Politechnika Poznańska

Wydział Elektroniki i Telekomunikacji

Katedra Telekomunikacji Multimedialnej i Mikroelektroniki

Rozprawa doktorska

Kompresja cyfrowych sygnałów fonicznych z łącznym wykorzystaniem rozszerzania widma i modelowania

Tomasz Żernicki

Promotor: prof. dr hab. inż. Marek Domański

Poznań, 2010

Spis treści

Spis treści **i**

1 Wstęp **5**

1.1. Zakres tematyczny pracy 5

1.2. Cele i teza pracy 10

1.2.1. Problem naukowy 10

1.2.2. Teza rozprawy 12

1.2.3. Cel naukowy 12

1.3. Metodologia badań 12

1.4. Pomiar jakości sygnałów fonicznych 14

1.5. Przegląd pracy 18

2 Przegląd rozwiązań dotyczących technik rozszerzania widma i modelowania sygnałów **20**

2.1. Wprowadzenie 20

2.2. Przegląd wybranych narzędzi z zakresu modelowania dźwięku 21

2.2.1. Modelowanie sinusoidalno-szumowe 22

2.2.2. Modelowanie obwiedni widma 26

2.3. Psychoakustyczne aspekty percepcji składowych dużych częstotliwości . . . 27

2.3.1. Wysokość dźwięku oraz struktura harmoniczna 27

2.3.2. Głośność 28

2.3.3. Barwa dźwięku 28

2.3.4. Podsumowanie 28

2.4. Klasyfikacja metod rozszerzania widma 29

2.4.1. Metody oparte na replikacji widma 29

2.4.2. Metody oparte na modelowaniu sinusoidalnym 31
2.4.3. Metody oparte na zniekształceniu nieliniowym 31
2.5. Spectral Band Replication - przykład techniki opartej na replikacji widma . 31
2.5.1. Wprowadzenie 31
2.5.2. Proces kodowania 32
2.5.3. Proces dekodowania 37
2.6. Zniekształcenia wprowadzanie przez techniki rozszerzania widma 42
2.6.1. Szorstkość dźwięku 43
2.6.2. Dudnienia 43
2.6.3. Składniki tonalne o zmiennej częstotliwości 44
2.6.4. Problem klasyfikacji składowych tonalnych i szumowych 45
2.6.5. Rekonstrukcja brakujących składowych tonalnych w dekoderze SBR 47
2.7. Poprawna rekonstrukcja składowych tonalnych w technice SBR 49
2.8. Accurate Spectral Replacement - przykład techniki opartej na modelowaniu sinusoidalno-szumowym 50
2.9. Ślepe techniki rekonstrukcji składowych dużych częstotliwości 51
2.10. Podsumowanie i uzasadnienie podjęcia tematu pracy 52

3 Metoda rozszerzania widma wykorzystująca modelowanie sinusoidalne 54
3.1. Wprowadzenie 54
3.2. Koncepcja metody 55
3.3. Opis proponowanej techniki 56
3.4. Separacja składowych tonalnych 58
3.4.1. Wejściowy filtr górnoprzepustowy 59
3.4.2. Wybór składowych tonalnych do separacji 59
3.4.3. Usuwanie składowych tonalnych 63
3.5. Ograniczenia separacji składowych tonalnych 68
3.6. Detektor tonalności 70
3.7. Strategia kodowania parametrów modelu sinusoidalnego 72
3.7.1. Wstęp 72
3.7.2. Sterowanie pracą kodera 72
3.7.3. Wybór trajektorii sinusoidalnych do zakodowania 77
3.7.4. Kodowanie Huffmana 80

3.8. Proces dekodowania sygnału . . . 83
3.9. Właściwości konstrukcyjne kodeka . . . 88
3.9.1. Zgodność proponowanej techniki BWE z MPEG-4 AAC HE . . . 88
3.9.2. Określenie parametrów wstępnych modelu sinusoidalnego . . . 90
3.9.3. Kwantyzacja amplitud i częstotliwości . . . 91
3.9.4. Proponowana strategia ograniczenia liczby trajektorii sinusoidalnych 92
3.10. Podsumowanie . . . 95

4 Metoda rozszerzania widma wykorzystująca skalowanie częstotliwości 96
4.1. Wprowadzenie . . . 96
4.2. Koncepcja metody . . . 97
4.3. Analiza sygnału w dekoderze . . . 99
4.4. Opis proponowanej techniki BWE . . . 102
4.4.1. Wprowadzenie . . . 102
4.4.2. Estymacja parametrów . . . 104
4.4.3. Skalowanie częstotliwości . . . 105
4.4.4. Skalowanie amplitudowej obwiedni widma . . . 106
4.5. Rekonstrukcja sygnału . . . 111
4.6. Kodowanie parametrów . . . 111
4.6.1. Kodowanie trajektorii częstotliwości podstawowych . . . 112
4.6.2. Kodowanie współczynników skalujących obwiednię widmową . . . 112
4.6.3. Reprezentacja składowej szumowej . . . 113
4.7. Pozostałe bloki funkcjonalne kodera . . . 114
4.7.1. Estymacja częstotliwości podstawowych f_0 . . . 114
4.7.2. Wybór trajektorii dla procesu modulacji . . . 115
4.7.3. Detektor tonalności . . . 116
4.8. Przykład działania techniki i wyniki eksperymentalne . . . 117
4.9. Podsumowanie . . . 120

5 Ocena eksperymentalna 121
5.1. Wprowadzenie . . . 121
5.2. Przygotowanie testów odsłuchowych . . . 121
5.3. Analiza statystyczna . . . 123

5.4. Wyniki - ocena subiektywna . . . 124
5.4.1. Porównanie kodeków wykorzystujących techniki rozszerzania widma dla sygnałów muzycznych . . . 124
5.4.2. Porównanie kodeków wykorzystujących techniki rozszerzania widma dla sygnałów zawierających silne składowe tonalne . . . 126
5.5. Wyniki - ocena obiektywna . . . 128
5.6. Dyskusja wyników . . . 130

6 Podsumowanie i dyskusja 134
6.1. Wnioski . . . 134
6.2. Oryginalne osiągnięcia . . . 136
6.3. Dalsze badania . . . 139

Spis literatury 140

A Kodowanie parametrów modelu sinusoidalnego 152
A.1. Badanie zależność prędkości bitowej od długości bloku ramek objętych kodowaniem tą samą tablicą Huffmana . . . 152
A.2. Badanie rozkładu funkcji prawdopodobieństwa dla skwantowanych parametrów modelu sinusoidalnego . . . 156

B Implementacja modelowania sinusoidalnego 157
B.1. Analiza sinusoidalna . . . 157
B.2. Śledzenie trajektorii sinusoidalnych . . . 157
B.3. Synteza sygnału . . . 159
B.4. Kompresja parametrów modelu sinusoidalnego . . . 159
B.4.1. Kodowanie trajektorii sinusoidalnych . . . 159
B.4.2. Kwantyzacja parametrów modelu sinusoidalnego . . . 160

C Testy odsłuchowe dla koderów sinusoidalnych 162

D Testy odsłuchowe dla proponowanych technik rozszerzania widma 166
D.1. Porównanie kodeków wykorzystujących techniki rozszerzania widma dla sygnałów muzycznych . . . 166

D.2. Porównanie kodeków wykorzystujących techniki rozszerzania widma dla sygnałów zawierających silne składowe tonalne 179
D.3. Sygnały foniczne wykorzystane w badaniach eksperymentalnych - płyta CD 180

Spis rysunków **181**

Spis tabel **190**

Podziękowania

Składam szczególne podziękowania mojemu promotorowi prof. dr hab. inż. Markowi Domańskiemu, za cenne wskazówki i pomoc w czasie pisania pracy doktorskiej.

Serdecznie dziękuję dr inż. Maciejowi Bartkowiakowi za wszelką pomoc, wsparcie i dobrą radę.

Dziękuję również wszystkim uczestnikom testów odsłuchowych za trud, który przyczynił się do zebrania materiałów do tej pracy.

Pragnę gorąco podziękować mojej Rodzinie, za cierpliwość i nieustanne wsparcie. Annie dziękuję za wiarę w to, że się uda.

Tomasz Żernicki

Abstract

This dissertation deals with low-bitrate audio coding using bandwidth extension techniques. Two new techniques have been proposed in order to improve encoding of high frequency tonal components.

First technique was implemented as an extension to MPEG-4 AAC HE. This technique exploits jointly spectral band replication and sinusoidal modeling. Second technique uses frequency scaling of narrowband signals for reconstruction of high frequency tonal components.

The author has implemented both codecs and software. Compression efficiency has been assessed in extensive experimental tests with usage of standard audio excerpts. This quality assessment has been performed using standard listening test according to BS.1534 ITU-R Recommendation.

Results obtained during listening tests proved that both prepared techniques outperform standard MPEG-4 AAC HE technique. This improvement is significant for excerpts with strong tonal components.

Spis symboli i skrótów

Skróty	Opis	Strona
3GPP	3rd Generation Partnership Project	13
AAC	Zaawansowane kodowanie dźwięku ang. *Advanced Audio Coding*	9
AMR-WB+	Extended Adaptive Multi Rate – WideBand	10
ANOVA	analiza wariancji (ang. *analysis of variance*)	123
ASR	Accurate Spectral Replacement	50
ATRAC	Adaptive Transform Acoustic Coding	6
BWE	technika rozszerzania widma (ang. *Bandwidth Extension*)	8
DRM	Digital Radio Mondiale	9
EBU	Europejska Unia Nadawców ang. *European Broadcasting Union*	9
f_{BWE}	granica podziału pomiędzy zakresem działania koderów podstawowych oraz koderów rozszerzania widma	8
LF	małe częstotliwości (ang. *low frequency*)	8
LPC	liniowe kodowanie predykcyjne (ang. *Linear Predictive Coding*)	
HF	duże częstotliwości (ang. *high frequency*)	8
MPEG	Moving Picture Experts Group	9
MPEG-4 AAC HE	MPEG-4 High Efficiency Advanced Audio Coding	9
MUSHRA	MUltiple Stimuli with Hidden Reference and Anchor	15
PEAQ	Perceptual Evaluation of Audio Quality	14

Skróty	Opis	Strona
QMF	zespół lustrzanych filtrów kwadraturowych ang. *Quadrature Mirror Filter*	32
QIFFT	Quadratically Interpolated FFT	60
SBR	technika replikacji pasma częstotliwościowego ang. *Spectral Band Replication*	9
SMR	stosunek sygnał-próg maskowania (ang. *signal-to-mask ratio*)	77
SSC	MPEG-4 SSC (ang. *Sinusoidal Coding*)	13
STFT	krótkookresowa transformacja Fouriera (ang. *short-time Fourier transform*)	23
STSA	krótkookresowe tłumienie widmowe (ang. *short-time Spectral Attenuation*)	64
TNR	stosunek enegrii tonalnej do szumowej (ang. *Tonal-to-Noise Ratio*)	36
t_R	czas swobodnego dostępu (ang. *random access*)	80
USAC	Unified Speech and Audio Coding	10

Rozdział 1

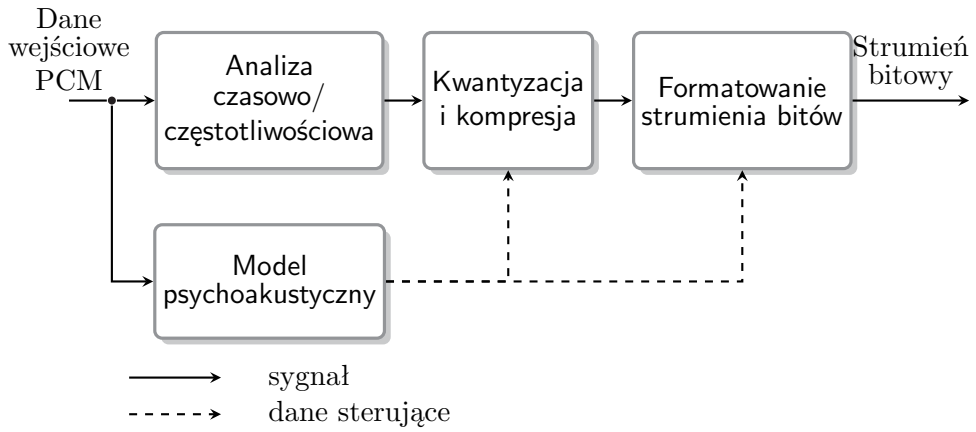
Wstęp

1.1. Zakres tematyczny pracy

W pracy poruszono zagadnienia kompresji dźwięku szerokopasmowego przesyłanego z małą prędkością bitową. W szczególności praca dotyczy ulepszeń zaawansowanych technik reprezentacji i kompresji cyfrowych sygnałów dźwiękowych dla zastosowań w systemach teleinformatycznych nowych generacji, w tym w systemach bezprzewodowych. Przez zaawansowane techniki kompresji cyfrowych sygnałów dźwiękowych, rozumie się najnowsze techniki rozwinięte po roku 2003. Opublikowano wówczas nową wersję standardu kompresji dźwięku szerokopasmowego ISO/IEC 14496-3 (określanego również jako MPEG-4 część 3) uwzględniającego kompresję z rozszerzaniem widma [68].

