przedstawimy możliwości konwersji głosu ludzkiego poprzez modyfikacje zespolonej pulsacji chwilowej obwiedni minimalnofazowej i fazora dodatnioskrętnego.

6.1. Głośność, wysokość i barwa głosu

Głośność jest atrybutem, który opisuje odbieraną przez narząd słuchu „moc" dźwięku, związaną z natężeniem rozchodzącej się fali akustycznej (średniej wartości strumienia energii akustycznej przepływającego w czasie 1 s przez jednostkowe pole powierzchni prostopadłej do kierunku rozchodzenia się fali). W największym uproszczeniu można powiedzieć, że głośność dźwięku wzrasta wraz z natężeniem. Szczegółowe badania [MO03] [MO95] (cytowane w [RO08]) pokazały jednak, że związek między fizycznymi parametrami dźwięku, a percepcją jego głośności jest bardziej złożony – zależy ona nie tylko od natężenia, ale również od charakterystyki widmowej i jej zmian w czasie. Dla opisu tych zależności powstały różne „modele głośności", których omówienie można znaleźć w literaturze [ZW65] [MO95][MO97][SU02] (cytowane w [RO08]). Jednym z pierwszych, ale do dziś stosowanych modeli są krzywe izofoniczne – jednakowej głośności [FL33] (cytowane w [RA07]). Pozwalają one określić liczbowo głośność czystego tonu o wybranej częstotliwości (z zakresu 20 Hz – 20 kHz), wyrażoną w fonach w odniesieniu do poziomu jego natężenia w decybelach. Poziom głośności dźwięku w fonach jest liczbowo równy poziomowi natężenia (w decybelach) tonu o częstotliwości 1 kHz, którego głośność jest taka sama, jak badanego dźwięku. Rys. 6.1 przedstawia przykładowy wykres krzywych izofonicznych. Krzywe izofoniczne są wyznaczane w subiektywnych eksperymentach i dlatego, a także ze względu na różny sposób przeprowadzania tych eksperymentów, znalezione w literaturze przykłady mogą się od siebie różnić.



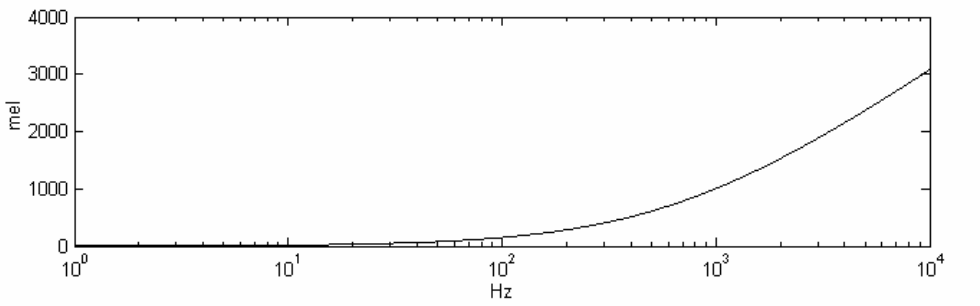
Rys. 6.1. Wykres krzywych izofonicznych [FL33].

Z rys. 6.1 wynika, że słuch jest bardziej czuły na częstotliwościach z zakresu ok. 500 do 7000 Hz. Przykładowo, ton o częstotliwości 100 Hz ma głośność 10 fonów, gdy poziom jego natężenia wynosi 30 dB. Ton o częstotliwości 1000 Hz ma tę samą głośność przy poziomie natężenia tylko 10 dB. Krzywa o głośności 0 fonów wyznacza próg słyszenia dla dźwięków o różnej częstotliwości.

Wysokość dźwięku jest atrybutem, dzięki któremu dźwięki mogą być uszeregowane od niskich do wysokich – tak w 1973 wysokość zdefiniował American National Standard Institute (ANSI). Francuski odpowiednik tej instytucji (AFNOR 1973) dodaje, że wysokość dźwięku związana jest z częstotliwością, czyli dźwięk może być wyższy lub niższy w zależności od tego, czy jego częstotliwość jest duża czy mała. Wysokość definiuje się głównie w odniesieniu do dźwięków, które mają strukturę harmoniczną, a więc również do mowy dźwięcznej i wiąże się ją z charakterystycznym dla takich sygnałów parametrem – częstotliwością podstawową. Zależność między perceptualnym atrybutem, jakim jest wysokość, a fizycznym parametrem – częstotliwością podstawową nie jest liniowa. Opisuje się ją wzorem:

$$P = 1127 \ln\left(1 + F_0 / 700\right) \tag{6.1}$$

gdzie P jest wysokością dźwięku w skali melowej, a F_0 jest częstotliwością podstawową wyrażoną w hercach [ST40] (cytowane w [RA07]). Skala melowa została skalibrowana tak, że 1000 herców odpowiada 1000 melom. Zależność P od F przedstawia rys. 6.2.



Rys. 6.2. Zależność między wysokością dźwięku w melach a częstotliwością podstawową w hercach

Tak zdefiniowana wysokość dźwięku nie tłumaczy jednak wyników eksperymentu opisanego w podrozdz. 4.6, w którym sygnały o tej samej częstotliwości podstawowej miały różną wysokość (brzmienie 4-tonowego sygnału minimalnofazowego było niższe niż maksymalnofazowego). Wyjaśnienie podaje de Cheveigné [CH05]. W sygnałach o strukturze formantowej, w widmie których amplitudy harmonicznych wokół pewnej częstotliwości F_r są większe od pozostałych, obok lub zamiast wysokości odpowiadającej częstotliwości F_0, percypowana może być wysokość odpowiadająca częstotliwości F_r. De Cheveigné nazywa ją wysokością widmową (ang. *spectral pitch*). Jest ona prawdopodobnie związana z jasnością barwy dźwięku. Jej percepcja może być różna u różnych słuchaczy oraz w różnych warunkach odsłuchowych.

Ostatni z omawianych atrybutów, barwa dźwięku, ma najbardziej złożoną, wielowymiarową naturę, a jego definicja jest najmniej ścisła. Według Encyklopedii Muzyki pod redakcją Andrzeja Chodkowskiego [CH01], "barwa jest jedną z podstawowych cech wrażeniowych dźwięku, pozwalającą na szeregowanie dźwięków pod względem ich jakości (np. ostrości, jasności, czy dźwięczności) i rozróżnianie dźwięków mimo ich jednakowej wysokości, głośności i czasu trwania". Podobnie barwę definiuje Amerykański Instytut Standardów (1994). Barwa jest więc tym, co pozwala nam odróżnić dźwięk gitary od skrzypiec, a znalezienie związków między barwą dźwięku a fizycznymi cechami sygnału

umożliwiło m.in. opracowanie algorytmów automatycznego rozpoznawania instrumentów muzycznych [KO01] [DA04]. Przeglądu najważniejszych prac dotyczących barwy dźwięków dokonała Donnadieu w [BE07]. Decydujące znaczenie dla percepcji barwy ma struktura widmowa sygnału. Przez wiele lat powszechne było uproszczenie, że na barwę wpływ ma wyłącznie widmo amplitudowe, tzn. że dwa dźwięki, będące sumą tonów prostych o tych samych częstotliwościach, brzmią tak samo, jeśli amplitudy odpowiadających sobie tonów w obu dźwiękach są sobie równe, bez względu na to, czy ich fazy różnią się między sobą czy też nie. W rzeczywistości jednak widmo fazowe ma wpływ na brzmienie dźwięku [PL69], ale na tyle słaby, że przy odsłuchu w pomieszczeniu, w którym pogłos modyfikuje zależności fazowe, są one niesłyszalne [DE99]. Tę prawidłowość można sprawdzić wykonując prosty test odsłuchowy z sygnałami będącymi sumą kilku tonów prostych (tak jak to opisaliśmy w podrozdz. 4.6). Określenie konkretnych parametrów sygnału, które decydują o barwie dźwięku, jest zadaniem niezwykle trudnym, właśnie ze względu na wielowymiarowość tego atrybutu. Zadania tego podjęli się również twórcy standardu MPEG7, w którym zdefiniowane są deskryptory, pozwalające na rozpoznawanie dźwięków [DA04]. Większość tych parametrów dotyczy widma amplitudowego sygnału. Część z nich to również parametry czasowe, co oznacza, że barwa dźwięku zależy też od dynamicznych właściwości sygnału.

6.2. Głos – cechy dystynktywne mówcy

Głos każdego człowieka jest niepowtarzalny, tak jak linie papilarne czy kształt uszu. Dlatego zalicza się on do cech biometrycznych, czyli takich, które umożliwiają odróżnienie osobnika od innych przedstawicieli tego samego gatunku. Spośród trzech atrybutów opisywanych w poprzednim podrozdziale, to barwa i wysokość są tymi, które decydują o różnym brzmieniu ludzkich głosów. Rozróżnianie głosów, które dla każdego człowieka jest naturalne, może być również wykonane przez systemy automatyczne, lecz w tym celu konieczne jest zdefiniowanie cech dystynktywnych mówcy, które, wydobyte z zarejestrowanego sygnału mowy, pozwalałyby na jego automatyczne rozpoznanie. Jednocześnie ich modyfikacja sprawi, że słyszany głos będzie brzmiał inaczej niż oryginalny.

Podstawowym parametrem, który charakteryzuje mówcę, jest częstotliwość podstawowa tonu krtaniowego F_0, związana z wysokością głosu. Jak pisaliśmy już w rozdz.

2, dla każdego mówcy można wyznaczyć zakres, w którym zmienia się częstotliwość podstawowa jego głosu, a także „naturalną wysokość głosu", czyli taką częstotliwość podstawową, której statystycznie używa najczęściej. Odpowiednio duża baza nagrań mówcy pozwala nam obliczyć wartość średnią estymowanej dla sygnału mowy częstotliwości F_0 oraz jej wartość maksymalną i minimalną, lub, jak proponują Rentzos i inni [RE04], rzykrotną wartość odchylenia standardowego od wartości średniej, która lepiej estymuje zakres F_0.

Na barwę dźwięku największy wpływ ma jego struktura widmowa, o czym pisaliśmy w poprzednim podrozdziale. Dla głosu są to częstotliwości środkowe oraz szerokości pasm poszczególnych formantów. Przy tym formant F3 i wyższe zależą głównie od budowy traktu głosowego, a więc od cech osobniczych mówcy, i ich położenia na osi częstotliwości zmieniają się nieznacznie podczas wypowiadania różnych głosek. Natomiast położenia formantów F1 i F2 mogą zmieniać się w bardzo szerokim zakresie w zależności od wypowiadanej głoski, ale zakres ten jest również charakterystyczny dla danego mówcy. W automatycznych systemach rozpoznawania mówcy czy konwersji głosu parametry te uzyskuje się pośrednio, poprzez obliczenie np. współczynników predykcji liniowej.

Poza wymienionymi cechami dystynktywnymi dla każdego mówcy można określić wzorce intonacji, długości poszczególnych dźwięków i pauz czy sposobu artykulacji. Te parametry nie są jednak związane z barwą głosu.

6.3. Możliwości modyfikacji głosu za pomocą ICF

Celem tej części pracy jest sprawdzenie, czy za pomocą modyfikacji zespolonych pulsacji chwilowych obwiedni minimalnofazowej i fazora dodatnioskrętnego można zmieniać brzmienie ludzkiego głosu, zachowując treść wypowiedzi, tempo, intonację, a także emocje. Zmiany brzmienia głosu nazywać będziemy wymiennie transformacją, modyfikacją lub konwersją głosu. Choć trzeci termin w języku angielskim (*conversion*) stosowany jest głównie dla takich zmian, które głos mówcy źródłowego przekształcają w głos innego, konkretnego mówcy docelowego, w polskich publikacjach nie znaleźliśmy takiego ograniczenia. Ponadto, słownik języka polskiego podaje, że konwersja to „przekształcenie postaci czegoś", a zatem termin ten pasuje jak najbardziej do opisu zmian brzmienia głosu.

Transformacje głosu mogą być wykorzystane w różnych zastosowaniach. Jednym z nich jest anonimizacja (depersonalizacja) mówcy, np. w materiałach szkoleniowych policji czy lekarzy, w sądzie lub w programach telewizyjnych. W zastosowaniach tych modyfikacje głosu pozwolą ukryć tożsamość mówcy, jednocześnie zachowując emocje i zapewniając naturalne brzmienie mowy. Można również wykorzystać konwersję głosu do uzyskania ciekawych efektów dla głosów postaci animowanych w filmach lub grach komputerowych.