Znaczna redukcja ilości informacji stała się możliwa dzięki efektom pracy nad poznaniem sposobu funkcjonowania ludzkiego narządu słuchu [161]. W rezultacie opracowano modele psychoakustyczne pozwalające oszacować wrażliwość ludzkiego ucha na różne bodźce słuchowe. Na podstawie danych z modelu psychoakustycznego można dokonać selekcji informacji percepcyjnie istotnej oraz odrzucić składowe sygnału mniej znaczące lub poddać jest silniejszej kwantyzacji. Wspomniane badania przyczyniły się do powstania perceptualnych technik kompresji sygnałów fonicznych.

Badania w dziedzinie perceptualnego kodowania dźwięku rozpoczęły się w latach 80-tych XX wieku [23, 86–88, 110, 127, 128]. Stworzono schemat kompresji, który wykorzystywał fakt, że ludzki narząd słuchu gromadzi informacje na podstawie krótkookresowej analizy sygnału fonicznego. Wykorzystano również tak zwany *efekt maskowania*

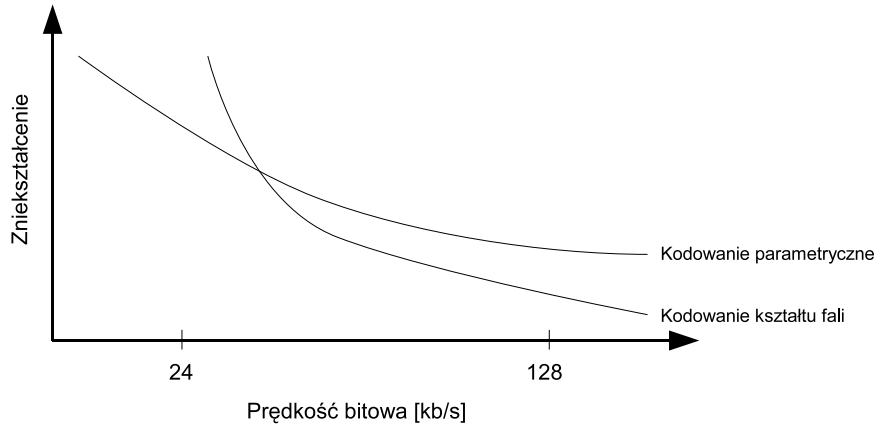


Rysunek 1.1: Ogólny schemat kodera perceptualnego.

[94, 123, 161], czyli zjawisko powodujące, że ludzkie ucho nie jest w stanie usłyszeć zniekształcenia w obecności innego silnego sygnału w bliskim sąsiedztwie czasowo-częstotliwościowym. Uwzględniając efekt maskowania można obliczyć *próg maskowania*, poniżej którego poszczególne składowe sygnału nie są słyszalne dla większości odbiorców. Zjawisko efektu maskowania wykorzystano do skonstruowania *kodera perceptualnego*, w którym szum kwantyzacji, wprowadzany na etapie kompresji, znajduje się poniżej progu maskowania.

Kodery perceptualne zaliczane są do grupy koderów kształtu fali (ang. *waveform coders*) [22, 54, 83, 84, 86], ponieważ starają się możliwie wiernie odtworzyć przebieg sygnału w dziedzinie czasu. Podgrupę technik kodowania kształtu fali stanowią metody kompresji transformatowych, w których sygnał przekształcany jest do dziedziny częstotliwości poprzez zespół filtrów lub transformację i poddawany dalszym operacjom w krótkich blokach danych. Możemy tutaj wymienić wiele komercyjnych technik kodowania wykorzystujących to podejście, takich jak Sony ATRAC (Adaptive Transform Acoustic Coding), Lucent PAC [89] lub Dolby AC-3 [40, 46]. W grupie technik wykorzystujących metody kompresji transformatowych znajdują się również ustandaryzowane schematy kodowania, jak MPEG-1 Layer 3 [64] znane również jako MP3 lub MPEG-2/4 AAC (Advanced Audio Coding) [65, 67]. Wszystkie przedstawione powyżej metody bazują na podobnym schemacie kodowania, który został przedstawiony na rysunku 1.1.

Alternatywą dla kodeków kształtu fali są metody parametrycznej reprezentacji sygnału [36, 130]. Techniki te znajdują swoje zastosowanie w kontekście bardzo małych prędkości bitowych, gdzie kodowanie transformatowe wprowadza nieakceptowalne zniekształcenia



Rysunek 1.2: Porównanie subiektywnej oceny zniekształcenia wprowadzanego przez kodery kształtu fali oraz parametryczne w funkcji prędkości bitowej (na podstawie pracy [119]).

sygnału. Koder parametryczny reprezentuje sygnał za pośrednictwem modelu, którego parametry są przesyłane do dekodera. Techniki kodowania parametrycznego pozwalają osiągnąć wyższą efektywność kompresji w stosunku do koderów kształtu fali dla małych prędkości bitowych poniżej 30 kb/s. Z drugiej strony zwiększając prędkość bitową nie można uzyskać zdekodowanego sygnału, który będzie percepcyjnie nieodróżnialny od oryginału. Dla odpowiednio dużych prędkości bitowych subiektywna ocena jakości sygnału uzyskanego na wyjściu dekodera parametrycznego, jest niższa niż w przypadku wykorzystania kodeka kształtu fali [35, 43, 119]. Ilustrację tego zjawiska przedstawia rysunek 1.2. Pokazuje on, że w stosunku do kodowania kształtu fali kodowanie parametryczne wprowadza mniejsze zniekształcenie dla małych prędkości bitowych. Jednak dla dużych prędkości bitowych, powyżej 128 kb/s, mamy do czynienia z sytuacją odwrotną. Ta obserwacja była podstawą do stworzenia hybrydowego kodera, łączącego techniki kodowania parametrycznego oraz kształtu fali (np. kodeka HILN [35]).

Mimo znaczącego postępu w dziedzinie kodowania sygnałów fonicznych wykorzystującego kodery kształtu fali oraz kodery parametryczne, w dalszym ciągu poszukuje się rozwiązań mogących zapewnić wysoką postrzeganą jakość zdekodowanego dźwięku szerokopasmowego oraz mowy. Ostatnie osiągnięcia w perceptualnym kodowaniu szerokopasmowych sygnałów fonicznych wykorzystują schemat kompresji z rozszerzaniem widma, przedstawiony na rysunku 1.3. *Koder i dekoder podstawowy* przetwarzają sygnał ograniczony pasmowo, natomiast brakujące składowe dużych częstotliwości generowane są za

pomocą techniki rozszerzania widma [68] BWE (ang. *bandwidth extension*) Tego typu podejście można zrealizować poprzez skalowanie częstotliwości [120], nieliniową filtrację [103] lub replikację widma [68].

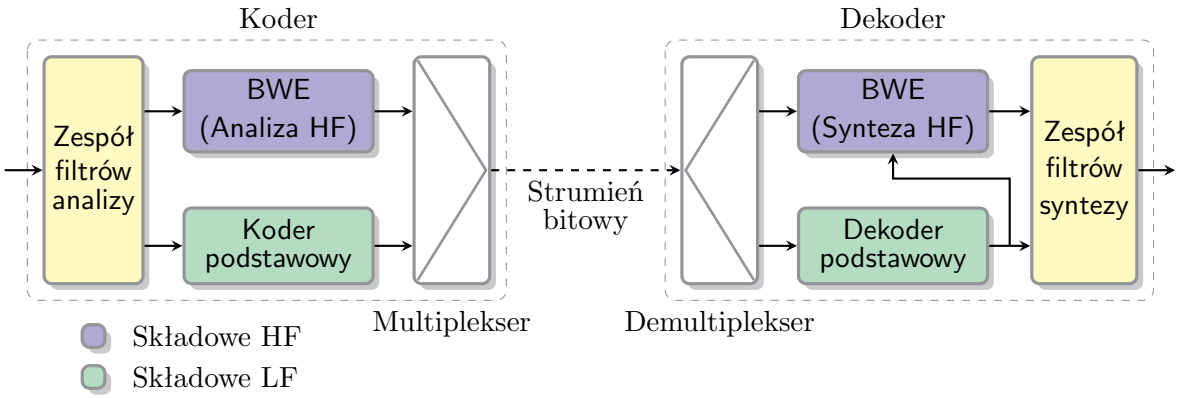
W rozprawie wykorzystuje się termin „rozszerzanie widma". Termin rozszerzanie (lub poszerzanie) widma użyto dotychczas w trzech publikacjach [8, 12, 164]. Stosowanie tego terminu, wykorzystanego także w prezentacjach konferencyjnych, nie prowadziło dotychczas do kontrowersji.

Terminem rozszerzanie widma w pracy zostaną opisane metody, które wykorzystują operację uzupełniania składowych widma, przez co pasmo sygnału wyjściowego jest szersze od pasma sygnału wejściowego.

Należy podkreślić, że w technikach BWE nie koduje się wprost brakujących składowych widma sygnału. W dekoderze proces rekonstrukcji bazuje na sygnale ograniczonym pasmowo oraz niewielkiej liczbie parametrów opisujących brakujące składowe. W ten sposób techniki BWE pozwalają na znaczną redukcję informacji w stosunku do metod kompresji transformatowej.

Schemat kompresji z rozszerzaniem widma wykorzystywany jest zarówno w zakresie kodowania szerokopasmowych sygnałów fonicznych oraz sygnałów mowy [82]. W rozprawie skupiono się na technikach BWE dotyczących szerokopasmowych sygnałów fonicznych.

Terminem małe częstotliwości LF (ang. *low frequency*) nazywać będziemy składowe sygnału w zakresie od 0 do f_{BWE}. Składowe LF kodowane są z wykorzystaniem kodera podstawowego (ang. *core encoder*). Terminem duże częstotliwości HF (ang. *high frequency*) nazywać będziemy składowe sygnału powyżej f_{BWE}. Przez f_{BWE} oznaczona będzie granica podziału pomiędzy pasmem małych i dużych częstotliwości. Ta częstotliwość stanowi jednocześnie granicę pomiędzy zakresem działania kodera podstawowego oraz kodera BWE. Dla obecnie stosowanych technik BWE [68] częstotliwość graniczne $f_{BWE} \approx 4$ kHz, czyli jest większa od częstotliwości podstawowych wydawanych przez instrumenty klasyczne muzyczne. Największa częstotliwość podstawowa dla instrumentów muzycznych wynosi 4186 Hz, co odpowiada dźwiękowi C5 na skali muzycznej. W muzyce praktycznie nie wykorzystuje się dźwięków o większych częstotliwościach. W zakresie częstotliwości powyżej 4 kHz spotykamy się tylko z dźwiękami reprezentującymi wyższe harmoniczne sygnału [34].