Większość znanych z literatury metod konwersji głosu modyfikuje oddzielnie częstotliwość podstawową oraz częstotliwości środkowe i szerokości pasm formantów. Do modyfikacji częstotliwości podstawowej wykorzystuje się najczęściej algorytmy SOLA [KA02] działające w dziedzinie czasu (TD-PSOLA) lub częstotliwości [FD-PSOLA]. Znane są również metody bazujące na sinusoidalnym modelu mowy (opisywanym w rozdz. 2) [QU92]. Zmianę częstotliwości środkowych i szerokości pasm formantów można uzyskać poprzez estymację a następnie modyfikację obwiedni widma sygnału mowy [TU00]. Innym sposobem jest wyznaczenie transmitancji traktu głosowego za pomocą analizy LPC, znalezienie jej biegunów (z których część odpowiada formantom mowy) i zmianę położenia tych biegunów na płaszczyźnie zespolonej [SL95].

Podstawą dla rozpoczęcia badań nad modyfikacją brzmienia za pomocą ICF były prace Hermanowicz i in. [HE88][HE89], dotyczące przesunięcia i obrotu widma sygnałów pasmowych za pomocą resyntezy przebiegu fazy chwilowej, oraz późniejsze [HE06a] [HE06b], w których zmieniana była wysokość dźwięków świergotowych, na przykład nagranych głosów ptaków, poprzez skalowanie zespolonej reprezentacji dynamicznej ($\lambda(t)$, $\omega(t)$). Uzyskane wyniki pokazały, że takie skalowanie pozwala uzyskać bardzo wysoką naturalność brzmienia zmodyfikowanego dźwięku nawet przy bardzo dużym współczynniku skalowania. W swoich pracach Hermanowicz i in. odnosili się również do kwestii spełnienia warunku Bedrosiana przez modyfikowane sygnały [HE07a] [HE07b], gdyż niespełnienie tego warunku wpływa ujemnie na jakość wynikowego sygnału. Jeśli jednak stosuje się bifaktoryzację V-K kwestię tę można pominąć, co zostało przedyskutowane w podrozdz. 4.3.

Z badań opisanych w rozdz. 3 i 4 wiemy, że treść wypowiedzi przenoszona jest przez obwiednię minimalnofazową $a_{mp}(t)$, która całkowicie zależy od obwiedni amplitudowej $a(t)$ sygnału mowy. Wiemy też, że sygnały wielotonowe, które mają taką samą obwiednię amplitudową, ale różnią się czynnikiem $\gamma_{pif}(t)$, mają różną barwę, a nawet różną wysokość

(choć ich częstotliwości podstawowe są takie same) [KA09]. Stąd wnioskujemy, że najbardziej znaczące zmiany brzmienia można uzyskać modyfikując częstotliwość chwilową $\omega_{pif}(t)$ fazora dodatnioskrętnego. Zbadamy również jak na brzmienie głosu wpłyną zmiany innych parametrów chwilowych czynników bifaktoryzacji V-K.

Jak pisaliśmy w rozdz. 3, para $(s_{mp}(t), s_{pif}(t))$ w pełni reprezentuje sygnał zespolony $u(t)$. Przy zachowaniu informacji o fazie początkowej można na podstawie tych przebiegów obliczyć $u(t)$. Jeśli wcześniej zmienimy któryś z przebiegów chwilowych, to oczywiście otrzymamy sygnał zmodyfikowany. Dla sygnału mowy chcemy, by wprowadzone modyfikacje zmieniły tylko brzmienie głosu, bez wpływu na pozostałe cechy mowy.

W niniejszym podrozdziale zaproponujemy dwa sposoby modyfikacji głosu: pierwszy, mniej złożony obliczeniowo, w którym czynniki bifaktoryzacji V-K oraz ich przebiegi chwilowe obliczane są i modyfikowane dla całego sygnału mowy (ten sposób konwersji głosu opisany zosał w artykułach konferencyjnych [KA10a][KA10b]), oraz drugi, w którym najpierw ekstrahowane są formanty, które potem modyfikowane są oddzielnie.

6.3.1. Proponowane modyfikacje ICF

Założeniem dla opracowywanych algorytmów było, by modyfikacje zespolonej pulsacji chwilowej były proste i możliwe do przeprowadzenia próbka po próbce, tak jak algorytmy bifaktoryzacji V-K i estymacji ICF. Z tego względu ograniczymy je do prostych operacji skalowania i przesunięcia. Poniżej opisane zostaną proponowane modyfikacje: skalowanie i przesuwanie IF fazora dodatnioskrętnego oraz skalowanie części rzeczywistej i urojonej ICF obwiedni minimalnofazowej. W poniższych formułach sygnały po modyfikacji oznaczone są indeksem dolnym „1”. Dla uproszczenia wzorów w zapisie pominięto fazy początkowe przebiegów $a_{mp}(t)$ i $\gamma_{pif}(t)$, należy jednak pamiętać, że zachowanie informacji o nich jest konieczne dla odtworzenia sygnału po modyfikacji. Rozważamy następujące sygnały:

$$u(t) = a_{mp}(t)\gamma_{pif}(t)$$
$$\gamma_{pif}(t) = \exp\left(j\varphi_{pif}(t)\right) = \exp\left(j\int_t \omega_{pif}(t)dt\right)$$
$$a_{mp}(t) = a(t)\exp\left(j\varphi_{mp}(t)\right) = a(t)\exp\left(j\int_t \omega_{mp}(t)dt\right)$$

1. Skalowanie IF fazora dodatnioskrętnego (wymnożenie przez stałą *c*)

$$\omega_{pif1}(t) = c\omega_{pif}(t)$$

$$u_1(t) = a_{mp}(t)\exp\left(j\int_t \omega_{pif1}(t)dt \right) = a_{mp}(t)\exp\left(jc\int_t \omega_{pif}(t)dt \right) = a_{mp}(t)\exp^c\left(j\int_t \omega_{pif}(t)dt \right) =$$
$$= a_{mp}(t)\gamma_{pif}^{c}(t) = u(t)\gamma_{pif}^{c-1}(t)$$

Wymnożenie IF fazora dodatnioskrętnego przez stałą *c* jest równoznaczne z podniesieniem PIFP do potęgi *c* (lub też z wymnożeniem sygnału analitycznego przez PIFP podniesiony do potęgi *c*–1). Stała *c* nie może być ujemna, gdyż spowoduje to przesunięcie widma sygnału analitycznego na ujemne częstotliwości. Po ponownej bifaktoryzacji zmieniony jest tylko przebieg PIFP (nie zmienia się amplituda sygnału, od której całkowicie zależy MPE).

2. Przesuwanie IF fazora dodatnioskrętnego (dodanie stałej pulscacji ω_c)

$$\omega_{pif1}(t) = \omega_{pif}(t) + \omega_c$$

$$u_1(t) = a_{mp}(t)\exp\left(j\int_t (\omega_{pif}(t) + \omega_c)dt \right) = a_{mp}(t)\exp\left(j\int_t \omega_{pif}(t)dt + j\int_t \omega_c dt \right) =$$
$$= a_{mp}(t)\exp\left(j\varphi_{pif}(t)\right)\exp(j\omega_c t) = u(t)\exp(j\omega_c t)$$

Przesunięcie dotyczy tylko części urojonej ICF, ponieważ część rzeczywista powinna pozostać równa zeru. Daje to ten sam efekt, co wymnożenie sygnału analitycznego przez sinusoidę zespoloną. W dziedzinie częstotliwości powoduje to przesunięcie widma na osi częstotliwości w prawo, gdy $\omega_c > 0$ lub w lewo, gdy $\omega_c < 0$. W tym drugim przypadku przesunięcie to nie może być zbyt duże, aby widmo sygnału analitycznego $u_1(t)$ nie znalazło się na ujemnych częstotliwościach. Po ponownej bifaktoryzacji zmieniony jest tylko przebieg PIFP.

3. Skalowanie części rzeczywistej ICF obwiedni minimalnofazowej (mnożenie $\sigma(t)$ przez stałą *c*)

$$s_{mp}(t) = \sigma(t) + j\omega_{mp}(t)$$

$$\sigma_{mp} = \frac{a'(t)}{a(t)} \Rightarrow a(t) = \exp\left(j\int_t \sigma(t)dt \right)$$

$$\sigma_1(t) = c\sigma(t) \Rightarrow a_1(t) = \exp\left(j\int_t \sigma_1(t)dt \right) = \exp\left(j\int_t c\sigma(t)dt \right) = \exp\left(jc\int_t \sigma(t)dt \right) =$$
$$= \exp^c\left(j\int_t \sigma(t)dt \right) = a^c(t)$$
$$u_1(t) = a^c(t)\exp\left(j\varphi_{mp}(t)\right)\gamma_{pif}(t) = a^{c-1}(t)u(t)$$

Wymnożenie części rzeczywistej ICF obwiedni minimalnofazowej przez stałą *c* równoważne jest podniesieniu do potęgi *c* obwiedni (amplitudy chwilowej) sygnału lub wymnożeniu sygnału analitycznego przez amplitudę podniesioną do potęgi *c*–1. Po ponownej bifaktoryzacji zmieniają się oba czynniki:

$$a_1(t) = a^c(t) \Rightarrow \lambda_1(t) = c\lambda(t) \Rightarrow \varphi_{mp1}(t) = c\varphi_{mp}(t)$$
$$a_{mp1}(t) = a_1(t)\exp\left(j\varphi_{mp1}(t)\right) = a^c(t)\exp^c\left(j\varphi_{mp1}(t)\right) = a^c_{mp}(t)$$
$$\gamma_{pif1} = \frac{u_1}{a_{mp1}} = \frac{a^c(t)\exp\left(j\varphi_{mp}(t)\right)\gamma_{pif}(t)}{a^c(t)\exp^c\left(j\varphi_{mp1}(t)\right)} = \gamma_{pif}\ \exp^{-(c-1)}\left(j\varphi_{mp}(t)\right) = \exp\left(j(\varphi_{pif}(t) - (c-1)\varphi_{mp}(t)\right)$$

4. Skalowanie części urojonej ICF

$$s_{mp}(t) = \sigma(t) + j\omega_{mp}(t)$$
$$\omega_{mp1}(t) = c\omega_{mp}(t)$$
$$u_1(t) = a(t)\exp\left(j\int_t c\omega_{mp}(t)dt \right)\gamma_{pif}(t) = a(t)\exp\left(j\int_t \omega_{mp}(t)dt + j\int_t (c-1)\omega_{mp}(t)dt \right)\gamma_{pif}(t) =$$
$$= a(t)\exp\left(j\int_t \omega_{mp}(t)dt \right)\exp\left(j\int_t (c-1)\omega_{mp}(t)dt \right)\gamma_{pif}(t) = u(t)\exp\left(j\int_t (c-1)\omega_{mp}(t)dt \right) =$$
$$= u(t)\exp\left(j(c-1)\varphi_{mp}(t)\right)$$

Po ponownej bifaktoryzacji MPE pozostanie niezmieniona (ponieważ niezmieniona jest amplituda chwilowa), zmieni się natomiast przebieg PIFP:

$$\gamma_{pif1}(t) = \frac{u_1(t)}{a_{mp}(t)} = \frac{u(t)\exp\left(j(c-1)\varphi_{mp}(t)\right)}{a_{mp}(t)} = \frac{a_{mp}(t)\gamma_{pif}(t)\exp\left(j(c-1)\varphi_{mp}(t)\right)}{a_{mp}(t)} =$$
$$= \exp\left(j\varphi_{pif}(t)\right)\exp\left(j(c-1)\varphi_{mp}(t)\right) = \exp\left(j(\varphi_{pif}(t) + (c-1)\,j\varphi_{mp}(t))\right)$$
$$\varphi_{pif1}(t) = \varphi_{pif}(t) + (c-1)\varphi_{mp}(t) \Rightarrow \omega_{pif1}(t) = \omega_{pif}(t) + (c-1)\omega_{mp}(t)$$

Wymnożenie części urojonej ICF obwiedni minimalnofazowej przez stałą *c* jest równoznaczne z dodaniem do IF fazora dodatnioskrętnego IF obwiedni minimalnofazowej pomnożonej przez *c*–1.

5. Wyeliminowanie z sygnału udziału IF obwiedni minimalnofazowej

Szczególnym przypadkiem ostatniej modyfikacji jest wymnożenie $\omega_{mp}(t)$ przez zero:

$$\varphi_{mp1}(t) = 0$$
$$u_1(t) = a(t)\gamma_{pif}(t) = a(t)\exp\left(j\varphi_{pif}(t)\right)$$

Po ponownej bifaktoryzacji MPE pozostanie niezmieniona, natomiast IF fazora dodatnioskrętnego będzie pomniejszona o wartość IF obwiedni minimalnofazowej.