Rysunek 1.3: Schemat kompresji z rozszerzaniem widma. W standardzie MPEG-4 AAC HE składowe LF kodowane są za pomocą techniki AAC (koder podstawowy), natomiast składowe HF - z wykorzystaniem techniki SBR.

Przykładem techniki uznawanej aktualnie za najlepsze rozwiązanie w dziedzinie rozszerzania widma jest Spectral Band Replication (SBR) [68]. SBR jest częścią międzynarodowego standardu ISO/IEC MPEG-4 część 3 [67] dotyczącego kodowania sygnałów fonicznych, zaproponowanego przez grupę ekspertów MPEG (ang. *Moving Picture Experts Group*) [117]. SBR w połączeniu z techniką kodowania transformatowego Advanced Audio Coding (AAC) tworzy osobną metodę kompresji MPEG-4 AAC High Efficiency. W rozprawie będzie wykorzystywany również skrót MPEG-4 AAC HE. Opisywana technika znalazła zastosowania w cyfrowym radiu, np. Digital Radio Mondiale (DRM+) [41], HD Radio [60] oraz XM Satellite Radio [157].

Ogólny schemat kompresji z rozszerzaniem widma zaprezentowano na rysunku 1.3, gdzie *koder podstawowy* odpowiada za kompresję składowych LF. W standardzie MPEG-4 AAC HE jest to technika kodowania transformatowego AAC. Analiza i synteza składowych HF realizowana jest przez technikę SBR, w której rozszerzanie widma zrealizowane zostało poprzez replikację widma na płaszczyźnie czasowo-częstotliwościowej.

Zastosowanie schematu kompresji z rozszerzaniem widma pozwala w znaczący sposób zredukować ilość informacji niezbędnych do reprezentacji składowych HF. Prędkość bitowa strumienia generowanego przez technikę SBR wynosi 1 – 3 kb/s, podczas gdy koder podstawowy wymaga kilkudziesięciu kb/s do zakodowania sygnału o tym samym paśmie częstotliwości z tą samą jakością. Potwierdzają to subiektywne testy odsłuchowe przeprowadzone w 2003 roku [1] przez grupę ekspertów z organizacji EBU (ang. *European Broadcasting Union*) [1]. Testy obejmowały sygnały zakodowane z prędkościami bitowymi rzędu 16 – 64 kb/s. Dla dwukrotnie mniejszej prędkości bitowej w stosunku do kodeków

podstawowych, kodeki wykorzystujące technikę SBR uzyskały tą samą ocenę subiektywną. W obydwu przypadkach kodowaniu podlegało pełne pasmo sygnału.

Po opublikowaniu standardu kompresji dźwięku szerokopasmowego MPEG-4 AAC HE (zawierającego SBR) w roku 2003 [67, 68], powstało wiele artykułów opisujących metody rozszerzania widma [44, 59, 82, 103, 108, 135, 153]. Jednak przeważnie są to techniki oferujące niską jakość dźwięku. Obecnie brakuje technik rozszerzania widma, które oferują wysoką (porównywalną z oryginałem) jakość zdekodowanego dźwięku szerokopasmowego. Dlatego opracowanie efektywnej techniki lub wskazanie dalszych kierunków badań w tym zakresie będzie miało duży wkład w rozwój nauki.

W 2007 roku komitet MPEG zdecydował o sformułowaniu wymagań [72] na nowy uniwersalny kodek mowy i muzyki z przeznaczeniem do dystrybucji muzyki w sieciach telefonii komórkowej, który przy prędkości bitowej 24kb/s będzie oferował lepszą jakość sygnału od najlepszych obecnie kodeków AMR-WB+ (ang. *Extended Adaptive Multi Rate – WideBand*) [16] oraz MPEG-4 AAC HE [67]. Nowy, obecnie redagowany standard nazywa się ang. *Unified Speech and Audio Coding* (USAC) [74]. Efektywna technika rozszerzania widma mogłaby w istotny sposób poprawić stopień kompresji uzyskiwany w przyszłych odmianach koderów nowego standardu kodowania szerokopasmowych sygnałów fonicznych. Dlatego ta rozprawa podejmuje tematykę rozszerzania widma dźwięku, kodowania parametrycznego oraz modelowania sinusoidalno-szumowego w zastosowaniu do kompresji dźwięku szerokopasmowego.

1.2. Cele i teza pracy

1.2.1. Problem naukowy

Algorytmy wykorzystujące techniki rozszerzania widma, takie jak Spectral Band Replication [68] bazują na spostrzeżeniu, że w dźwiękach naturalnych składowe o małych częstotliwościach są silnie skorelowane ze składowymi o dużych częstotliwościach, a widmo amplitudowe sygnału z rosnącą częstotliwością łagodnie zanika do zera. Dodatkowo zakładamy, że w widmie sygnału występują prążki harmoniczne w równych odstępach częstotliwości, przez co składowe harmoniczne HF stanowią naturalne uzupełnienie składowych LF.

Schematy kompresji z rozszerzaniem widma wykorzystując powyższe właściwości potrafią zakodować dobrej jakości sygnał składowych HF z prędkością bitową rzędu 1 – 3 kb/s. W zastosowaniu razem z technikami kodowania perceptualnego [63–65, 67], techniki BWE pozwalają koderowi podstawowemu (np. MP3, AAC) przeznaczyć większą pulę bitów na reprezentację składowych LF.

Występują jednak sygnały dla których aktualnie stosowane techniki BWE nie są w stanie poprawnie odtworzyć brakujących składowych HF, nawet przy zwiększeniu puli bitów kodeka BWE. W widmie składowych dużych częstotliwości mogą występować składniki, których nie da się odtworzyć poprzez mechanizm replikacji widma (np. źródła dźwięku o częstotliwości podstawowej powyżej częstotliwości podziału f_{BWE}). Dodatkowo istotnym problemem jest zachowanie harmonicznej struktury składników tonalnych, których częstotliwości powinny być wielokrotnościami częstotliwości tonów podstawowych. W wielu technikach [40, 68, 135, 153] pojawia się również problem związany z poprawną reprezentacją szybkich zmian częstotliwości (efekty glissando i vibrato) oraz ich odtworzeniem.

Dotychczas opisane w literaturze prace związane z kompresją dźwięku szerokopasmowego z rozszerzaniem widma dotyczą głównie techniki SBR oraz jej modyfikacji. Technika SBR jest powszechnie uznanym standardem, jak również dostępna jest jej referencyjna implementacja [2]. Z tych powodów została wybrana do badań przeprowadzonych przez autora rozprawy.

Dla techniki SBR otwartym problemem jest jednak poprawna rekonstrukcja tonalnych składowych dużych częstotliwości, a zwłaszcza odtworzenie szybkich zmian w dziedzinie czasu i częstotliwości. W rozprawie pokazano, jak rozwiązać powyższy problem poprzez zastosowanie modelowania sinusoidalno-szumowego. Odpowiednie modele dają bardzo elastyczną reprezentację sygnału fonicznego, ponieważ funkcjami bazowymi są sinusoidy o dowolnej częstotliwości w odróżnieniu od transformacji Fouriera, gdzie funkcje bazowe mają stałe częstotliwości.

Istotą prac badawczych opisanych w rozprawie było poszukiwanie odpowiedniego modelu sinusoidalno-szumowego oraz jego parametrów, które w połączeniu z aktualnie stosowanymi technikami rozszerzania widma, pozwoliłyby otrzymać sygnał dźwiękowy wysokiej jakości subiektywnej przy jednoczesnej małej prędkości bitowej strumienia, rzędu 20 – 40 kb/s.

1.2.2. Teza rozprawy

Można stworzyć efektywną metodę kompresji cyfrowych sygnałów fonicznych, która wykorzystuje w tym samym paśmie częstotliwości rozszerzanie widma oraz modelowanie sinusoidalno-szumowe. Dla prędkości bitowych mniejszych od 30 kb/s, zastosowanie nowej metody prowadzi do uzyskania sygnału fonicznego o jakości lepszej niż przy zastosowaniu techniki MPEG-4 AAC HE. W szczególności poprawie ulegnie odtwarzanie składowych tonalnych dużych częstotliwości.

1.2.3. Cel naukowy

Celem pracy jest stworzenie efektywnej metody kompresji wykorzystującej techniki rozszerzania widma oraz modelowania sinusoidalno-szumowego, które łącznie pozwolą uzyskać dobrą subiektywną jakość zrekonstruowanego sygnału przy małych prędkościach bitowych.

1.3. Metodologia badań

W rozprawie, punktem wyjścia w badaniach były aktualne techniki rozszerzania widma oraz kompresji składowych HF, z głównym naciskiem położonym na technikę SBR będącą częścią standardu MPEG-4 AAC HE. Aktualnie jest to najbardziej zaawansowany algorytm kompresji składowych HF.

Technika SBR została dokładnie zbadana pod kątem poprawnej rekonstrukcji składowych harmonicznych dużych częstotliwości. Wyniki własnych badań (punkt 2.6) potwierdzają znane z literatury trudności dotyczące poprawnej rekonstrukcji składowych harmonicznych HF.

Dla uniknięcia wymienionych trudności, w rozprawie zaproponowano dwie nowe techniki łącznego wykorzystania modelowania sinusoidalno-szumowego i techniki rozszerzania widma. Nowe rozwiązania poddane zostały badaniom eksperymentalnym. W celu ich przeprowadzenia przygotowano modele kodeków związane z nowymi technikami. Jednym z elementów pracy była implementacja modelu sinusoidalno-szumowego, a następnie zbadanie możliwości zastosowania modelowania sinusoidalnego wraz z techniką rozszerzania widma.

Podczas prac wykorzystano najnowsze implementacje techniki MPEG-4 AAC HE oraz parametrycznej reprezentacji sygnałów fonicznych:

- Implementacja [2] kodeka MPEG-4 AAC HE opracowana przez konsorcjum 3rd Generation Partnership Project (3GPP)* [3].

  Kod źródłowy pochodzący od 3GPP powstawał przy współpracy firmy Coding Technologies [27] twórcy techniki SBR [68] i zawiera wzorcową implementację narzędzia rozszerzania widma [2, 18].

- Implementacja kodeka MPEG-4 SSC (ang. *Sinusoidal Coding*) [69] opracowana przez firmę Philips oraz wybrana przez MPEG na wzorcową implementację techniki kodowania sinusoidalno-szumowego.

- Autorska implementacja kodera sinusoidalno-szumowego w środowisku Matlab.

Spośród dostępnych implementacji kodeka SBR, autor wybrał implementację 3GPP, którego kod źródłowy jest publicznie dostępny. Dzięki temu możliwa była analiza algorytmów kodowania oraz ich modyfikacja w celach eksperymentalnych. Kod źródłowy algorytmu SBR jest na bieżąco aktualizowany przez 3GPP i Coding Technologies. Implementacja kodera MPEG-4 AAC HE opublikowana przez konsorcjum 3GPP uznawana jest jako jedna z najlepszych w tej klasie, co potwierdzają testy odsłuchowe [21] oraz opinie [31] twórców standardu MPEG-4 AAC HE.

Wymienione wyżej implementacje posłużyły do zbudowania modeli dwóch kodeków implementujących oryginalne metody kompresji zaproponowane w rozprawie. Autor zrealizował dwa pełne kodeki, czyli dla obu metod powstały zarówno koder jak i dekoder. Dzięki temu możliwe było wszechstronne zbadanie właściwości zaproponowanych technik kompresji.

W trakcie badań skonfrontowano efektywność znanych z literatury technik rekonstrukcji składowych HF z zaproponowanymi technikami. Zbadano również możliwość rozszerzenia aktualnego standardu kodowania dźwieku szerokopasmowego MPEG-4 AAC HE

*3GPP jest wspólnym projektem kilku organizacji standaryzacyjnych mający na celu rozwój systemów telefonii komórkowej trzeciej generacji 3G. Organizacje uczestniczące w projekcie to: ARIB, CWTS, ETSI, T1, TTA, TTC.

Tabela 1.1: Sygnały testowe.

Sygnał	Opis	Czas trwania	źródło
accordion	pojedynczy instrument	9,3 s	EBU SQAM [42]
altosax	pojedynczy instrument	8,4 s	EBU SQAM [42]
glissando	pojedynczy instrument	4,6 s	Uniwersytet Iowa [151]
violin2	pojedynczy instrument	5,6 s	
discofunk	muzyka	7,2 s	MPEG [73]
philipspop	muzyka	9,5 s	MPEG [73]
strings	muzyka	10,4 s	

o techniki zaproponowane w pracy doktorskiej. W tym celu przeprowadzono szereg testów subiektywnych według zaleceń ITU-R/EBU BS.1534 [78], w których zbadano różne warianty proponowanych rozwiązań oraz ich wpływ na jakość zdekodowanego sygnału.

Krytycznym elementem przeprowadzonych badań jest wybór odpowiednich sygnałów fonicznych. W tym celu wykorzystano sekwencje testowe pochodzące z płyty EBU SQAM (ang. *Sound Quality Assessment Material*) [42], bazy danych Uniwersytetu Iowa [151] oraz sekwencje pochodzące z zestawu testowego zaproponowanego przez komitet standaryzacyjny MPEG [73]. Dodatkowo wykorzystano nagrania, które uwypuklają problemy zniekształceń wprowadzanych przez technik rozszerzania widma. Opis sygnałów testowych znajduje się w tabeli 1.1. Nagrania pochodzące ze wspomnianych źródeł zostały zarejestrowane w profesjonalnych studiach nagraniowych. Wszystkie nagrania są sygnałami monofonicznymi próbkowanymi z częstotliwością 44100 Hz. W rozprawie nie będą poruszane zagadnienia związane z kodowaniem dźwięku wielokanałowego.

1.4. Pomiar jakości sygnałów fonicznych

W pracy badano sygnały foniczne reprezentujące dźwięk szerokopasmowe w tym muzykę. W literaturze zaproponowano szereg standardowych miar jakości sygnałów fonicznych, które możemy podzielić na dwie grupy:

- miary obiektywne - do oceny różnicy między sygnałami wykorzystują modele statystyczne, które mogą bazować na modelu ludzkiego słuchu [80, 81, 156, 159]. Aktualnie standardem w dziedzinie pomiarów obiektywnych dla dźwięku szerokopasmowego jest miara PEAQ (ang. *Perceptual Evaluation of Audio Quality*) [77].

- miary subiektywne - sygnały testowe oceniane są przez określoną liczbę osób według skali jakości ustalonej w danej metodologii [76, 78, 79]. Po przeprowadzeniu testów odsłuchowych uzyskane dane poddawane są obróbce statystycznej.

Dla oceny efektywności kodeków fonicznych powszechnie stosowaną praktyką jest stosowanie miary subiektywnej.

Obecne standardy subiektywnej oceny jakości sygnałów fonicznych zdekodowanych przez dekoder jest metodologia BS.1116 [76] dotycząca małych zniekształceń sygnału. Dla średnich i dużych zniekształceń sygnału stosuje się BS.1534 [78] MUSHRA (ang. *MUltiple Stimuli with Hidden Reference and Anchor*). Metodologie testów subiektywnych zostały zaproponowane przez ITU-R (ang. *International Telecommunication Union Radiocommunication*).

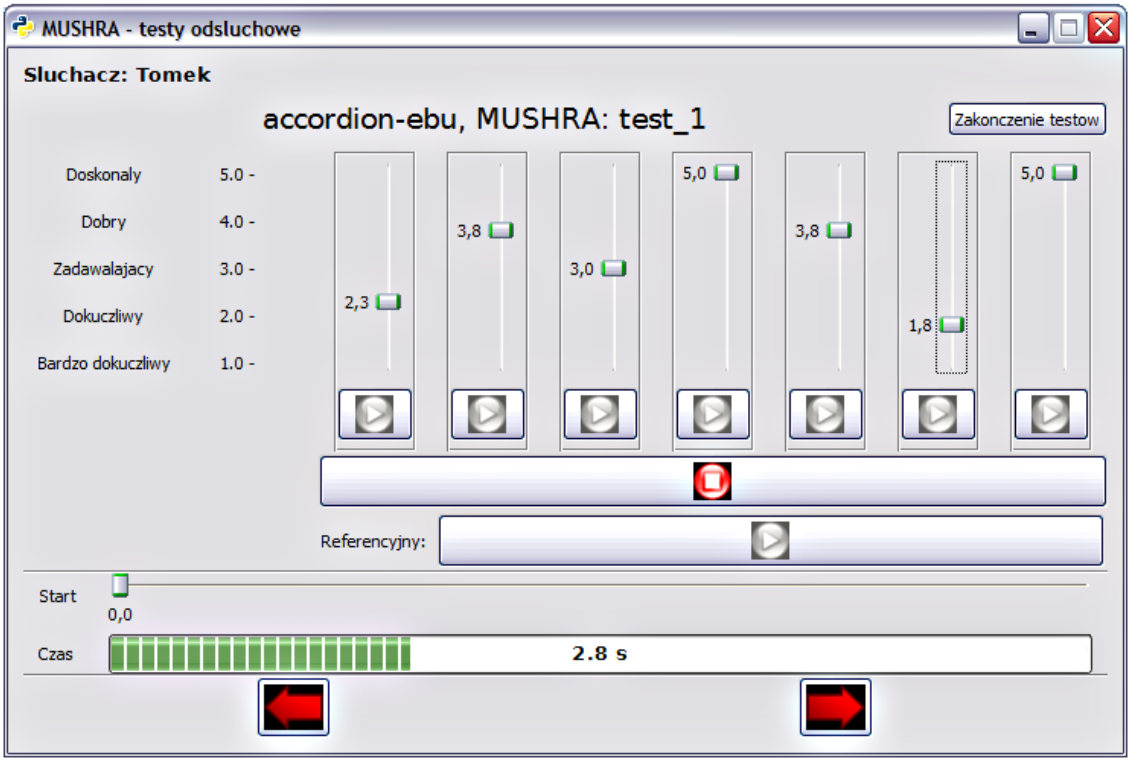
Według zaleceń [76] i [78] małe zniekształcenia dotyczą będących na granicy percepcji różnić między sygnałami w badanych systemach. Średnie i duże zniekształcenia odpowiadają wyraźnym słyszalnym różnicom pomiędzy badanymi systemami.

Techniki badane w rozprawie doktorskiej operują w zakresie małych prędkościach bitowych rzędu 16-24 kb/s, co może prowadzić do powstawania słyszalnych zniekształceń w rekonstruowanych sygnałach. Dlatego autor zdecydował się wykonać porównanie badanych systemów, stosując subiektywną miarę jakości dla średnich i dużych zniekształceń MUSHRA [78].

Metodologia MUSHRA definiuje szereg warunków związanych z wykonaniem testów subiektywnych, takich, jak akustyka pomieszczenia, w którym wykonywane są testy odsłuchowe, parametry techniczne sprzętu nagłaśniającego, dobór materiałów testowych oraz sam przebieg eksperymentu. Każdy słuchacz musi przejść wstępną fazę treningową podczas której zostaje zaznajomiony z procesem testów odsłuchowych.

W czasie trwania testów odsłuchowych ustalona liczba osób ma za zadanie porównać i ocenić jakość prezentowanych sygnałów testowych w odniesieniu do sygnału oryginalnego. Każdy uczestnik eksperymentu wyraża swoją opinię w ustalonej wcześniej skali, za pomocą odpowiedniej aplikacji kontrolującej przebieg badania. Następnie indywidualne noty są uśredniane i poddawane obróbce statystycznej, w celu oszacowania całościowych ocen jakości. Sygnały oceniane są w skali od 1 do 5 według tabeli 1.2.

Testy odsłuchowe są podzielone na sesje, w których bierze udział jedna osoba. Sesja odsłuchowa składa się z pewnej liczby zestawów testowych, których liczba zależy od ilości sygnałów testowych, badanych systemów oraz prędkości bitowych. Uczestnik ma za zadanie dokonać oceny określonej liczby sygnałów fonicznych reprezentujących badane systemy, wchodzących w skład pojedynczego zestawu testowego. Dodatkowo w zestawie testowym umieszczane są trzy ukryte sygnały odniesienia nazwane kotwicami (ang. *anchor*), czyli sygnał oryginalny oraz dwa sygnały ograniczone pasmowo do 3,5 kHz i 7 kHz. Do badań subiektywnych wykorzystano autorską aplikację, której interfejs przedstawiony jest na rysunku 1.4.



Rysunek 1.4: Interfejs graficzny aplikacji wykorzystanej podczas testów subiektywnych MUSHRA.

Tabela 1.2: Subiektywna skala ocen – MUSHRA.

Jakość	**Ocena**	**Zniekształcenie**	
Doskonała	5	Niesłyszalne	} *Małe zniekształcenia*
Dobra	4	Słyszalne, ale akceptowalne	} *Średnie zniekształcenia*
Zadowalająca	3	Mało dokuczliwe	
Słaba	2	Dokuczliwe	} *Duże zniekształcenia*
Dokuczliwa	1	Bardzo dokuczliwe	

Otrzymane wyniki pomiaru subiektywnego pozwalają na miarodajne porównanie systemów, jednak przeprowadzanie badań wymaga dużych nakładów pracy związanej z przygotowaniem i organizacją testów. Pojedyncza sesja odsłuchowa trwa około 30-40 minut. Aby wyniki były wiarygodne, standard MUSHRA wymaga uczestnictwa około 15 – 20 osób. Tak więc, czas trwania jedynie samych sesji odsłuchowych, dla określonych warunków eksperymentalnych, wynosi około 10 godzin. Biorąc pod uwagę fakt, że testy muszą odbywać się w specjalnie do tego celu przeznaczonym pomieszczeniu odsłuchowym, w którym w danej chwili może znajdować się tylko jedna osoba, ich realny czas realizacji wydłuża się do kilku dni.

Podczas przygotowania rozprawy doktorskiej przeprowadzono wiele sesji odsłuchowych, mających na celu sprawdzenie różnych parametrów badanych systemów. Dlatego łączny czas wszystkich sesji odsłuchowych wynosił kilkaset godzin.

Autor rozprawy nie zdecydował się na przeprowadzenie badań wykorzystując metodologię testów obiektywnych, ze względu na istotną wadę utrudniającą ich stosowanie do pomiaru zniekształceń sygnałów fonicznych. Testy obiektywne są relatywnie słabo skorelowane z wynikami testów subiektywnych. Wynika to ze zbyt całościowego rozpatrywania błędu – bez rozdzielenia go na składowe słyszalne i te, które można zaniedbać z perspektywy układu słuchowego [14]. Metody obiektywne są bardzo wrażliwe na nawet niewielkie modyfikacje sygnału, które są niedostrzegalne przez słuchaczy.