Nie jest analizowane przesuwanie ICF obwiedni minimalnofazowej, gdyż przesunięcie części rzeczywistej spowoduje eksponencjalny wzrost amplitudy sygnału w czasie, natomiast przesunięcie części urojonej da taki sam efekt jak przesunięcie IF fazora dodatnioskrętnego, ponieważ $\omega(t) = \omega_{mp}(t) + \omega_{pif}(t)$.

6.3.1.1. Synteza sygnału mowy po modyfikacjach

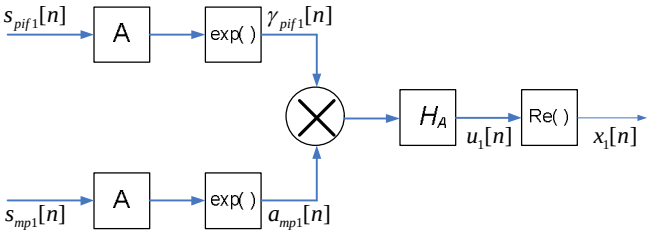
Wiemy, że ICF sygnału analitycznego w pełni go reprezentuje, tzn. z przebiegu ICF można odzyskać oryginalny sygnał. Również bifaktoryzacja V-K jest w pełni odwracalna. Aby z przebiegów ICF obwiedni minimalnofazowej i fazora dodatnioskrętnego uzyskać sygnał analityczny, wystarczy wykonać działanie odwrotne niż przy liczeniu zespolonej pulsacji chwilowej, otrzymując czynniki bifaktoryzacji, a następnie czynniki te należy wymnożyć przez siebie. Po modyfikacji ICF uzyskany w ten sposób sygnał analityczny będzie oczywiście różny od oryginalnego. Ponieważ ICF obliczamy za pomocą wzoru:

$$s(t) = \frac{d}{dt}\ln\left(u(t)\right) \tag{6.2}$$

działaniem odwrotnym jest:

$$u(t) = \exp\left(\int_t s(t)dt \right) \tag{6.3}$$

W implementacji dyskretnej całkę po czasie zastępujemy akumulatorem. W tym miejscu warto również przypomnieć, że aby poprawnie odtworzyć sygnał analityczny musimy zachować informację o fazie początkowej. Najłatwiej to zrobić obliczając pierwsze próbki zespolonych pulsacji chwilowych $s_{mp}[1]$ i $s_{pif}[1]$ jako odpowiednio $\mathrm{Ln}(a_{mp}[1])$ oraz $\mathrm{Ln}(\gamma_{pif}[1])$. Schemat blokowy syntezy mowy po modyfikacji ICF przedstawia rys. 6.3. Na rysunku przez A oznaczono blok akumulatora.



Rys. 6.3. Synteza sygnału mowy po modyfikacji ICF

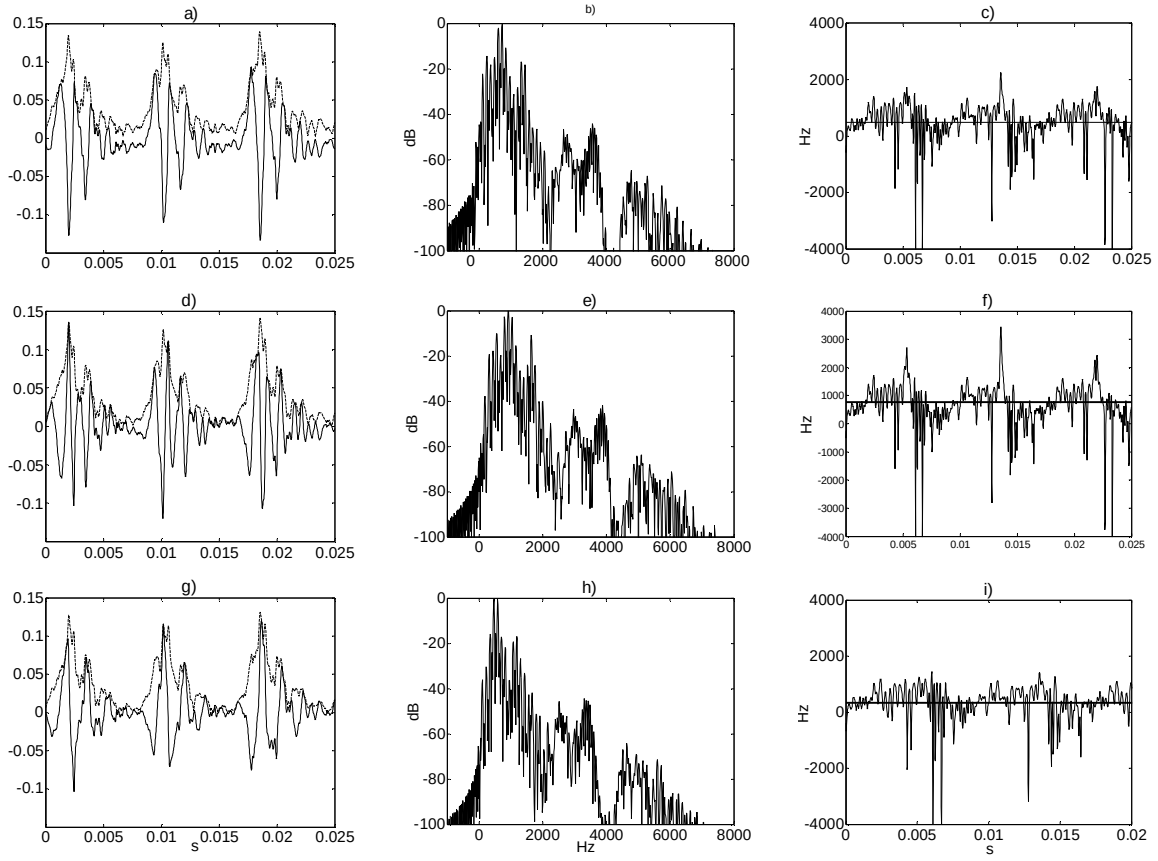
Po wymnożeniu czynników $a_{mp1}[n]$ i $\gamma_{pif1}[n]$ uzyskany sygnał zespolony filtrowany jest za pomocą zespolonego filtru Hilberta H_A (przyczynowego filtru FIR aproksymującego idealny zespolony filtr Hilberta), by sygnał otrzymany w wyniku wprowadzonych modyfikacji był analityczny.

6.3.2. Modyfikacje sygnału mowy

Zaproponowane w poprzednim podrozdziale modyfikacje ICF wykorzystamy do konwersji głosu, przy czym czynniki bifaktoryzacji i ich ICF liczone będą dla całego sygnału mowy (w kolejnym punkcie pokażemy, że modyfikacje te można zastosować również do poszczególnych formantów). Zmiany, jakie modyfikacje ICF wprowadzają do sygnału mowy zostaną omówione na przykładzie głoski /a/. W dalszej części przedstawimy również wyniki testów odsłuchowych oceniających brzmienie zmodyfikowanego głosu.

1. Skalowanie IF fazora dodatnioskrętnego.

Wyniki dla skalowania IF fazora dodatnioskrętnego zostaną przedstawione dla dwóch współczynników skalujących: c=1.5 oraz c=1/1.5. Wykresy przed i po modyfikacjach przedstawia rys. 6.4.



Rys. 6.4. Zmiany w sygnale mowy wynikające ze skalowania IF fazora dodatnioskrętnego: wykresy dla sygnału oryginalnego (pierwszy wiersz), wykresy po skalowaniu ze współczynnikiem 1.5 (środkowa drugi wiersz), wykresy po skalowaniu ze współczynnikiem 1/1.5 (trzeci wiersz); oscylogramy sygnału mowy (lewa kolumna) wraz z obwiednią (linia przerywana), periodogramy sygnału mowy (środkowa kolumna), przebiegi IF sygnału mowy (prawa kolumna).

Skalowanie IF przede wszystkim przesuwa jej wartość średnią w górę (gdy $c > 1$) lub w dół (gdy $c < 1$), co powoduje przesunięcie widma sygnału mowy odpowiednio w prawo lub w lewo na osi częstotliwości. Ponadto zakres zmian IF zwiększa się wówczas lub zmniejsza, co również można zauważyć w widmie sygnału. Im większy jest współczynnik skalujący, tym mniej wyraźna jest struktura formantowa mowy. Amplitudy formantów wyrównują się – pierwszy formant jest coraz mniej dominujący. Zmniejsza się tym samym stopień minimalnofazowości sygnału. Dla wartości c większych niż ok. 2.5 (w zależności od

modyfikowanego głosu) struktura formantowa praktycznie zanika, co znacznie zmniejsza zrozumiałość mowy. Dla $c<1$ efekt zmiany struktury formantowej jest odwrotny. Analizując oscylogramy sygnału mowy widzimy, że obwiednia amplitudowa nie zmienia się (zgodnie z ustaleniami z poprzedniego punktu). Zauważamy natomiast, że sygnał pod obwiednią ma więcej lub mniej przejść przez zero (odpowiednio dla $c>1$ i $c<1$). Częstotliwość podstawowa sygnału pozostaje niezmieniona.

Jeśli chodzi o zmianę brzmienia głosu, to przede wszystkim zauważane jest podniesienie ($c>1$) lub obniżenie ($c<1$) percypowanej wysokości. Jest to efekt opisywanej przez de Cheveigné wysokości widmowej, o której pisaliśmy w podrozdz. 6.1, gdyż częstotliwość podstawowa sygnału nie zmienia się. Jednocześnie dźwięk staje się bardziej lub mniej jasny, co wynika z większego lub mniejszego udziału wysokich częstotliwości. Problemem jest to, że przy mnożeniu IF przez czynnik nie będący liczbą całkowitą uzyskujemy przesunięcie widma sygnału o częstotliwość, która nie jest wielokrotnością F_0. Przez to harmoniczne sygnału zmodyfikowanego nie znajdują się na częstotliwościach, które są wielokrotnością F_0. W efekcie słyszymy dwie wysokości głosu jednocześnie – tę związaną z F_0 oraz widmową. Efekt ten jest bardzo mało zauważalny, gdy c niewiele różni się od liczby całkowitej (np. 1.1 lub 0.9). Ale też wtedy uzyskane brzmienie jest prawie takie samo, jak w przypadku zastosowania współczynnika całkowitego. Najlepszym współczynnikiem skalowania okazuje się więc $c=2$ (dla niektórych głosów również $c=3$, choć wtedy głos jest bardzo nienaturalny). Jeśli wybierzemy $c=0$ to wyeliminujemy udział PIFP z sygnału, pozostanie tylko MPE. Podczas odsłuchiwania zmodyfikowanych fraz zauważyliśmy również, że opisywana modyfikacja IF powoduje uwydatnienie sybilantów, przez co mowa jest sepleniąca. Można tego uniknąć modyfikując IF tylko dźwięcznych fragmentów mowy. Do klasyfikacji mowy na dźwięczną i bezdźwięczną można użyć prostego klasyfikatora opisanego w p. 5.2.1.