Miary obiektywne bazują często na porównaniu stosunku sygnału do szumu SNR, co w przypadku np. technik wykorzystujących modelowanie sinusoidalno-szumowe jest niewystarczające. Przykładowo wiele koderów parametrycznych (np. MPEG-4 SSC) nie przesyła informacji o fazie składowych, zmniejszając tym samym docelową prędkość bitową, a jednocześnie uzyskując wysoką jakość zdekodowanego sygnału, co potwierdzają słuchacze. Dla tak zakodowanego sygnału wynik uzyskany za pomocą miary obiektywnej byłby niski. Skuteczność miar obiektywnych jest również kwestionowana w kontekście pomiaru jakości systemów, w których występują średnie lub duże zniekształcenia [147].

1.5. Przegląd pracy

Rozprawa składa się z sześciu rozdziałów, które łącznie wchodzą w skład dwóch podstawowych części pracy. Pierwsza część dotyczy studiów literaturowych na temat zaawansowanych technik kompresji sygnałów fonicznych, ze szczególnym uwzględnieniem metod rozszerzania widma oraz modelowania sinusoidalno-szumowego. W tej części pracy zawarte są również wyniki badań dotyczące efektywności kompresji techniki MPEG-4 AAC HE, uzyskanych przez autora rozprawy. Druga część pracy zawiera opis nowych, autorskich algorytmów kompresji sygnałów fonicznych oraz wyniki przeprowadzonych eksperymentów.

Rozdział 2 zawiera:

- Przegląd wybranych zagadnień z dziedziny modelowania sinusoidalno-szumowego.
- Klasyfikację technik rozszerzania widma wraz ze szczegółowym opisem metody SBR wchodzącej w skład standardu MPEG-4 AAC HE.
- Określenie głównych problemów pojawiających się przy implementacji technik BWE, tj. poprawnej rekonstrukcji składowych tonalnych.

Rozdział 3 przedstawia:

- Oryginalną, zaproponowaną przez autora, metodę rozszerzania widma bazującą na modelowaniu sinusoidalno-szumowym, która może stanowić uzupełnienie aktualnych technik BWE, zwiększając ich efektywność kompresji.
- Sposób zaadoptowania nowej techniki do aktualnego standardu kodowania MPEG-4 AAC HE.
- Spektrogramy przykładowych zdekodowanych sygnałów.

Rozdział 4 opisuje:

- Zaproponowane przez autora rozprawy nowatorskie podejście do rekonstrukcji składowych HF, w którym składowe tonalne reprezentowane są przez sygnały wąskopasmowe.
- Proces kodowania oraz rekonstrukcji składowych HF na podstawie sygnału prototypowego uzyskiwanego w dekoderze.

Rozdział 5 koncentruje się na:

- Opisie i analizie wyników otrzymanych podczas testów odsłuchowych.
- Porównaniu technik zaproponowanych w rozprawie ze standardową techniką MPEG-4 AAC HE.

Rozdział 6 podsumowuje pracę i zawiera:

- Najważniejsze wnioski.
- Oryginalne osiągnięcia.
- Wskazanie kierunku dalszych badań w tej dziedzinie.

Rozdział 2

Przegląd rozwiązań dotyczących technik rozszerzania widma i modelowania sygnałów

2.1. Wprowadzenie

Zastosowanie technik rozszerzania widma BWE (ang. *Bandwidth Extension*) pozwala na efektywną rekonstrukcję brakujących składowych HF. Tego typu techniki bazują na informacji o podobieństwie składowych małych i dużych częstotliwości. Poprzez zastosowanie metod BWE możliwe jest osiągnięcie wysokiego stopnia kompresji danych odpowiadających składowym dużych częstotliwości.

W rozdziale tym algorytmy rozszerzania widma zostaną podzielone na następujące grupy [103]:

- Ślepe techniki - sygnał składowych HF rekonstruowany jest na podstawie ograniczonego pasmowo sygnału wejściowego otrzymanego w dekoderze. Dekoder nie posiada dodatkowych informacji mogących ułatwić rekonstrukcję brakujących składowych HF. Techniki tego typu znajdują głównie zastosowanie w poprawieniu subiektywnej jakości sygnałów ograniczonych pasmowo. Mogą również zostać wykorzystane w dekoderach sygnałów fonicznych w celu rozszerzenia pasma sygnału.

- Algorytmy posiadające *a priori* wiedzę o strukturze sygnału. Podejście to znajduje zastosowanie w kompresji dźwięku ze względu na stosunkowo dobrą jakość zdekodowanego sygnału.

Dodatkowo zostaną omówione podstawy modelowania sinusoidalno-szumowego sygnałów fonicznych. Techniki te pozwalają opisać sygnał za pomocą grupy parametrów, których przesłanie wymaga jedynie małej prędkości bitowej. Elementy modelowania sinusoidalno-szumowego znalazły również zastosowanie w proponowanych w rozprawie algorytmach BWE.

2.2. Przegląd wybranych narzędzi z zakresu modelowania dźwięku

Modelowanie sygnałów fonicznych odgrywa obecnie istotną rolę w dziedzinach związanych z kompresją, rozpoznawaniem i analizowaniem sygnałów mowy, usuwaniem zakłóceń oraz modyfikacją sygnałów. Narzędzia z tego zakresu wykorzystują parametryczne modele sygnału, które bazują na właściwościach ludzkiego słuchu oraz aktualnej wiedzy na temat struktury sygnałów muzycznych oraz mowy. W tego typu modelach sygnał reprezentowany jest przez grupę parametrów. W zależności od modelu parametry opisują istotne w danym kontekście właściwości sygnału (np. głośność, barwę) i pozwalają na syntezę sygnału równoważnego z punktu widzenia wybranych cech.

Parametryczne modele sygnałów fonicznych możemy podzielić na dwie grupy:

- **wysokopoziomowe** - w których parametry bazują na semantyce zapisu muzycznego, tj. nuty, czas trwania dźwięku, tempo, itd. Kodery foniczne zbudowane w oparciu o tego typu modele generują strumień binarny o bardzo małej prędkości bitowej rzędu 1 – 10 kb/s. Przykładem tego typu technik może być standard MPEG-4 część 3 opisujący dźwięk syntetyczny (ang. *MPEG-4 Structured Audio*) [66, 137]. Grupa MPEG zaproponowała również standard MPEG-7 [70] służący między innymi do klasyfikacji wysokopoziomowych cech sygnałów fonicznych.

- **niskopoziomowe** - w których parametry opisują np. widmo sygnału, obwiednię, parametry poszczególnych prążków, itd. Modele tego typu rozpatrują źródła sygnałów

fonicznych jako całość, wykorzystując tylko krótko-czasowe zależności. Wykorzystując tego typu model można precyzyjnie kontrolować wysokość oraz czas trwania poszczególnych dźwięków. Przykładem niskopoziomowej reprezentacji sygnału fonicznego jest modelowanie sinusoidalno-szumowe, w którym sygnał reprezentowany jest przez składowe sinusoidalne, szum oraz krótkotrwałe transjenty.

W rozprawie wykorzystane zostaną elementy niskopoziomowej reprezentacji sygnału, bazujące na modelowaniu sinusoidalno-szumowym, którego podstawy zostaną zaprezentowane w kolejnym punkcie.

2.2.1. Modelowanie sinusoidalno-szumowe

Modelowanie sinusoidalno-szumowe opiera się na założeniu, że każdy sygnał można przedstawić jako sumę składowej deterministycznej (sinusoidy) $d(t)$ oraz stochastycznej (szum) $n(t)$:

$$s(t) = d(t) + n(t). \tag{2.1}$$

McAulay i Quatieri (MQ) zaproponowali model sinusoidalny sygnału [113], który reprezentuje deterministyczną część sygnału $d(t)$ jako sumę składowych sinusoidalnych, zgodnie ze wzorem (2.2):

$$d(t) = \sum_{k=1}^{K} A_k(t) \cos\left(\varphi_k(t)\right), \tag{2.2}$$

$$\varphi_k(t) = \varphi_k(0) + 2\pi \int_0^t f_k(\tau) d\tau, \tag{2.3}$$

gdzie:

k – numer sinusoidy,

K – liczba sinusoid,

$A_k(t)$ – amplituda chwilowa,

$\varphi_k(t)$ – faza chwilowa,

$\varphi_k(0)$ – faza początkowa,

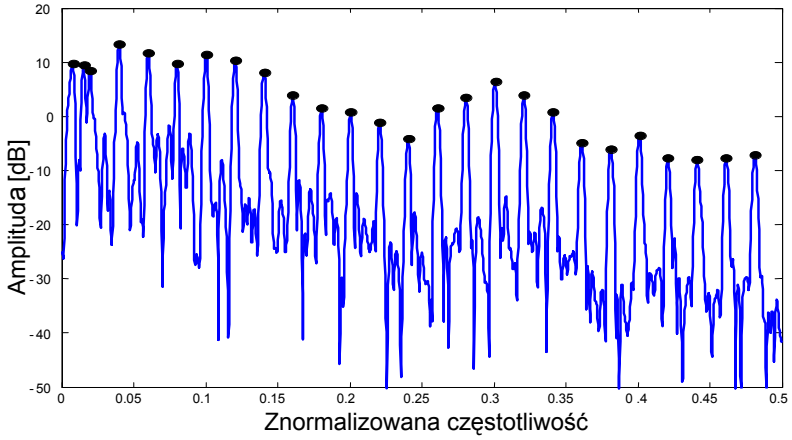
$f_k(t)$ – częstotliwość chwilowa.

Poszczególne składniki sumy we wzorze (2.2) określane są jako *składowe sinusoidalne* (ang. *sinusoidal partials*), ponieważ lokalnie można je przybliżyć funkcją sinusoidalną o wolno zmiennych w czasie parametrach amplitudy i częstotliwości. W rzeczywistości są to funkcje sinusoidalne z modulowaną amplitudą i częstotliwością, których energia może

być rozłożona w szerokim zakresie częstotliwości. Efekt ten można zaobserwować zwłaszcza dla niestacjonarnych fragmentów sygnału.

Parametry chwilowe amplitudy $A_k(t)$ i częstotliwości $f_k(t)$ są wyznaczane za pomocą analizy czasowo-częstotliwościowej, na podstawie krótkookresowej transformacji Fouriera STFT [149]. Rekonstrukcja sygnału polega na wyznaczeniu wartości pośrednich amplitud i częstotliwości pomiędzy zmierzonymi z odstępem t parametrami modelu. W tym celu wykorzystuje się proste funkcje interpolujące, np. wielomian niskiego rzędu.

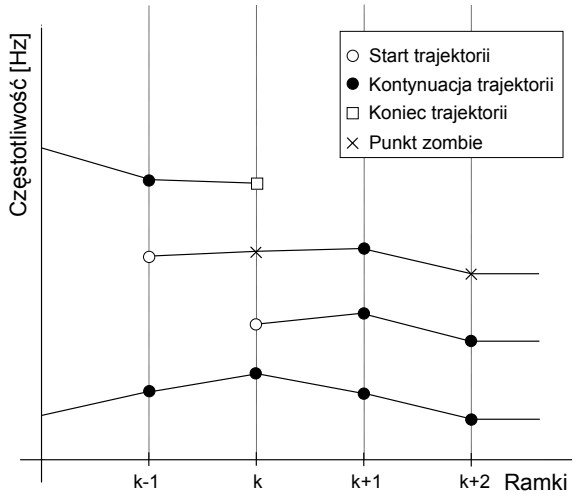
Detekcja parametrów odbywa się dla lokalnych szczytów widma amplitudowego w danej ramce analizy. Proces ten określać będziemy *detekcją sinusoid* lub *analizą sinusoidalną*. Przykładowy wynik detekcji sinusoid pokazany jest na rysunku 2.1. Dobry przegląd technik związanych z detekcją parametrów sinusoid znajduje się w pracy [92].



Rysunek 2.1: Wynik analizy sinusoidalnej dla widma amplitudowego pojedynczej ramki sygnału.

Wyznaczone podczas analizy sinusoidalnej sekwencje parametrów chwilowych amplitudy i częstotliwości łączone są w trajektorie sinusoidalne, natomiast opisujący to proces nazywa się *śledzeniem trajektorii sinusoidalnych* i został pokazany na rysunku 2.2. Jeśli sinusoida nie może zostać połączona z sinusoidą w poprzedniej ramce, wówczas mówimy o starcie lub narodzinach trajektorii. Jeśli można połączyć sinusoidy w kolejnych ramkach wówczas mamy do czynienia z kontynuacją trajektorii. Jeśli trajektoria sinusoidalna nie może zostać połączona z sinusoidą w kolejnej ramce, wówczas następuje koniec trajektorii (śmierć trajektorii). Dodatkowo trajektoria może zostać uzupełniona przez pozorną (nieistniejącą) sinusoidę nazwaną *zombie*. Termin ten został wprowadzony przez McAulaya i Quatieriego [113]. Punkt zombie pozwala na kontynuację trajektorii, co umożliwia

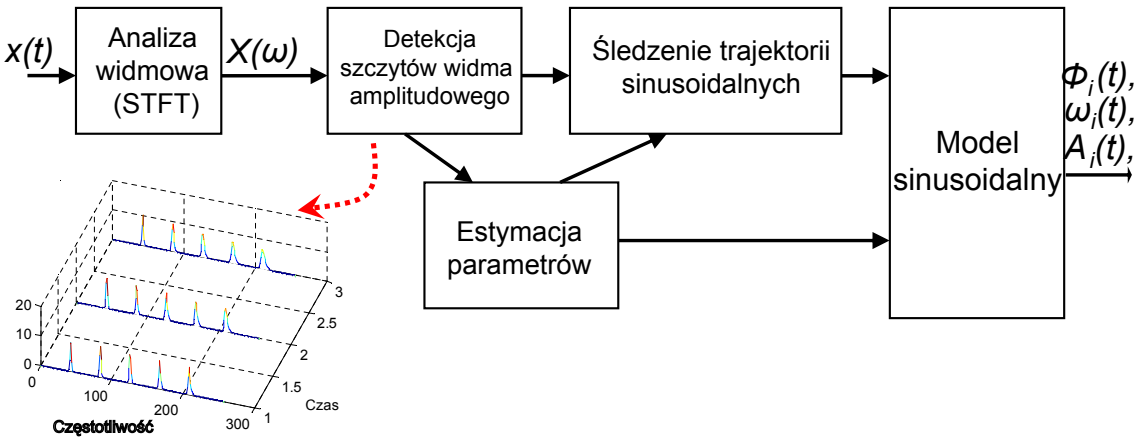
zmniejszenie liczby trajektorii, a przez to wyższą efektywność kompresji kodera sinusoidalnego.



Rysunek 2.2: Śledzenie trajektorii sinusoidalnych.

Decyzja, na podstawie której łączone są sinusoidy zależy od tzw. kryterium kontynuacji zdefiniowanego dla odpowiedniego algorytmu śledzącego. Obecnie dostępnych jest wiele metod rozwiązania tego problemu, np. stosujących predykcję [98, 99] lub ukryte modele Markowa [33].

Schemat modelowania sinusoidalnego pokazany jest na rysunku 2.3. Sygnał wejściowy poddawany jest krótkookresowej transformacji Fouriera STFT. Następnie w dziedzinie DFT wyznaczane są parametry sinusoid i łączone w trajektorie sinusoidalne.

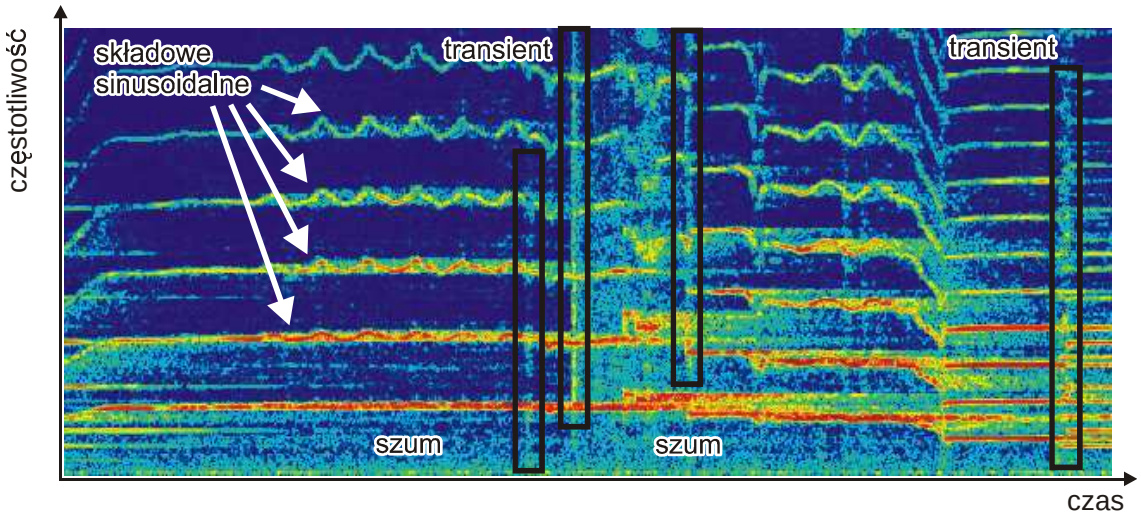


Rysunek 2.3: Schemat modelowania sinusoidalnego

Możliwe jest także reprezentowanie sygnałów szumowych za pomocą sinusoid, jednak wymaga to generowania znaczącej liczby parametrów. Sensownym wydaje się być zastosowanie odrębnego modelu parametrycznego dla składowych szumowych. Serra i Smith zaproponowali modelowanie sinusoidalne [139, 141] dla części deterministycznej oraz modelowanie szumowe dla części stochastycznej. Widmo składowej stochastycznej wyznacza się jako różnica pomiędzy widmem sygnału oryginalnym a widmem części deterministycznej [15, 19]. Część deterministyczna sygnału reprezentowana jest za pomocą modelowania szumowego (rozdział 2.2.2).

Podział na składową deterministyczną i stochastyczną wykorzystali również Fitz i Hacken [48, 49]. W ich rozwiązaniu składowym sinusoidalnym towarzyszy wąskopasmowy szum generowany przez zespół oscylatorów.

Obecnie w parametrycznym kodowaniu sygnałów fonicznych stosuje się modele hybrydowe [37, 69, 132, 152], w których dokonuje się dekompozycji sygnału na składniki szumowe, szerokopasmowe transjenty oraz składniki sinusoidalne (rysunek 2.4).



Rysunek 2.4: Model hybrydowy sygnału. Spektrogram pokazujący detekcję poszczególnych elementów modelu.

Składowe tonalne

W pracy składowe sinusoidalne określane będą również jako *składowe tonalne*. Należy jednak zwrócić uwagę, że pojęcie *tonalności* określa cechę sygnału, która pozwala rozróżnić składowe tonalne od szumu [161]. Tonalność przyjmuje wartości z zakresu $< 0,1 >$, w którym 0 oznacza, że dana składowa lub sygnał mają charakter całkowicie szumowy. Natomiast tonalność na poziomie 1 oznacza, że składowa sygnału może być przybliżona

jako pojedyncza sinusoida. Efekt tonalności związany jest z modulacją parametrów sinusoidy, przez co zmienia się pasmo składowej. Nie ma więc ścisłej granicy określającej czy dana składowa sinusoidalna jest tonalna. Wprowadzono jednak szereg miar [52, 63, 125], które w połączeniu ze zdefiniowanym progiem mogą służyć jako klasyfikatory tonalności. Wartość przy dla które klasyfikator tonalności podejmuje decyzję o uznaniu składowej za szumową lub tonalną w pracy nazywać będziemy *progiem tonalności*.

2.2.2. Modelowanie obwiedni widma

Przez *obwiednię widma* rozumieć będziemy funkcję niskiego rzędu przechodzącą przez maksima widma amplitudowego.

Sygnał będący różnicą sygnału oryginalnego i części deterministycznej nazywany jest residuum szumowym. W obecnie stosowanych modelach sinusoidalno-szumowych residuum szumowe modeluje się za pomocą obwiedni widma amplitudowego.

Motywacją zastosowania modelowania szumowego jest spostrzeżenie, że ludzki narząd słuchu jest wrażliwy na chwilowe zmiany energii w pasmach krytycznych[28, 161]. Natomiast jest mniej wrażliwy na rozkład energii wewnątrz poszczególnych pasm krytycznych. Na tej zasadzie działa model szumowy zastosowany przez Goodwina [51]. W koderze mierzona jest energia w pasmach krytycznych, natomiast w dekoderze generowany jest szum pasmowy o energii odpowiadającej sygnałowi oryginalnemu.

Inne rozwiązanie dla modelowania szumowego zastosował Serra [139, 141]. W swojej pracy proponuje wykorzystanie modelu predykcji liniowej LP [71] (ang. *Linear Predictive*) do modelowania obwiedni widma. Po stronie analizy stosowany jest filtr IIR generujący rezonanse przy pewnych częstotliwościach. LP znajduje współczynniki tego filtru poprzez predykcję p poprzednich próbek sygnału [71]. Proces modelowania LP wykonywany jest w krótkich oknach analizy w celu oddania charakteru zmienności widma w czasie. Po stronie syntezy generowany jest szerokopasmowy szum, którego widmo jest ukształtowane na podstawie odtworzonej obwiedni widma modelu LP. Podobny efekt można osiągną stosując model ARMA (ang. *Auto-Regressive Moving Average*) [32].

2.3. Psychoakustyczne aspekty percepcji składowych dużych częstotliwości

W tym rozdziale zostanie opisana percepcja składowych HF. W szczególności omówione zostaną aspekty związane z wysokością, głośnością oraz barwą dźwięku. Informacje te są szczególnie istotne z punktu widzenie kodera, który musi zdecydować jakie elementy sygnału HF są istotne, a jakie można odrzucić.

2.3.1. Wysokość dźwięku oraz struktura harmoniczna

Wysokość dźwięku zależna jest od częstotliwości drgań. Ludzki narząd słuchu jest w stanie odebrać dźwięki z zakresu od około 16 - 20 Hz do około 16 - 20 kHz. Zakres wykorzystywany przez typowe dźwięki muzyczne jest zwykle mniejszy. Współczesna muzyka posługuje się jednak niekiedy dźwiękami o częstotliwościach spoza tego zakresu, zwykle generowanymi przez elektroniczne instrumenty muzyczne.

Wysokość wielotonu złożonego z wielu składowych harmonicznych jest w różny sposób postrzegana przez ludzki układ słyszenia. Wrażenie wysokości tonów zmienia się przy dużych natężeniach. Dla małych częstotliwości wysokość obniża się wraz ze wzrostem natężenia, zaś dla dużych częstotliwości wzrasta.

Chcąc oddać to samo wrażenie wysokości dźwięku techniki BWE powinny poprawnie zrekonstruować strukturę harmoniczną sygnału. Jeśli znamy częstotliwość podstawową f_0 danego szeregu, wówczas pozostałe harmoniczne można wygenerować jako wielokrotności f_0. Jednak dla niektórych źródeł dźwięku częstotliwość k-tej harmonicznej ma wartość $f_k = kf_0\sqrt{1+\beta k^2}$, gdzie β jest współczynnikiem nieharmoniczności [11, 13, 93, 158] w granicach od 0 do 1. Generując szeregi harmoniczne w technikach BWE należy mieć te aspekty na uwadze. Zaburzenie struktury harmonicznej może powodować, że postrzegana wysokość dźwięku będzie różnić się od oryginalnej. W skrajnych przypadkach może to prowadzić do powstawania przykrych zniekształceń.