2. Przesuwanie IF fazora dodatnioskrętnego

Przesuwanie IF fazora dodatnioskrętnego powoduje jedynie zwiększenie lub zmniejszenie jego wartości średniej, a kształt jego przebiegu pozostaje taki sam. Tym samym widmo sygnału jest przesuwane w prawo lub w lewo na osi częstotliwości, przy zachowaniu struktury formantowej. Ponownie, jeżeli widmo zostanie przesunięte o częstotliwość nie

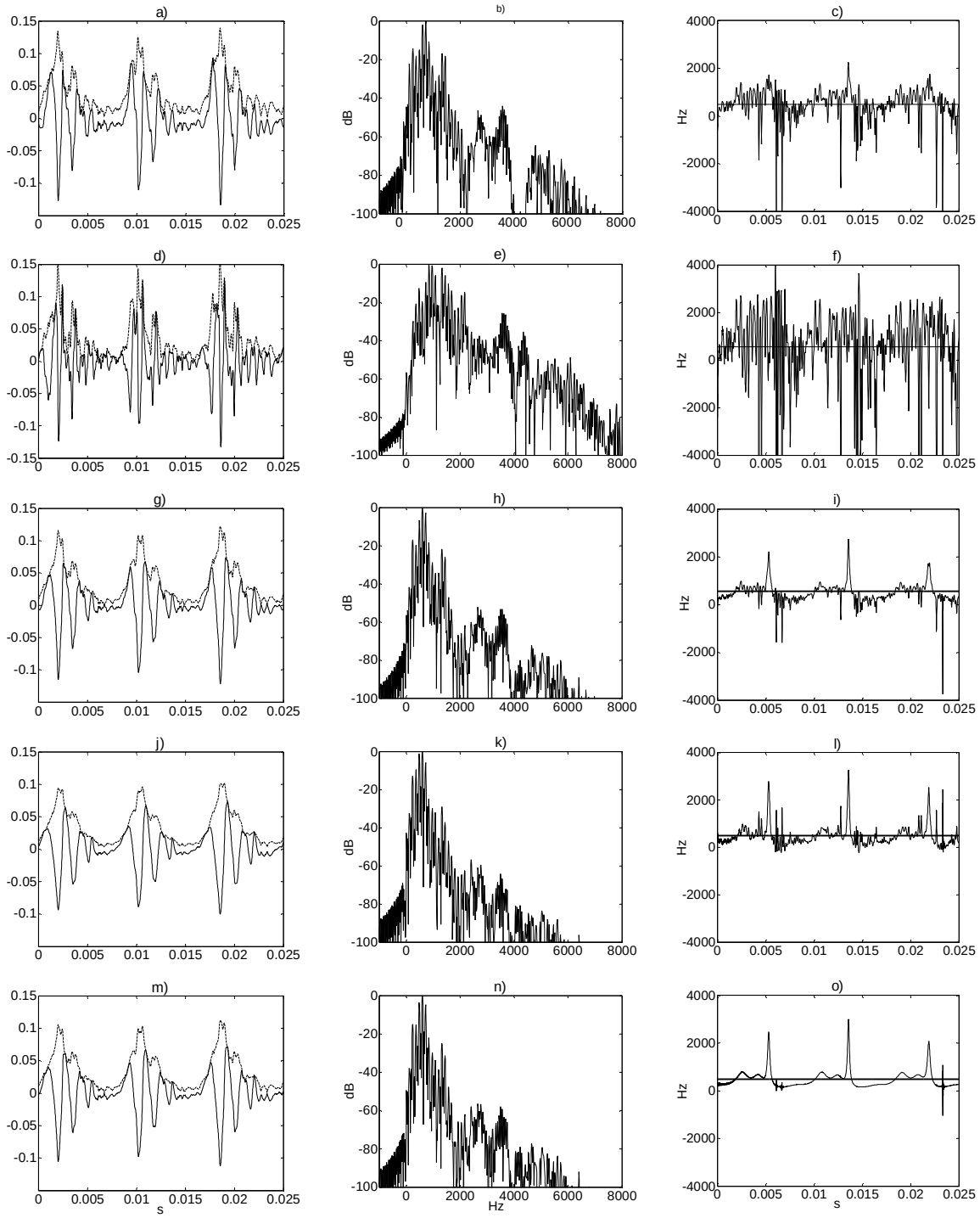
będącą wielokrotnością F_0, to usłyszymy jednocześnie dwie wysokości głosu. Najlepszy wynik uzyskamy przesuwając IF nie o stałą wartość ω_c, ale dodając zmieniającą się w czasie, zgodnie z częstotliwością podstawową, pulsację $\omega_0[n] = 2\pi k F_0[n]$, gdzie k jest liczbą całkowitą. Estymację częstotliwości podstawowej należy oczywiście przeprowadzić dla każdej próbki, gdyż algorytm konwersji głosu jest algorytmem potokowym, wykorzystując estymator opisany w podrozdz. 5.2. Eksperymenty pokazały, że najlepiej brzmiący głos uzyskamy dla k nie mniejszych niż –2 i nie większych niż 3. Nie należy wybierać mniejszych k, by widmo zmodyfikowanej mowy nie znalazło się po stronie ujemnych częstotliwości. Dla k większych od 3 głos jest coraz mniej naturalny.

3. Skalowanie części rzeczywistej ICF obwiedni minimalnofazowej

Skalowanie $\sigma[n]$ zmienia obwiednię amplitudową sygnału. Zmiany te są nieliniowe, gdyż wymnożenie $\sigma[n]$ przez współczynnik c powoduje podniesienie do potęgi c amplitudy chwilowej (modyfikacji liniowych można dokonać oczywiście bezpośrednio na przebiegu $a[n]$). Można rozróżnić dwa aspekty tej modyfikacji: zmianę głośności dźwięku oraz zmianę jego dynamiki. Dla $c > 1$ zakres dynamiki (rozumiany jako różnica między maksimum i minimum logobwiedni) zwiększa się, a dla $c < 1$ się zmniejsza. Współczynnik c nie powinien być zbyt mały ani zbyt duży, gdyż w pierwszym przypadku szum jest za bardzo wzmacniany, a w drugim – dźwięki o niskiej amplitudzie są zbyt mocno tłumione. Ponieważ część rzeczywista jest również wyznacznikiem chwilowej szerokości pasma sygnału, to skalowanie $\sigma[n]$ zmienia szerokość pasma zajmowanego przez sygnał mowy.

4. Skalowanie części urojonej ICF obwiedni minimalnofazowej

Rys. 6.5 przedstawia wyniki modyfikacji sygnału mowy poprzez skalowanie IF obwiedni minimalnofazowej ze współczynnikiem skalowania c=–1/3, 0, 1/3, 3. Skalowanie $\omega_{mp}[n]$ nie zmienia wartości średniej IF sygnału mowy, widmo nie jest więc przesuwane na osi częstotliwości. Zmienia się natomiast kształt widma. Dla $c < 1$ wyższe formanty są tłumione, dla $c > 1$ wzmacniane. Mniejszy udział wysokich częstotliwości w widmie sygnału sprawia, że jego przebieg jest gładszy i odwrotnie.



Rys. 6.5. Zmiany w sygnale mowy wynikające ze skalowania IF obwiedni minimalnofazowej: wykresy dla sygnału oryginalnego (pierwszy wiersz), wykresy po skalowaniu ze współczynnikiem c=3 (drugi wiersz), c=1/3 (trzeci wiersz), c=–1/3 (czwarty wiersz) i c=0 (piąty wiersz); oscylogramy sygnału mowy (lewa kolumna) wraz z obwiednią (linia przerywana), periodogramy sygnału mowy (środkowa kolumna), przebiegi IF sygnału mowy (prawa kolumna).

Zmiany brzmienia głosu dotyczą wyłącznie jego barwy. Im większy jest udział $\omega_{mp}[n]$ tym jaśniejszy jest dźwięk. Percypowana wysokość pozostaje natomiast niezmieniona. Dla $c > 3$ głos staje się coraz mniej naturalny, a powyżej c=5 jakość i zrozumiałość mowy są znacznie degradowane. Nie należy też wybierać współczynników mniejszych niż –1, gdyż wtedy wyższe formanty są za bardzo tłumione.

5. Łączenie modyfikacji

Opisane modyfikacje mogą być łączone, by uzyskać jednocześnie różne zmiany głosu. Eksperymenty pokazały, że najlepsze wyniki uzyskuje się, gdy łączy się modyfikacje $\omega_{pif}[n]$ i $\omega_{mp}[n]$. W przypadku, gdy $\omega_{pif}[n]$ jest podwyższana (poprzez skalowanie lub przesunięcie), można zmniejszyć udział $\omega_{mp}[n]$ tak, by głos nie był zbyt jasny. I odwrotnie, jeśli przesuwamy $\omega_{pif}[n]$ w dół, to zwiększenie udziału $\omega_{mp}[n]$ dodaje głosowi jasności, przez co jest on bardziej naturalny.

Interesujący efekt można też uzyskać modyfikując jednocześnie część rzeczywistą i urojoną $s_{mp}[n]$. Nie zmienia ona znacząco barwy głosu, jednak mnożąc $\sigma[n]$ przez $c < 1$, a $\omega_{mp}[n]$ przez $1/c$ możemy uzyskać efekt, jakby głos nagrywany był dalej od mikrofonu niż oryginalny.

Żadna z proponowanych modyfikacji ani ich kombinacji nie zmienia czasowej struktury wypowiadanych fraz (zachowane jest tempo, długości poszczególnych dźwięków i pauz). Niezmienione pozostają również przebiegi częstotliwości podstawowej, a więc zachowane są akcenty wyrazowe i zdaniowe, nawet, jeżeli percypowana wysokość głosu zmienia się. Algorytm konwersji głosu wykorzystujący takie modyfikacje jest szybki i mało złożony obliczeniowo. Ponadto działa on próbka po próbce w przeciwieństwie do tradycyjnych algorytmów, które przetwarzają mowę w ramkach [BA96] [KA02] [ST98] [TU00].

6.3.2.1. Testy odsłuchowe

Efekty brzmieniowe zaproponowanych modyfikacji przy użyciu różnych współczynników skalowania i przesunięcia zostały ocenione w testach odsłuchowych.

Poszczególnym modyfikacjom oraz ich kombinacjom poddane zostały dwa głosy męskie i dwa żeńskie. Zestaw nagrań testowych zawierał 20 fraz. Frazy były odtwarzane w przypadkowej kolejności. Uczestnicy testu odsłuchiwali je dwukrotnie. Za pierwszym razem oceniali naturalność głosu w skali od 1 do 5 z krokiem 0.5. Na drugim etapie ich zadaniem była ocena ogólnej jakości nagrania, przy czym każdy z uczestników mógł sam zdefiniować, co rozumie pod pojęciem „jakość". Skala oceny była taka, jak wyżej. W zestawie odsłuchowym znalazły się również oryginalne nagrania każdego mówcy. Stanowiły one sygnały referencyjne. Ponieważ subiektywna skala oceny jakości czy naturalności mowy może być różna u różnych słuchaczy, wyniki takich testów mogą być bardzo rozbieżne. Wprowadzenie do zestawu odsłuchowego nagrań, których ocena jest z góry znana, pozwala wyeliminować ten efekt. Odchylenie ocen każdego słuchacza od oceny standardowej jest wykorzystywane do zrównoważenia wyników testów dla pozostałych wyrażeń.

Pierwszym wnioskiem, jaki można wyciągnąć na podstawie wyników oceny naturalności mowy jest to, że modyfikacje głosów o niższej częstotliwości podstawowej (mężczyzn) oceniane były wyżej. Średnia ocen od wszystkich słuchaczy dla żadnej z modyfikacji głosów męskich nie wyniosła mniej niż 3, podczas gdy dla głosów kobiecych najniższa średnia ocen wyniosła 1.7. Najwyżej oceniane były nagrania, w których modyfikowana była IF obwiedni minimalnofazowej. Żaden z uczestników nie ocenił ich poniżej 3.5. Oceny były tym niższe im wyższy był współczynnik skalujący.

Modyfikacją, dzięki której można uzyskać największe zmiany brzmienia, było przesunięcie IF fazora dodatnioskrętnego o wielokrotność częstotliwości podstawowej. Wyniki oceny tych nagrań zawierały się w przedziale 2–4.5, gdy IF była przesuwana w górę, oraz 1.7-3.5, gdy była przesuwana w dół. Oceny były wyższe, gdy przesunięcie $\omega_{pif}[n]$ występowało jednocześnie ze skalowaniem $\omega_{mp}[n]$ (od 2.5 do 4.5, wyższe dla głosów męskich). Jeszcze wyżej oceniono nagrania, gdzie modyfikowana była tylko mowa dźwięczna.

Drugi etap testów pokazał, że dla ok. 75% słuchaczy jakość dźwięku jest silnie skorelowana z naturalnością głosu. Nagrania, które oceniane były jako nienaturalne otrzymywały również niską ocenę jakości i odwrotnie. Wyjątkiem były głosy, w których $\omega_{mp}[n]$ skalowana była ze współczynnikiem większym niż dwa. Tu, mimo wysokiej oceny

naturalności, jakość oceniana była nisko ze względu na słyszalne zniekształcenia fazy w postaci trzasków.

Wyniki testów odsłuchowych pokazują, że modyfikacją, która najbardziej znacząco zmienia brzmienie głosu, a jednocześnie zachowuje satysfakcjonującą naturalność i jakość mowy jest przesuwanie IF fazora dodatnioskrętnego o wielokrotność częstotliwości podstawowej z jednoczesnym skalowaniem IF obwiedni minimalnofazowej. Chociaż za pomocą takich modyfikacji nie można przeprowadzić konwersji głosu jednego mówcy na głos innego mówcy, to można jednak wykorzystać je w działającym online algorytmie konwersji głosu w zastosowaniach takich, jak depersonalizacja mówcy.