Wysokie częstotliwości mają również duże znaczenie dla lokalizacji źródeł dźwięku w przestrzeni [24, 30]. Ograniczenie pasmowe sygnału do 7 kHz powoduje, że słuchacze muszą wykonać więcej ruchów głową aby poprawnie zlokalizować źródło dźwięku.

2.3.2. Głośność

Głośność harmonicznych jest ściśle związana z ich amplitudami. Należy jednak zauważyć, że ludzkie ucho grupuje składowe sygnału, więc głośność danego szeregu harmonicznego postrzegamy jako całość, bez wyodrębniania poszczególnych składowych szeregu.

Standardowy model głośności dźwięku, taki jak ISO532 [62], lub zaproponowany przez Glasberg'a i Moora'a [50] model głośności dla dźwięków zmiennych w czasie, biorą pod uwagę tylko amplitudę składowych. Oznacza to, że jeśli technika BWE ma odtworzyć tą samą głośności dla składowych HF, to musi zrekonstruować dokładnie takie samo widmo amplitudowe jak w sygnale oryginalnym.

Biorąc pod uwagę globalny próg słyszenia [123, 161] możemy zauważyć, że składowe o dużych częstotliwościach podlegają silnemu efektowi maskowania. Wobec tego możemy założyć, że idealne odwzorowanie obwiedni widma dla składowych HF nie jest krytyczne.

2.3.3. Barwa dźwięku

Barwa dźwięku jest subiektywną cechą dźwięku, która pozwala odróżnić brzmienia różnych instrumentów lub głosu [122]. Z punktu widzenia techniki BWE, barwa dźwięku decyduje o kształcie obwiedni widma, a więc jednego z podstawowych parametrów. O kształcie obwiedni widma decyduje z kolei struktura harmoniczna oraz widmo amplitudowe i fazowe sygnału. Z tego powodu zrekonstruowane w technice BWE składowe HF powinny mieć precyzyjnie określone amplitudy oraz częstotliwości.

Na podstawie prac Plomp'a i Steeneken'a [126] można przyjąć, że wpływ parametru fazy na barwę dźwięku dużych częstotliwości jest niewielki w porównaniu do wpływu amplitudy. Jest to powodem, dla którego w technikach BWE nie przesyła się informacji o widmie fazowym sygnału. Po stronie dekodera faza może być wygenerowana w sposób syntetyczny, np. generatorem losowej fazy.

2.3.4. Podsumowanie

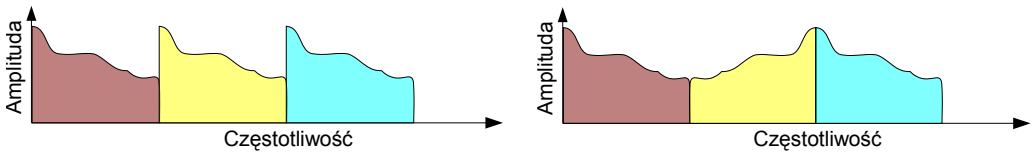
Poprawna rekonstrukcja sygnału z wykorzystaniem technik BWE wymaga właściwego odwzorowania obwiedni widma składowych HF. Pozwoli to uzyskać prawidłowe wrażenie głośności oraz barwy dźwięku składowych HF. Istotne jest również zachowanie poprawnej struktury harmonicznej sygnału odpowiadającej za prawidłowe postrzeganie wysokości dźwięku.

2.4. Klasyfikacja metod rozszerzania widma

W rozdziale tym przedstawiona zostanie idea rozszerzania widma sygnałów fonicznych z uwzględnieniem aktualnych prac badawczych koncentrujących się na tej dziedzinie. Zaprezentowane zostaną parametryczne metody rozszerzania widma bazujące na replikacji oraz modelowaniu sinusoidalno-szumowym. Przedstawiona zostanie również idea nieparametrycznego rozszerzania widma, czyli tzw. ślepe techniki.

2.4.1. Metody oparte na replikacji widma

Pierwszym etapem dekodowania sygnału jest kontrolowane przesunięcie wybranych części widma składowych LF do pasma składowych HF. W efekcie uzupełniane są brakujące fragmenty widma. Proces ten realizowany jest poprzez replikację (kopiowanie) fragmentów widma, tak jak w technice SBR [68]. Inny sposób generacji składowych HF polega na zastosowaniu układu nieliniowego [104] lub nadpróbkowania [16]. Rysunek 2.5 przedstawia porównanie technik generowania składowych HF dla metod BWE.



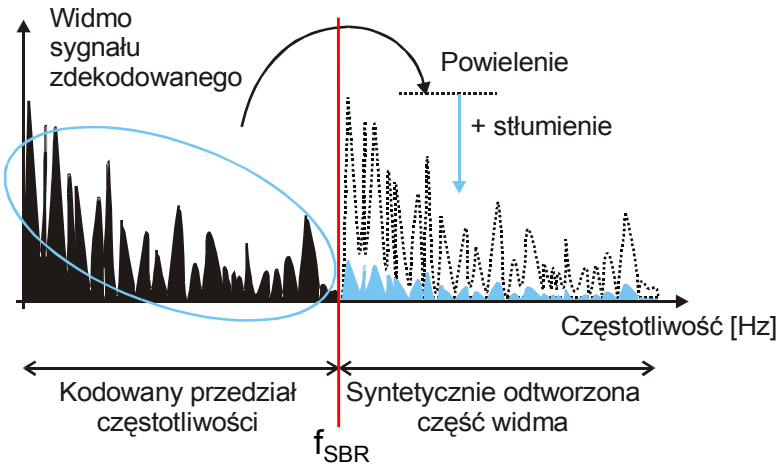
(a) Kopiowanie lub modulacja (b) Nadpróbkowanie

Rysunek 2.5: Widmo amplitudowe powstałe po zastosowaniu dwóch różnych metod rozszerzania widma.

Kolejne etapy rekonstrukcji składowych HF w parametrycznych technikach BWE przedstawiono na przykładzie techniki SBR (rysunek 2.6). Granica pomiędzy małymi częstotliwościami oraz rekonstruowanym pasmem składowych HF wyznacza częstotliwość podziału f_{SBR}. Częstotliwość podziału zależy w głównej mierze od techniki kompresji oraz docelowej prędkości bitowej. Powielone składowe HF poddane są skalowaniu, którego celem jest odtworzenie poprawnej energii poszczególnych składowych sygnału. Należy podkreślić, że kodowaniu nie podlegają poszczególne składowe widma, tak jak ma to miejsce w technikach transformatowych (np. w AAC). Do dekodera przesyłane są współczynniki skalujące energię obwiedni widma.

Jeśli widmo składowych LF i HF jest silnie ze sobą skorelowane, to przedstawiony powyżej proces syntetycznej generacji widma okazuje się być wystarczający. W innym

przypadku koder może przesłać informacje o dodatkowych składnikach tonalnych, które mają być odtworzone po stronie dekodera. W dekoderze mogą być zastosowane również inne operacje mające na celu nadanie odtworzonemu sygnałowi charakteru podobnego do oryginału.



Rysunek 2.6: Idea rekonstrukcji składowych dużych częstotliwości w technice SBR.

Parametryczną technikę rozszerzania widma opartą o kopiowanie składowych widmowych zastosowano w wielu nowoczesnych kodekach sygnałów fonicznych. Aktualnym standardem i rozwiązaniem uważanym aktualnego za najlepsze w tej dziedzinie jest technika Spectral Band Replication (SBR) będąca częścią standardu MPEG-4 AAC HE [68]. Powstało również kilka innych technik, których zasada działania jest zbliżona do SBR i opiera się na kontrolowanym przesunięciu składowych widma. Najważniejsze z nich to PlusV [153, 154], technika zaproponowana przez firmę VSLI Solution. Oferowała ona jakość zdekodowanego sygnału zbliżoną do SBR, jednak nie jest już wykorzystywana. Inną techniką jest Spectral Extension [47], będąca częścią standardu EAC-3 [40] wprowadzonego przez firmę Dolby Digital. W technice ARM-WB+ zaproponowanej przez konsorcjum 3GPP, w module generacji dużych częstotliwości wykorzystano nadpróbkowanie [16], tak jak to zostało pokazane na rysunku 2.5. D. Sinha [142] zaproponował technikę kopiowania fragmentów widma bazując na wzajemnym podobieństwie już zrekonstruowanych składowych widmowych. Technika ta umożliwia zastosowanie większej ilości kopiowanych fragmentów widma w stosunku do np. techniki SBR.

2.4.2. Metody oparte na modelowaniu sinusoidalnym

Obecnie istnieje niewiele rozwiązań technik rozszerzania widma wykorzystujących modelowanie sinusoidalno-szumowe dla muzyki i dźwięku szerokopasmowego. Tego typu podejście pozwala precyzyjnie kontrolować parametry sinusoid, a przez to zapewnić poprawną rekonstrukcję struktury harmonicznej sygnału, co nie zawsze jest możliwe w przypadku technik wykorzystujących replikację widma.

Najistotniejszym rozwiązaniem w tej dziedzinie jest technika Accurate Spectral Replacement (ASR) zaproponowana przez Ferreirę [44]. Dokładniejszy opis algorytmu znajduje się w rozdziale 2.8.

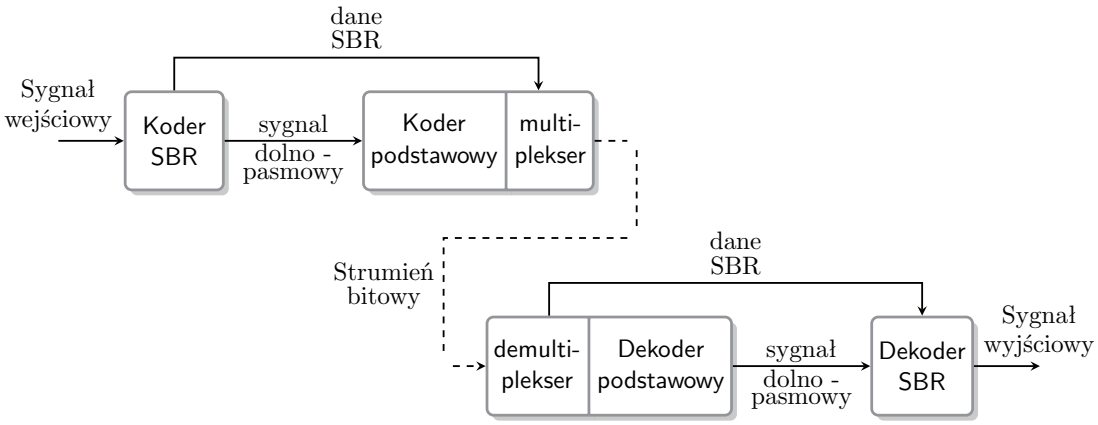
2.4.3. Metody oparte na zniekształceniu nieliniowym

Algorytmy rozszerzania widma, które nie korzystają z dodatkowych informacji przesłanych przez koder określane są w literaturze jako techniki ślepe (ang. *blind bandwidth extension*). Znajdują one zastosowanie zarówno w kodowaniu dźwięku szerokopasmowego i mowy. W przypadku kodowania sygnału mowy można zastosować modele opisujące parametry mowy i na ich podstawie zrekonstruować składowe HF. Trudno jednak znaleźć model, który jest w stanie opisać muzykę, w której występować może wiele źródeł dźwięku o różnej charakterystyce. Z tego powodu stosując ślepe techniki rekonstrukcji składowych HF wykorzystuje się proste metody bazujące na generacji harmonicznych w układzie nieliniowym [103, 104, 109] lub ekstrapolację liniową obwiedni widma [25, 57, 59, 108]. Techniki [25, 57, 59, 108] znalazły zastosowanie jako rozszerzenie schematu kompresji perceptualnej [28] po stronie dekodera.