6.3.3. Modyfikacje ICF poszczególnych formantów

Ponieważ na brzmienie ludzkiego głosu bardzo duży wpływ mają: położenie poszczególnych formantów na osi częstotliwości oraz ich amplitudy i szerokości pasm, sensowna wydaje się ich oddzielna modyfikacja. Transformacje opisane w poprzednim punkcie pozwalały wpływać na kształt i położenie całego widma, a nie konkretnych formantów. W niniejszym punkcie opiszemy sposób konwersji głosu poprzez modyfikację ICF poszczególnych formantów. Skorzystamy przy tym z algorytmu ekstrakcji formantów opisanego w podrozdz. 5.3. Jest to metoda bardziej złożona obliczeniowo niż ta opisana w poprzednim punkcie, ale jednocześnie umożliwia ona uzyskanie bardzo naturalnych głosów przy dużym stopniu zmiany ich brzmienia.

W opisywanym algorytmie konwersji głosu każdy formant traktowany jest jako oddzielny przebieg zespolony. IF poszczególnych formantów modyfikowane są indywidualnie z wykorzystaniem modyfikacji opisanych w poprzednim punkcie. Możemy więc przesunąć formant na osi częstotliwości poprzez przesunięcie go IF w górę lub w dół. Da to najbardziej słyszalne zmiany w brzmieniu głosu. Jak zaznaczono w poprzednim punkcie, IF należeży przesuwać o wartość będącą wielokrotnością częstotliwości podstawowej (najlepsze wyniki daje przesunięcie o $2F_0$ lub $3F_0$). W najprostszym przypadku przesuwamy IF o stałą wielokrotnośc F_0. Można jednak zmieniać tą wartość zgodnie z przyjęta formułą (np. przesunięcie o niższą wielokrotność F_0 formantów znajdujących się poniżej 2000 Hz i wyższą dla formantów powyżej 2000 Hz). Modyfikowane są wyłacznie dźwięczne fragmenty

mowy, gdyż fragmenty bezdźwięczne nie mają wpływu na brzmienie głosu, a ich modyfiakcja może wpływac negatywnie na jakość i naturalność głosu, o czym pisaliśmy w p. 6.3.2.

Bardziej delikatne zmiany możemy uzyskać skalując ICF obwiedni minimalnofazowej formantu (lub tylko obwiednię chwilową). W ten sposób możemy poszerzyć lub zawęzić pasmo zajmowane przez formant, zwiększyć jego amplitudę lub zmienić kształt widma (to ostatnie poprzez skalowanie części urojonej). Modyfikacje można oczywiście łączyć oraz można modyfikować pojedynczy formant lub kilka formantów. Wstępne eksperymenty pokazały, że najbardziej naturalnie brzmiący głos uzyskamy modyfikując ICF formantów: drugiego i trzeciego. Naturalność i jakość zmodyfikowanych głosów poddane zostały ocenie w testach odsłuchowych.

6.3.3.1. Testy odsłuchowe

Dla oceny naturalności i jakości zmodyfikowanej mowy wykorzystano subiektywny test parametryczny ACR (ang. *Absolute Category Rating*), w którym miarą oceny parametrów był MOS (ang. *Mean Opinion Score*). Zmiany w brzmienu głosu, wprowadzane zaproponowane modyfikacje mogą się bardzo różnić w zależności od charakteru modyfikowanego głosu. Dlatego najistostniejszym w przeprowadzonych testach było ocenienie, czy poprzez modyfikacje ICF można uzyskać głos o wysokim stopniu naturalności i czy nie obniżają one znacząco jakości sygnału mowy. Aby to potwierdzić wybrano dwa przykładowe głosy (męski i żeński), z których każdy poddany został czterem modyfikacjom:

1) przesunięcie częstotliwości środkowej formantu F2 w górę o $2F_0$;
2) przesunięcie częstotliwości środkowej formantu F2 w dół o $2F_0$;
3) przesunięcie częstotliwości środkowej formantu F2 w górę o $2F_0$ z jednoczesnym zmniejszeniem jego amplitudy chwilowej oraz przesunięciem częstotliwości środkowej formantu F3 w dół o $2F_0$;
4) przesunięcie częstotliwości środkowej formantu F2 w dół o $2F_0$ z jednoczesnym zmniejszeniem jego amplitudy chwilowej oraz przesunięciem częstotliwości środkowej formantu F3 w górę o $2F_0$.

Modyfikacje 1 i 2 wybrano, gdyż dają one najbardziej znaczące zmiany w brzmieniu głosu. Dodanie do nich modyfikacji amplitudy chwilowej i przesunięcia formantu F3 daje dodatkowe zmiany w brzmieniu i może poprawić naturalność głosu. Do 8 uzyskanych w ten sposób sygnałów testowych dodano 2 nagrania, w których głosy mówców nie zostały zmodyfikowane. Czas trwania sygnałów testowych wynosił od 3 do 8 sekund. Nagrania materiałów do testów przeprowadzone zostały podczas jednej sesji (nie było więc zmiany warunków). Wypowiedź każdego mówcy zapisana została w pojedynczym pliku, a następnie w całości przetworzona z zastosowaniem każdej modyfikacji. Do testów wycięte zostały próbki z każdego z uzyskanych w ten sposób nagrań. Dla każdej modyfikacji i oryginału wycięty został inny fragment nagrania mówcy. Zrobiono to celowo, by nie wskazywać słuchaczom, ile było mówców, a ile głosów jest zmodyfikowanych. Sposób przygotowania próbek testowych daje pewność, że różnice w próbkach, w zakresie brzmienia głosu i jakości nagrania, wynikają wyłacznie z wprowadzonych modyfikacji.

Test przeprowadzono w trzech seriach. W pierwszych dwóch seriach zadaniem słuchaczy była ocena parametrów związanych bardziej z brzmieniem głosu: jasności, szorstkości i stłumienia w pierwszej serii oraz, w jakim stopniu głos jest nosowy, gardłowy lub świszczący w drugiej serii. W obu seriach oceniana była również naturalność głosu (jeśli dla jakiegoś słuchacza ocena tego parametru różniła się znacznie w pierwszej i drugiej serii, wyniki jego testu nie były wyłączane z porównania). W trzeciej serii oceniono jakość zmodyfikowanej mowy na podstawie czterech parametrów: stopnia występowania szumu szerokopasmowego, szumu gaussowskiego, szumu niskoczęstotliwościowego i trzasków. Parametry wybrane zostały z grupy parametrów zaproponowanych przez Becha i Zacharova do oceny mowy [BE06]. Autorzy [BE06] proponują wykorzystanie większej liczby parametrów, jednak, aby ograniczyć czas trwania testów wybrano te, które według doświadczenia autora niniejszej pracy mogły być istotne. W teście zrezygnowano z oceny ogólnej jakości mowy, gdyż, jak pokazały wstępne testy odsłuchowe, opisane w podp. 6.3.2.1, tak oceniana jakość mowy jest silnie skorelowana z naturalnością głosu i nie może posłużyć do określenia stopnia obniżenia jakości poprzez wprowadzane modyfikacje.

W normie ITU-T P.800 [ITU-T,96], która definiuje test ACR oraz MOS zalecana jest skala pięciostopniowa, tzn. słuchacz ma możliwość wystawienia oceny 1, 2, 3, 4 albo 5. Jednak, aby umożliwić słuchaczom bardziej dokładne różnicowanie ocen różnych nagrań, w testach

przyjęto skalę 11-stopniową, od 0 do 10, przy czym 0 oznaczało, odpowiednio, głos bardzo nienaturalny, ciemny, chropawy, nosowy, gardłowy oraz w dużym stopniu świszczący i przytłumiony. 10 natomiast oznaczało odpowiednio: głos bardzo naturalny, jasny, gładki, w bardzo małym stopniu (lub w ogóle) nosowy, świszczący, gardłowy i przytłumiony. Test przeprowadzono z wykorzystaniem słychawek studyjnych. Kolejność odtwarzania nagrań była przypadkowa. Słuchacz mógł odsłuchać każde nagranie dowolną ilość razy przed wystawieniem oceny. W teście wzięło udział 24 słuchaczy, studentów i pracowników Katedry Systemów Multimedialnych PG.

Opracowując wyniki testów odsłuchowych zauważono, że oceny poszczególnych parametrów tego samego głosu wystawione przez różnych słuchaczy mogą się znacznie różnić. Natomiast rozkład różnic pomiędzy oceną poszczególnych parametrów dla głosu naturalnego i zmodyfikowanego charakteryzuje się znacznie mniejszym odchyleniem standardowym. Z tego względu wymienione poniżej wnioski dotyczą różnic pomiędzy głosem oryginalnym a zmodyfikowanym. W tab. 6.1 zamieszczono średnie różnice pomiędzy ocenami dla głosów oryginalnych i poszczególnych głosów zmodyfikowanych uzyskanymi w serii 1 i 2 testów. Numery modyfikacji odpowiadają tym z listy zamieszczonej na str. 139. Znak minus oznacza, że średnia ocena dla głosu zmodyfikowanego była niższa niż dla głosu oryginalnego.

TAB. 6.1. WYNIKI TESTU MOS DLA MODYFIKACJI GŁOSÓW

	Jasność	Szorstkość	Stopien Stłumienia	Nosowość	Stopień świszczenia	Gardłowość	Naturalność
Głos męski							
Modyfikacja 1	1.92	2.08	0.67	-0.33	0.33	4.67	1.42
Modyfikacja 2	-1.33	1.5833	-4.08	-5.75	-0.08	-0.17	-0.21
Modyfikacja 3	1.42	-0.5	-2.9	-3.67	0.33	1.25	-0.29
Modyfikacja 4	-0.58	-0.17	-3.42	-2.25	-0.42	0.92	-0.96
Głos żeński							
Modyfikacja 1	1.42	-1.08	-1.42	-3.75	0.08	3.08	-2.13
Modyfikacja 2	-3.42	-0.5	-5.08	-5.58	0.42	-2.17	-3.42
Modyfikacja 3	1.08	0.67	-0.58	-0.33	0.08	2.92	-1.54
Modyfikacja 4	-2.83	-1.08	-1.42	-5.67	0.08	2.08	-2.71

1. Modyfikacje 1 i 3 wpływają na zwiększenie jasności głosu, a 2 i 4 na jej obniżenie (w modyfikacjach 1 i 3 formant F2 przesuwany był w górę, a w 2 i 4 – w dół). Przy tym zmiany te są większe dla modyfikacji 1 i 2 niż 3 i 4, co wynika z przesunięcia formantu F3 w stronę przeciwną niż przesunięcie formantu F2 w modyfikacjach 3 i 4.
2. Głosy po modyfikacji 1 i 3 ocenione zostały jako mniej gardłowe niż oryginały, przy czym dla modyfikacji 1 zmiana ta jest głębsza.
3. Modyfikacje 2 i 4 wprowadzają znaczące stłumienie głosu (stłumienie pewnych fonemów lub zakresów częstotliwości, co może prowadzić do pogorszenia zrozumiałości mowy). Może to wynikać ze zbyt dużego zbliżenia formantu F2 do formantu F1 w fonemach, gdzie odstęp między nimi już naturalnie był mały. Jednocześnie dla tych modyfikacji głosy były oceniane jako bardziej nosowe niż oryginały.
4. Średnia ocena naturalności zmodyfikowanego głosu męskiego była dla trzech modyfikacji (2, 3 i 4) taka jak dla głosu oryginalnego (nie różniła się o więcej niż 1 punkt). Dla modyfikacji 1 naturalność głosu zmodyfikowanego została oceniona wyżej niż głosu oryginalnego (głos oryginalny był ciemny i gardłowy, modyfikacja 1 zmniejszyła te parametry co prawdopodobnie wpłynęło na wyższą ocenę naturalności). Naturalność zmodyfikowanego głosu żeńskiego została w testach oceniona niżej niż głosu oryginalnego dla wszystkich modyfikacji (przy czym modyfikacje 2 i 4 zostały ocenione wyżej niż, odpowiednio, 1 i 3). Powodem jest przypuszczalnie wyższa częstotliwość podstawowa tego głosu, co spowodowało, że formanty były przesuwane o większą wartość niż w przypadku głosu męskiego, czego skutkiem były głębsze zmiany głosu, zmniejszające jego naturalność. Jest to przesłanka do tego, by wielokrotność F_0, o którą przesuwamy położenie formantu, uzależniać od jej wartości, a także od chwilowej częstotliwości środkowej modyfikowanego formantu i formantów sąsiednich (do czego przesłanką jest wniosek przedstawiony w poprzednim punkcie).

Wyniki testu odsłuchowego pokazały, że stosując zaproponowaną metodę konwersji głosu można uzyskać głos o wysokiej naturalności (w niektórych przypadkach nawet wyższej niż głosu oryginalnego). Przy tym warto zaznaczyć, że z uwag słuchaczy zebranych po przeprowadzeniu testu wynika, że słyszeli oni więcej niż dwa głosy, a więc zastosowane modyfikacje wprowadziły na tyle duże zmiany brzmienia głosu, że mówca stał się nierozpoznawalny.

Analiza ocen parametrów dotyczących jakości mowy pokazała, że zastosowane modyfikacje nie obniżają również znacząco jakości mowy. Różnice pomiędzy ocenami pierwszych trzech parametrów (występowania różnego rodzaju szumów) dla nagrań zmodyfikowanych i oryginalnych były tak małe (nie przekraczały 1 punkta), że można je uznać za nieistotne statystycznie. Większe różnice zauważono natomiast w ocenach czwartego parametru (występowania trzasków) dla trzech nagrań. Różnica w ocenach tego parametru między nagraniem oryginalnym a zmodyfikowanym wyniosła nawet 5 punktów. Występowanie trzasków nie było jednak związane z konkretną modyfikacją czy konkretnym głosem. Najprawdopodobniej jest ono wynikiem błędów estymacji ICF. Pamiętajmy, że modyfikacja ICF wprowadza zmiany w fazie sygnału, więc błędy jej estymacji mogą prowadzić do zniekształceń fazowych, objawiających się właśnie w postaci trzasków.

7. Podsumowanie

Głównym celem rozprawy było opracowanie nowych metod analizy głosu za pomocą zespolonej pulsacji chwilowej – ICF. Jest to narzędzie, które do tej pory nie było stosowane w analizie głosu. W literaturze można tu znaleźć wyłącznie przykłady zastosowań pulsacji chwilowej (IF), która jest częścią urojoną ICF. Tymczasem część rzeczywista ICF niesie również informację, a mianowicie o chwilowej szerokości pasma sygnału (IB). Jak pokazaliśmy w rozdz. 5, IB można wykorzystać do estymacji szerokości pasm formantów. Ponadto, ICF stanowi pełną reprezentację sygnału, tzn. na podstawie ICF można jednoznacznie odtworzyć reprezentowany przez nią sygnał przy zachowaniu informacji o jego fazie początkowej, a dla IF tak nie jest.

W rozprawie wykorzystaliśmy bifaktoryzację Voelckera-Kumaresana (V-KB), czyli faktoryzację sygnału na obwiednię minimalnofazową i fazor dodatnioskrętny. V-KB jest alternatywą do powszechnie stosowanej faktoryzacji AM·FM. Ma ona tę zaletę, że oba jej czynniki są zawsze analityczne, podczas gdy czynnik FM bywa analityczny tylko wtedy, gdy czynniki faktoryzacji AM·FM spełniają założenia twierdzenia Bedrosiana. Ponadto, jeżeli do estymacji częstotliwości chwilowej sygnału wykorzystamy IF fazora dodatnioskrętnego, to zamiast IF czynnika FM uzyskamy przebieg, który przyjmuje wartości dodatnie dla każdej chwili czasu. Również, $\omega_{pif}(t)$ ma bardziej gładki przebieg niż $\omega(t)$, gdyż powstaje przez usunięcie z $\omega(t)$ udziału pulsacji chwilowej obwiedni minimalnofazowej. W ten sposób eliminowaliśmy wpływ modulacji amplitudy na przebieg IF.

Analizę sygnału mowy poprzez bifaktoryzację V-K zaproponowali i jako jedyni stosowali Kumaresan i in. [KU99]. W niniejszej rozprawie przebadaliśmy właściwości czynników V-KB oraz ich zespolonych pulsacji chwilowych dla sygnałów syntetycznych, należących do klasy sygnałów 4-tonowych o tej samej obwiedni chwilowej. Wnioski z przeprowadzonych testów omówiliśmy w p. 4.6, odnosząc je również do innych metod estymacji IF, które mają na celu wyeliminowanie wpływu modulacji amplitudy na przebieg IF [LO96][OL00]. Pokazaliśmy, że zasteosowanie V-KB w analizie mowy jest uzasadnione po pierwsze tym, że sygnał mowy (a ściślej głoski dźwięczne) są prawie minimalnofazowe. Po drugie, MPE w dużym stopniu zachowuje strukturę formantową sygnału mowy. I po trzecie, IF fazora dodatnioskrętnego wskazuje na częstotliwość środkową dominującego formantu.

Zauważyliśmy również, że IF wykazuje lepsze właściwości dla sygnałów o wyższym stopniu minimalnofazowości i pokazaliśmy, że stosując filtr deemfazy można wymusić większy stopień minimalnofazowości, a tym samym poprawić właściwości estymowanej IF. Pomysł ten wykorzystaliśmy w algorytmie ekstrakcji formantów oraz estymacji F_0.

Częstotliwość podstawowa F_0, związana z percypowaną wysokością głosu, jest jednym z najważniejszych parametrów w analizie głosu. W rozprawie zaproponowaliśmy algorytm estymacji F_0 wykorzystujący zespoloną pulsację chwilową (opisany w p. 5.2). Jego zaletą jest to, że działa on potokowo, estymując F_0 dla każdej próbki sygnału mowy. Bazuje on na algorytmie zaproponowanym przez Bloka i in. [BL04], w którym estymacja F_0 przeprowadzana jest w kilku gałęziach, a pierwszym blokiem każdej gałęzi jest pasmowy filtr Hilberta. W algorytmie opisanym w niniejszej rozprawie zastosowaliśmy inny bank filtrów, zaprojektowany przez autorkę tak, by zminimalizować liczbę gałęzi algorytmu przy estymacji F_0 zmieniającej się w zakresie od 90 do 500 Hz. Zakres ten można rozszerzyć poprzez dołożenie kolejnych gałęzi algorytmu z odpowiednio zaprojektowanymi pasmowymi filtrami Hilberta. Odróżnia to zaproponowany algorytm od tradycyjnych metod, przetwarzających mowę w ramkach, w których zakres estymowanych poprawnie czestotliwości jest ograniczony przez szerokość ramki. Opracowaliśmy również metodę wyboru najlepszej estymaty, bazującą na IB przebiegów w każdej gałęzi algorytmu. Ponadto dołożyliśmy nową gałąź algorytmu, w której przeprowadzana jest klasyfikacja mowy na dźwięczną i bezdźwięczną. Zaproponowany algorytm został przetestowany pod względnem skuteczności klasyfikacji na mowę dźwięczną i bezdźwięczną oraz dokładności estymacji F_0.

Kolejnym ważnym zadaniem w analizie głosu jest estymacja częstotliwości środkowych (bądź rezonansowych) formantów. Tradycyjne metody przeprowadzające taką analizę bazują na modelu liniowym „źródło-filtr" [RA07]. W niniejszej rozprawie odeszliśmy od niego na rzecz modelu zaproponowanego przez Maragosa i in. [MA95], w którym sygnał mowy jest sygnałem wielokomponentowym, będącym superpozycją formantów. Każdy formant jest natomiast modelowany jako monokomponentowy sygnał o modulowanej amplitudzie i częstotliwości. Inaczej niż w modelu „źródło-filtr", w którym rozdziela się pobudzenie od transmitancji filtru modelującego trakt głosowy, w tym przypadku wyodrębnia się poszczególne formanty. Należało więc rozwiązać problem dekompozycji sygnału

wielokomponentowego na pojedyncze komponenty. W p. 3.3.1 przedstawiliśmy wybrane metody dekompozycji sygnałów wielokomponentowych, wskazując, że w analizie głosu najczęściej stosowane jest podejście, w którym sygnał mowy przetwarzany jest przez bank równoległych filtrów, których częstotliwości środkowe adaptują się do częstotliwości środkowych formantów. Podejście to jest popularne z praktycznego powodu. A mianowicie, wielu autorów uważa, że IF estymowana dla sygnałów szerokopasmowych i wielokomponentowych jest trudna lub wręcz niemożliwa do interpretacji. W konsekwencji stosują oni prostsze podejście, polegające na tym, że sygnał mowy jest najpierw odfiltrowywany za pomocą banku filtrów, a dopiero potem estymowana jest IF w każdym podpaśmie. Jednak tu, w p. 5.1. pokazaliśmy, że IF estymowana dla mowy dźwięcznej o paśmie ograniczonym do 7200 Hz wskazuje na częstotliwość środkową dominującego formantu, a IB jest dobrą estymatą szerokości jego pasma. Wnioski te wykorzystaliśmy do opracowania nowego, iteracyjnego algorytmu ekstrakcji formantów, w którym formanty wyodrębniane są po kolei.

Opracowany tu algorytm bazuje na metodzie HVD zaproponowanej przez Feldmana [FE06] [FE11], którą zmodyfikowaliśmy dla potrzeb analizy mowy, co opisano w p. 5.3. Częstotliwość środkowa wyodrębnionych formantów jest estymowana za pomocą IF (części urojonej ICF), a szerokości ich pasm – za pomocą IB (części rzeczywistej ICF). Tak, jak estymacja F_0, ekstrakcja formantów jest przeprowadzana próbka po próbce. Poprawność działania algorytmu została zweryfikowana eksperymentalnie (wyniki porównano z wynikami otrzymanymi za pomocą algorytmu bazującego na LP). Mimo że w zaproponowanej metodzie opóźnienia poszczególnych iteracji sumują się, odpowiednie zaprojektowanie użytych filtrów umożliwiło działanie algorytmu w czasie rzeczywistym (całkowite opóźnienie nie przekraczało 20 ms). Ograniczeniem zaproponowanej tu metody jest możliwość jej użycia wyłącznie do mowy dźwięcznej. Dla ścisłości dodajmy też, że odejście od tradycyjnego modelu „źródło-filtr” sprawia, że nie można analizować oddzielnie przebiegu pobudzenia.

Nowe algorytmy analizy najważniejszych parametrów sygnału mowy, opisane w p. 5.2 i 5.3, wykorzystujące nie stosowane dotychczas w analizie głosu narzędzie jakim jest ICF, potwierdzają pierwszą część postawionej w rozprawie tezy, że zespolona pulsacja chwilowa jako reprezentacja sygnału mowy daje nowe, dotychczas nieznane możliwości jego analizy. Dodatkowo, w rozdz. 5 zaproponowaliśmy wykorzystanie ICF do estymacji stopnia

minimalnofazowości sygnału mowy. Może to znaleźć zastosowania w fonetyce i foniatrii, a prawdopodobnie również w badaniach nad mową zaburzoną. Badania takie jak dotąd nie były prowadzone.

Wyniki analizy sygnału mowy za pomocą ICF były podstawą do osiągnięcia drugiego celu rozprawy i zarazem potwierdzenia drugiej części tezy, a mianowicie, że proste modyfikacje zespolonej pulsacji chwilowej czynników bifaktoryzacji Voelckera-Kumaresana sygnału mowy pozwalają na konwersję głosu mówcy, czyli zmianę jego brzmienia.

W rozdz. 6 zaproponowaliśmy dwie metody konwersji głosu. W pierwszej modyfikowana jest ICF estymowana dla całego sygnału mowy, w drugiej – ICF jest estymowana dla poszczególnych formantów. Pierwsza metoda jest znacznie mniej złożona obliczeniowo, gdyż nie wymaga ekstrakcji formantów. Można za jej pomocą uzyskać zarówno delikatne zmiany brzmienia (np. bardziej jasny głos), jak również zmienić głos tak, by nie był rozpoznawalny, a więc pozwalający np. na anonimizację mówcy. Jednak, jak pokazały testy odsłuchowe, głębsze zmiany powodują tu znaczne obniżenie naturalności uzyskanego głosu. W przeciwieństwie do tego druga metoda, chociaż bardziej złożona obliczeniowo, pozwoliła nie tylko uzyskać delikatne i głębokie zmiany głosu, ale także zachować jego wysoką naturalność, a to było drugim celem rozprawy.

Bibliografia

[BA96] G. Baudoin, Y. Stylianou, “On the transformation of the speech spectrum for voice conversion,” *Proc. Int. Conf. Spoken Language Processing*, 1996, pp. 1405-1408.

[BE06] A. Bech, N. Zacharov, *Perceptual Audio Evaluation: Theory, Method and Application*. Chichester, England: John Wiley and Sons, 2006.

[BE07] J.W. Beauchamp, *Analysis, Synthesis and Perception of Musical Sound: the Sound of Music*. New York, NY: Springer, 2007.

[BE63] E. Bedrosian, “A product theorem for Hilbert transforms.” *Proceedings of the IEEE*, vol. 51, pp.868-869, May 1963.

[BL04] M. Blok, M. Rojewski, A. Sobociński, “Nowy estymator tonu krtaniowego.” *Zeszyty Naukowe Wydziału Elektroniki Telekomunikacji i Informatyki PG*, vol. 2, pp. 125-134, 2004

[BO04] A. Bouzid, N. Ellouze, “Empirical mode decomposition of voiced speech signal,” *Proc. 1st Int. Symp. Control, Communications and Signal Processing*, 2004, pp. 603-606.

[BO92a] B. Boashash, “Estimating and interpreting the instantaneous frequency of a signal – part 1: fundamentals.” *Proceedings of the IEEE*, vol. 80, no.4, pp. 520-538, April 1992.

[BO92b] B. Boashash, “Estimating and interpreting the instantaneous frequency of a signal – part 2: algorithms and applications.” *Proceedings of the IEEE*, vol. 80, no.4, pp. 540-568, April 1992.

[BR11] S. Braun, M. Feldman, “Decomposition of non-stationary signals into varying time scales: Some aspects of the EMD and HVD methods.” *Mechanical Systems and Signal Processing*, vol. 25, pp.2608-2630, 2011.

[BR74] J.L. Brown, “Analytic signals and product theorems for Hilbert transforms.” *IEEE Transactions on Circuits and Systems*, vol. 21, pp. 790–792, November 1974.

[BR86] J.L. Brown, “A Hilbert transform product theorem.” *Proceedings of the IEEE*, vol. 74, pp.520-521, March 1986.

[CA37] J. Carson, T. Fry, "Variable frequency electric circuit theory with application to the theory of frequency modulation." *Bell System Technical Journal*, vol. 16, pp.513-540, 1937.

[CA73] G.D. Cain, "Hilbert transform relations for products." *Proceedings of the IEEE*, vol. 61, pp. 663-664, May 1972.

[CH01] A. Chodkowski, *Encyklopedia Muzyki*. Warszawa, Polska: Wydawnictwo Naukowe PWN, 2001.

[CH02] A. de Cheveigné, H. Kawahara, "YIN, A fundamental frequency estimator for speech and music." *Journal of the Acoustical Society of America*, vol. 111, pp. 1917-1930, April 2002.

[CH05] A. de Cheveigné, *Pitch: Neural Coding and Perception*. New York, NY: Springer, 2005

[CO85] L. Cohen, T. Posch, "Positive time-frequency distribution functions." *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, no. 1, pp. 31-37, 1985.

[CO89] L. Cohen, L. Chongmoon, "Instantaneous frequency and time-frequency distributions," *Proc. IEEE Int. Symp. Circuits and Systems*, 1989, pp.1231-1234.

[CO92] L. Cohen, "What is a multicomponent signal?," *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 1992, pp. 113-116.

[CO95] L. Cohen. *Time-Frequency Analysis*. Englewood Cliffs, NJ: Prentice Hall, 1995.

[CO99] L. Cohen, P. Loughlin, D. Vakman, "On an ambiguity in the definition of the amplitude and phase of a signal." *Signal Processing*, vol. 79, no. 3, pp. 301-307, December 1999.

[CZ01] A. Czyżewski, *Dźwięk Cyfrowy: Wybrane Zagadnienia Teoretyczne, Technologia, Zastosowania*. Warszawa, Polska: Akademicka Oficyna Wydawnicza EXIT, 2001.

[DA04] P. Dalka, M. Dąbrowski, "System rozpoznawania dźwięków instrumentów muzycznych." *Zeszyty Naukowe Wydziału Elektrotechniki i Automatyki Politechniki Gdańskiej,* no. 20, pp. 29-34, 2004.

[DE93] J.R. Deller, J.G. Proakis, J.H.L. Hansen, *Discrete-Time Processing of Speech Signals*. New York, NY: Macmillan Publishing Company, 1993.

[DE99] D. Deutch, *The Psychology of Music*. San Diego, CA: Gulf Professional Publishing, 1999.

[FA60] G. Fant, *Acoustic Theory of Speech Production*. Haga, Holandia: Mouton & Co, 1960.

[FE06] M. Feldman, “Time-varying vibration decomposition and analysis based on the Hilbert transform.” *Journal of Sound and Vibration*, vol. 295, pp. 518-530, 2006.

[FE08] M. Feldman, “Theoretical analysis and comparison of Hilbert transform decomposition methods.” *Mechanical Systems and Signal Processing*, vol. 28, pp. 509-519, 2008.

[FE11] M. Feldman, *Hilbert Transform Applications in Mechanical Vibration*. Chichester, UK: John Wiley and Sons, 2011.

[FL33] H. Fletcher, W.J. Munson. “Loudness, its definition, measurement and Calculation.” *Journal of the Acoustical Society of America*, vol. 5, no. 2, pp. 82– 108, October 1933.

[GA46] D. Gabor, “Theory of communication.” *Journal of the IEE*, vol. 93, pp. 429-457, 1946.

[GA07] Y. Gao, Z. Yang, “Pitch modification based on syllable units for voice morphing system,” *Proc. of IFIP Int. Conf. Network and Parallel Computing*, 2007, pp. 135-139.

[GE03] D. Gerhard, “Pitch Extraction and Fundamental Frequency: History and Current Techniques.” Technical report, Dept. of Computer Science, University of Regina, 2003.

[GI05] F. Gianfelici, G. Biagetti, P. Crippa, C. Turchetti, “AM-FM decomposition of speech signals: an asymptotically exact approach based on the iterated Hilbert transform,” *Proc. 13th Workshop on Statistical Signal Processing*, 2005, pp. 333-338.

[GI07] F. Gianfelici, G. Biagetti, P. Crippa, C. Turchetti, “Multicomponent AM-FM representations: an asymptotically exact approach.” *IEEE Transactions on*

Audio, Speech and Language Processing, vol. 15, no. 2, pp. 823-837, 2007.

[HA03] S.L. Hahn, "On the uniqueness of the definition of the amplitude and phase of the analytic signal." *Signal Processing*, vol. 83, no. 8, pp. 1815-1820, August 2003.

[HA07] S.L. Hahn, "The history of applications of analytic signals in electrical and radio engineering." *Proc. EUROCON Int. Conf. Computer as a Tool*, 2007, pp. 2627-2631.

[HA59] S.L. Hahn, "The instantaneous complex frequency concept and its application to the analysis of building up of oscillations in oscillators." *Proceedings of Vibration Problems*, no. 1, pp. 24-46, 1959.

[HA64] S.L. Hahn, "Complex variable frequency in electric circuit theory." *Proceedings of the IEEE (Letters)*, vol. 52, no. 6, pp. 735-736, June 1964.

[HA94] H.M. Hanson, P. Maragos, A. Potamianos, "A system for finding speech formants and modulations via energy separation." *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 3, pp. 436-443, July 1994.

[HA95] S.L. Hahn, *Hilbert Transforms in Signal Processing*. Norwood, MA: Artech House, 1995.

[HE06a] E. Hermanowicz, M. Rojewski, D. Tkaczuk, "Modyfikacja wysokości brzmienia dźwięku świergotowego na podstawie jego zespolonej reprezentacji dynamicznej," *Krajowe Sympozjum Telekomunikacji i Teleinformatyki*, 2006, dokument elektroniczny.

[HE06b] E. Hermanowicz, M. Rojewski, "Pitch shifter based on complex dynamic representation rescaling and direct signal synthesis." *Bulletin of the Polish Academy of Sciences: Technical Sciences*, vol.54, no.4, pp. 499-504, December 2006.

[HE07a] E. Hermanowicz, M. Rojewski, "Application of Bedrosian Condition to pitch-shifting performance evaluation for chirp sounds." *Proc. 15th Int. Conf. Digital Signal Processing*, 2007, pp. 523-526.

[HE07b] E. Hermanowicz, M. Rojewski, "On Bedrosian condition in application to chirp sounds." *Proc. 15th European Signal Processing Conference EUSIPCO*, 2007, pp. 1221-1225.

[HE88] E. Hermanowicz, M. Rojewski, “Resyntezer przebiegu fazy chwilowej jako transmodulator cyfrowy.” *XI Krajowa Konferencaj Teoria Obwodów i Układy Elektroniczne*, 1988, tom 2, pp. 170-175.

[HE89] E. Hermanowicz, M. Rojewski, “Moduły cyfrowego transmodulatora opartego o resyntezę przebiegu fazy chwilowej.” *Przegląd Telekomunikacyjny*, r. 62, nr 8, pp. 236-238, 1989.

[HE91] E. Hermanowicz, M. Rojewski, “The minimum-phase demodulate and its application to autoregressive analysis of bandpass signal.” *Signal Processing*, vol. 25, no. 1, pp. 1-9, 1991.

[HU98] N. E. Huang et al., “The empirical mode decomposition and Hilbert spectrum for nonlinear and non-stationary time series analysis”, *Proceedings of the Royal Society London A*, vol. 454, pp. 903-995, 1998.

[ITU-T96] ITU-T P.800, “Methods for subjective determination of transmission quality.” 1996.

[JA07] S. Jang, S. Choi, H. Kim, H. Choi, Y. Yoon, “Evaluation of performance of several established pitch detection algorithms in pathological voices.” *Proc. 29th Annual Int. Conf. IEEE EMBS*, 2007, pp. 620-623.

[JO03] K. Johnson, *Acoustics and Audiotory Phonetics.* Malden, MA: Blackwell Publishing, 2003.

[JO90] G. Jones, B. Boashash, “Instantaneous frequency, instantaneous bandwidth and the analysis of multicomponent signals,” *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, 1990, pp. 2467-2470.

[KA02] K. Kahrs, K. Brandenburg, *Applications of Digital Signal Processing to Audio and Acoustics*. New York, NY: Kluwer Academic Publishers, 2002.

[KA06] M. Kaniewska. “Ekstraktor Tonu Krtaniowego do Protezy Mowy.” Praca dyplomowa magisterska, Katedra Systemów Multimedialnych PG, Gdańsk, 2006.

[KA07] M. Kaniewska, “Porównanie działania metod YIN i MAWT w estymacji tonu krtaniowego mowy zaburzonej,” *Zeszyty Naukowe Wydziału ETI Politechniki Gdańskiej, seria Technologie Informacyjne, tom 14, 691-698*, 2007.

[KA08a] M. Kaniewska, “Speech formant frequency and pitch estimation using

instantaneous complex frequency," *Proc. Int. Conf. Signals and Electronic Systems*, 2008, pp. 493-496.

[KA08b] M. Kaniewska, "On the use of instantaneous complex frequency for pitch and formant tracking," *Proc. Int. Conf. NTAV/SPA*, 2008, pp. 61-65..

[KA09] M. Kaniewska, "On the use of instantaneous complex frequency for analysis and modification of simple sounds," *Proc. Conf. Ph.D. Research in Microelectronics and Electronics*, 2009, pp. 340-343.

[KA10a] M. Kaniewska, "Human voice modification using instantaneous complex frequency," *Proc. 128th AES Convention*, 2010, dokument elektroniczny.

[KA10b] M. Kaniewska, "Voice transformations through instantaneous complex frequency modifications," Proc. EUSIPCO, 2010, pp. 90-94.

[KA10c] M. Kaniewska, "Instantaneous complex frequency for pipeline pitch estimation," *Proc. Int. Conf. NTAV/SPA*, 2010, pp. 83-88.

[KA11] M. Kaniewska, "On-line pitch estimation using instantaneous complex frequency," *Proc. 20th ECCTD*, 2011, dokument elektroniczny.

[KI98] B.E.D. Kingsbury, N. Morgan, S. Greenberg, "Robust speech recognition using the modulation spectrogram." *Speech communication*, vol. 25, pp. 117-132, 1998.

[KL03] W.B. Kleijn, T. Bäckström, P. Alku, "On line spectral frequencies." *IEEE Signal Processing Letters*, vol. 10, no. 3, pp. 75-77, March 2003.

[KO01] B. Kostek, A. Czyżewski, Representing Musical Instrument Sounds for Their Automatic Classification." Journal of Audio Engineering Society, vol. 49, no. 9, pp. 768-785, 2001.

[KU03a] A. Kumar, A. Verma, "Using phone and diphone based acoustic models for voice conversion: a step towards creating voice fonts," *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2003, pp. I-720-723.

[KU03b] R. Kumaresan, G.K. Allu, J. Swaminathan, Y. Wang, "Decomposition of a bandpass signal and its applications to speech processing," *Proc. 37th Asilomar Conf. Signals, Systems and Computers*, 2003, pp. 2078-2082.

[KU04] A. Kumar, A. Verma, "Articulatory class based spectral envelope representation for voice fonts," *Proc. Int. Conf. Multimedia and Expo*, 2004,

pp. 1647-1650.

[KU99] B. Kumaresan, A. Rao, “Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications.” *Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1912-1924, March 1999.

[LI58] D.A. Linden, “A note concerning instantaneous frequency.” *Proceedings of the IRE (Correspondence)*, vol. 46, p. 1970, December 1958.

[LO94] P.J. Loughlin, J. Pitton, L. Atlas, “Construction of positive time-frequency distributions.” *IEEE Transactions on Signal Processing*, vol. 42, no. 10, pp. 2607-2705, October 1994.

[LO96] P.J. Loughlin, B. Tacer, “On the amplitude- and frequency-modulation decomposition of signals.” *Journal of the Acoustical Society of America*, vol. 100, no. 3, pp. 1594-1601, September 1996.

[LO98] P.J. Loughlin, “The time-dependent weighted average instantaneous frequency,” *Proc. IEEE-SP Int. Symp. Time-Frequency and Time-Scale Analysis*, 1998, pp. 97-100.

[LU96] S. Lu, P.C. Doershik, “Nonlinear modeling and processing of speech based on sums of AM-FM formant models.” *IEEE Transactions on Signal Processing*, vol. 44, pp. 773-782, April 1996.

[MA05] J. Malkin, Xiao Li, J. Bilmes, “A graphical model for formant tracking,” *IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2005, pp. 913-916.

[MA74] L. Mandel, “Interpretation of instantaneous frequency.” *American Journal of Physics*, vol. 42, pp. 840-846, 1974.

[MA93] P. Maragos, J.F. Kaiser, T.F. Quatieri, “On amplitude and frequency demodulation using energy operators.” *IEEE Transactions on Signal Processing*, vol. 41, no. 4, pp. 1532-1550, 1993.

[MA95] P. Maragos, A. Potamianos, “Speech formant frequency and bandwidth tracking using multiband energy demodulation,” *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, 1995, pp. 784-787.

[MO03] B.C.J. Moore, *An Introduction to the Psychology of Hearing*. London, UK: Academic Press, 2003.

[MO95] B.C.J. Moore, *Hearing*. San Diego, CA: Academic Press, 1995.

[MO97] B.C.J. Moore, B.R. Glasberg, T. Baer, "A model for prediction of thresholds, loudness and partial loudness." *Journal of the Acoustical Society of America*, vol. 45, pp. 224-240, 1997.

[NU66] A. Nuttal, "On the quadrature approximation to the Hilbert transform of modulated signals." *Proceedings of the IEEE*, vol. 54, pp. 1458-1459, October 1966.

[OL00] P.M. Oliveira, V. Barroso, "Definitions of instantaneous frequency under physical constraints," *Journal of the Franklin Institute*, vol. 337, pp. 303-316, 2000.

[OL98a] P.M. Oliveira, V. Barroso, "Instantaneous frequency of mono and multicomponent signals," *Proc. IEEE-SP Int. Symp. Time-Frequency and Time-Scale Analysis*, 1998, pp. 105-108.

[OL98b] P.M. Oliveira, V. Barroso, "On the concept of instantaneous frequency," *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 1998, pp. 2241-2244.

[OL99] P.M. Oliveira, V. Barroso, "Instantaneous frequency of multicomponent signals." *IEEE Signal Processing (Letters)*, vol. 6, no. 4, pp.81-83, April 1999.

[OP89] A.V. Oppenheim, R.W. Schafer, J.R. Buck, *Discrete-Time Signal Processing*. Upper Saddle River, NJ: Prentice Hall, 1989.

[PL69] R. Plomp, H.J.M. Steeneken, "Effect of phase on the timbre of complex tones." *Journal of the Acoustical Society of America*, vol. 46, no. 2, pp. 409-421, 1969.

[PL95] F. Plante, G. Meyer, W.A. Ainsworth, "A pitch extraction reference database," *Proc. EUROSPEECH*, 1995, pp. 837-840.

[PO46] B. Van der Pol, "The fundamental principles of frequency modulation." *Proc of the IEE*, vol. 93 (III), pp.153-158, 1946.

[PO95] A. Potamianos, "Speech Processing Applications Using an AM-FM Modulation Model." Ph.D. Thesis, The Division of Applied Sciences, Harvard University, 1995.

[PO97] M.A. Poletti, “The homomorphic analytic signal.” *IEEE Transactions on Signal Processing*, vol. 45, no. 8, pp. 1943-1953, August 1997.

[QU92] T.F. Quatieri, R.J. McAulay, “Shape invariant time-scale and pitch modification of speech.” *IEEE Transactions on Signal Processing*, vol. 40, pp. 497 – 510, March 1992.

[RA00] A. Rao, R. Kumaresan, “On decomposing speech into modulated components.” *IEEE Transactions on Speech and Audio Processing*, vol. 8, no.3, pp. 240-254, May 2000.

[RA07] L.R. Rabiner, R.W. Schafer. *Introduction to Digital Speech Processing*. Boston, MA: NOW Publishers, 2007.

[RA76] L.R. Rabiner, M.J. Cheng, A.E. Rosenberg, C.A. McGonegal, “A comparative performance study of several pitch detection algorithms.” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, no. 5, pp. 399-418, October 1976.

[RE04] D. Rentzos, S. Vaseghi, Qin Yan, Ching-Hsiang Ho, “Voice conversion through transformation of spectral and intonation features,” *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2004, pp. I-21-24.

[RE07] B. Resch, M. Nilsson, A. Ekman, W.B. Kleijn, “Estimation of the instantaneous pitch of speech.” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 3, March 2007.

[RI66] A.W. Rihaczek, “Hilbert transforms and the complex representation of real signals.” *Proceedings of the IEEE*, vol. 54, pp.434-435, March 1966.

[RO06] R.M. Roark, “Frequency and voice: Perspectives in the time domain.” *Journal of Voice*, vol. 20, pp. 325-354, 2006.

[RO08] D. Rochesso, P. Polotti, *Sound to Sense, Sense to Sound: a State of the Art in Sound and Music Computing*, Berlin, Germany: Logos Verlag, 2008.

[RO10] M.Rojewski, Notatka niepublikowana z dn. 5.06.2010.

[RO94] M. Rojewski, “Nowa definicja i bezbłędna estymacja dyskretnej zespolonej pulsacji chwilowej,” *X Krajowe Sympozjum Telekomunikacji*, 1994, pp. 453-460.

[SA00] S. Saliu, “Definition of instantaneous frequency on real signals,” *Proc.*

European Signal Processing Conference, 2000, pp. 343-346.

[SH09] D. Sharma, P.A. Naylor, "Evaluation of pitch estimation in noisy speech for application in non-intrusive speech quality assesment," *Proc. European Signal Processing Conference*, 2009.

[SL95] J. Slifka, T.R. Anderson, "Speaker modification with LPC pole analysis." *Proc IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 1995, pp. 211-226.

[ST40] S.S. Stevens, J. Volkman, "The relation of pitch to frequency: a revised scale." *American Journal of Psychology*, vol. 53, no. 3, pp. 329-353, 1940.

[ST98] Y. Stylianou, O. Cappé, E. Moulines, "Continuous probabilistic transform for voice conversion." *IEEE Transactions on Speech and Audio Processing*, vol. 6, pp. 131-142, March 1998.

[SU02] P. Susini P., S. McAdams, B. Smith, "Global and continous loudness estimation of time-varying levels." *Acta Acustica*, vol. 88, pp. 536-548, 2002.

[SU06] H. Suzuki et al., „Instantaneous frequencies of signals obtained by the analytic signal method." *Acoustical Science and Technology*, vol. 27, no. 3, pp. 163-170, 2006.

[ŚW10] K. Świder, "Wykorzystanie Zespolonej Pulsacji Chwilowej do Odtwarzania Synchronizacji Symbolowej Sygnałów QPSK." Rozprawa doktorska, Wydział Elektroniki, Telekomunikacji i Informatyki Politechniki Gdańskiej, Gdańsk 2010.

[TA09] L. Tan, L. Yang, D. Huang, "Necessary and sufficient conditions for the Bedrosian identity." *Journal of Integral Equations and Applications*, vol. 21, no. 1, pp. 77-94, Spring 2009.

[TA88] R. Tadeusiewicz, *Sygnał Mowy*. Warszawa, Polska: Wydawnictwa Komunikacji i Łączności, 1988.

[TA97] K. Tanaka, M. Abe, "A new fundamental frequency modification algorithm with transformation of spectrum envelope according to F0," *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 1997, pp. 951-954.

[TU00] O. Türk, "New Methods for Voice Conversion." M.S. Thesis, Electrical and Electronics Engineering, Boğaziçi University, 2000.

[VA78] D. Vakman, L.A. Vaĭnshteĭn, “Amplitude, phase, frequency – fundamental concepts of oscillation theory.” *Soviet Physics Uspekhi*, vol. 20, pp. 1002–1016, 1978.

[VA96] D. Vakman, “On the analytic signal, the Teager-Kaiser energy algorithm, and other methods for defining amplitude and frequency.” *IEEE Transactions on Signal Processing*, vol. 43, no. 4, pp.797-797, April 1996.

[VA98] D. Vakman, *Signals, Oscillations and Waves: A Modern Approach*. Boston, MA: Artech House 1998.

[VI48] J. Ville, “Theorie et application de la notion de signal analytic.” *Cables et Transmissions,* vol. 2A, no. 1, pp. 61-74, 1948.

[VO66a] H.B. Voelcker, “Toward a unified theory of modulation part I: Phase-envelope relationships.” *Proceedings of the IEEE*, vol. 54, no. 3, pp. 340–354, March 1966.

[VO66b] H.B. Voelcker, “Towards a unified theory of modulation part II: Zero manipulation.” *Proceedings of the IEEE,* vol. 54, no. 5, pp. 735–755, May 1966.

[WE98] D. Wei, A.C. Bovik, “On the instantaneous frequencies of multicomponent AM-FM signals.” *IEEE Signal Processing Letters*, vol. 9, no. 4, pp. 84-86, April 1998.

[WO99] Wonchul Nho, P.J. Loughlin, “When is instantaneous frequency the average frequency at each time?” *IEEE Signal Processing (Letters)*, vol. 6, no.4, pp. 78-80, April 1999.

[WR00] A.A. Wrench, “A multi-channel/multi-speaker articulatory database for continous speech recognition research.” Phonus, Research Report No.5, pp.1-13, 2000.

[XU06] Y. Xu, D. Yan, “The Bedrosian identity for the Hilbert transform of product functions.” *Proceedings of the American Mathematical Society*, vol. 134, no. 9, pp. 2719-2728, September 2006.

[YE09] B. Yegnanarayana, K.S.R. Murty, “Event-based instantaneous fundamental frequency estimation from speech signals.” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 614-624, May 2009.

[ZA10] C. Zawidzki, "Estymator wysokości chwilowej dźwięku w oparciu o jego pulsację chwilową." Praca dyplomowa magisterska, Katedra Systemów Multimedialnych PG, Gdańsk, 2010.

[ZW65] E. Zwicker, B. Scharf, "A model of loudness summation." *Psychological Review*, vol. 72, pp. 3-26, 1965.

[WWW1] Strona University College London, Division of Psychology and Language Sciences: http://www.langsci.ucl.ac.uk/ipa, 28.10.2011

[WWW2] Centrum Logopedyczne PL: http://www.logopedia.pl, 28.10.2011

[WWW3] Strona domowa G. Morrisona: http://geoff-morrison.net/, 28.10.2011

Załącznik A – zawartość płyty CD

1. Rozprawa doktorska w formacie PDF.
2. Opis rozprawy zawierający streszczenie w języku polskim i angielskim.
3. Materiał do testów odsłuchowych, opisanych w podp. 6.3.3.1.
    a) wykorzystane nagrania w kolejności odtwarzania wraz z opisem;
    b) formularz dla słuchaczy.