2.5. Spectral Band Replication - przykład techniki opartej na replikacji widma

2.5.1. Wprowadzenie

Technika Spectral Band Replication stanowi część standardu MPEG-4 AAC HE [67], który jest wynikiem najwyższych osiągnięć w dziedzinie kodowania dźwięku szerokopasmowego. W rozprawie autor prezentuje również technikę usprawnienia algorytmu SBR.



Rysunek 2.7: Ogólny schemat techniki SBR.

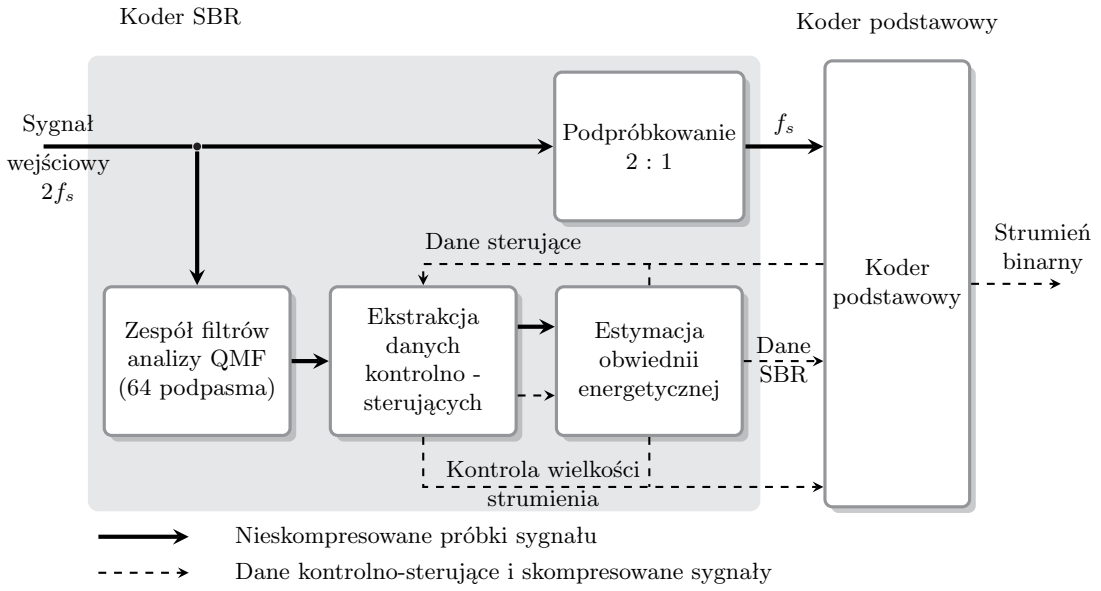
Podstawowy schemat kompresji wykorzystujący technikę SBR przedstawiony został na rysunku 2.7. Sygnał wejściowy jest najpierw przetwarzany przez koder SBR na którego wyjściu otrzymujemy sygnał dolnopasmowy oraz strumień binarny SBR. Następnie sygnał foniczny poddawany jest kompresji w koderze podstawowym. Ostatecznie dane z kodera podstawowego i SBR są umieszczane we wspólnym strumieniu bitowym.

Składowe HF kodowane przez koder SBR reprezentowane są przez stosunkowo niewielki strumień danych rzędu 1 – 3 kbit/s. Jest to znaczna redukcja informacji w stosunku do tradycyjnego kodowania transformatowego (AAC). Dzięki zmniejszeniu puli bitów na reprezentację składowych HF, koder podstawowy może przeznaczyć więcej bitów na składowe LF. W ten sposób otrzymujemy lepszą reprezentację składowych LF, w stosunku do kodowania pełnopasmowego przy założonej tej samej docelowej prędkości bitowej.

Po stronie dekodera dane ze strumienia bitowego są demultipleksowane i rozdzielane odpowiednio do dekodera podstawowego i SBR. W pierwszej kolejności rekonstruowany jest sygnał LF w dekoderze podstawowym. Na podstawie otrzymanego w ten sposób sygnału oraz danych pochodzących ze strumienia SBR rekonstruowany jest sygnał pełnopasmowy. Zaletą tego typu podejścia jest możliwość zdekodowania strumienia przez dekoder nie obsługujący strumienia SBR. Otrzymany w ten sposób sygnał jest jednak ograniczony pasmowo.

2.5.2. Proces kodowania

Podstawowy schemat kodera SBR przedstawiony został na rysunku 2.8. Głównym elementem zarówno kodera jak i dekodera jest polifazowy zespół filtrów QMF (ang. *Quadrature Mirror Filter*) o modulacji zespolonej. Zespół filtrów QMF ma za zadanie podzielić sygnał



Rysunek 2.8: Schemat kodera SBR.

wejściowy na grupę sygnałów podpasmowych. W dalszych etapach kodowania sygnału, koder SBR operuje w dziedzinie podpasm QMF.

Sygnał na wejściu kodera SBR jest poddany filtracji w zespole filtrów QMF, który dzieli sygnał na 64 zespolone sygnały podpasmowe. Dla każdych 64 wejściowych próbek w dziedzinie czasu na wyjściu zespołu filtrów otrzymujemy 64 próbki podpasmowe. Przyjmując, że częstotliwość próbkowania f_s wynosi 44,1 kHz otrzymana szerokość pasma QMF wynosi odpowiednio 344 Hz, natomiast rozdzielczość czasowa 1,4 ms. W dalszych blokach przetwarzania koder SBR operuje na wartościach zespolonych.

Kolejnym elementem procesu kodowania jest ekstrakcja danych kontrolno - sterujących. Na proces ten składa się wykrycie transjentów (szybkozmiennych elementów sygnału), wyznaczenie tonalności w pasmach skalowania, określenie poziomu szumu oraz detekcja składowych sinusoidalnych w widmie składowych HF.

Zakres działania techniki SBR

Koder SBR rekonstruuje częstotliwości z zakresu od f_{SBR} do $f_{SBR_{max}}$. Wartość f_{SBR} ustalana jest w koderze na podstawie docelowej prędkości bitowej i stanowi wielokrotność częstotliwości granicznej pasma QMF. Zależność tą można opisać wzorem:

$$f_{SBR} = k_{SBR} \cdot \frac{f_s}{2M}, \tag{2.4}$$

Tabela 2.1: Charakterystyczne częstotliwości techniki SBR dla kodeka MPEG-4 AAC HE – częstotliwość podziału f_{SBR} i górna granica częstotliwości $f_{SBR_{max}}$ dla $f_s = 44100$ Hz. Wartości pochodzą z implementacji opracowanej przez konsorcjum 3GPP[2] oraz firmę Coding Technologies [116]. Przedstawione prędkości bitowe odpowiadają zakresowi badań ujętych w rozprawie.

prędkość bitowa $[kb/s]$	$\mathbf{f_{SBR}}$ $[kHz]$	$\mathbf{f_{SBR_{max}}}$ $[kHz]$
16	3,44	10,6
20	4,13	11,7
24	5,50	14,8

gdzie k_{SBR} to indeks pasma QMF, od którego koder SBR rozpoczyna analizę sygnału HF, f_s - częstotliwość próbkowania, M - liczba pasm QMF. Współczynnik k_{SBR} przesyłany jest do dekodera.

Górna granica pracy dekodera SBR - $f_{SBR_{max}}$ zależy od liczby pasm QMF rekonstruowanych po stronie dekodera, zgodnie ze wzorem:

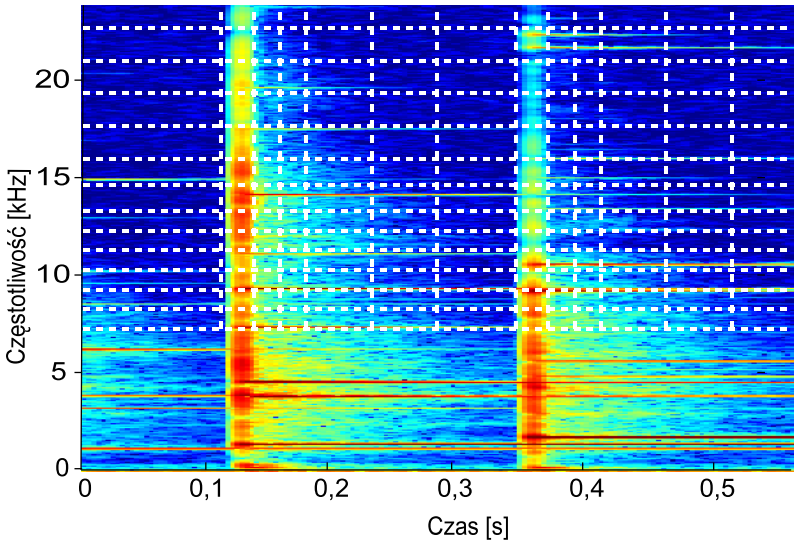
$$f_{SBR_{max}} = p \cdot \frac{f_s}{2M}, \tag{2.5}$$

gdzie p to liczba rekonstruowanych pasm QMF.

Wartości f_{SBR} oraz $f_{SBR_{max}}$, przedstawione w tabeli 2.1, są wspólne dla większości obecnych implementacji kodera MPEG-4 AAC HE [2, 116]. Przykładowo, implementacje komitetu standaryzacyjnego 3GPP oraz firmy Nero korzystają z implementacji SBR firmy Coding Technologies [2, 18].

Siatka czasowo-częstotliwościowa

Dane kontrolno-sterujące zostają wykorzystane do reprezentacji obwiedni widma składowych HF dla danej ramki (blok danych o długości 2048 próbek). Parametry te są następnie grupowane na siatce czasowo-częstotliwościowej, której przykład został pokazany na rysunku 2.9. Pojedynczy segment siatki może składać się z kilku pasm QMF na przestrzeni kilku ramek analizy czasowej. Segmenty siatki czasowo-częstotliwościowej łączone są w grupy, które tworzą wektory opisujące obwiednie widma energii dla danego zakresu czasu (ang. *SBR envelope*).



Rysunek 2.9: Rysunek przedstawia siatkę czasowo-częstotliwościową (białe przerywane linie) nałożoną na spektrogram sygnału. Siatka ilustruje granice segmentów energetycznych. Można zauważyć zmienną rozdzielczość częstotliwościową dla dużych częstotliwości oraz zmienną rozdzielczość czasową (krótkie okna w okolicy transjentów).

Poszczególne segmenty obwiedni energetycznej wyznaczane są jako średnia energia próbek podpasmowych w danym zakresie czasowo-częstotliwościowym. Zakres czasu pojedynczego segmentu energetycznego określany jest głównie na podstawie wykrytych transjentów w sygnale. W chwili, gdy sygnalizowany jest transjent, koder SBR ustawia początek nowego segmentu energetycznego w miejscu wystąpienia transjentu. W momencie wykrycia transjentu czas trwania segmentu energetycznego jest skracany do kilku ramek. W ten sposób zwiększana jest rozdzielczość czasowa analizy, co pozwala uniknąć efektu pre-echa.

Następnie, w zależności od czasu trwania transjentu, koder SBR stopniowo zwiększa czas trwania segmentu energetycznego. Dłuższe segmenty energetyczne obejmują czas wybrzmiewania transjentu.

Kolejnym krokiem jest przełączenie analizy SBR na segmenty energetyczne obejmujące długi czas trwania sygnału. Segmenty tego typu znajdują zastosowanie dla stacjonarnych framgmentów sygnału, gdzie nie ma gwałtownych zmian amplitudy.

Poza rozdzielczością czasową segmentów energetycznych koder SBR musi podjąć również decyzję o ich rozdzielczości częstotliwościowej. W koderze SBR stosowane są dwa schematy grupowania podpasm QMF w celu uzyskania wysokiej i niskiej rozdzielczości częstotliwościowej. Segmenty energetyczne o mniejszej rozdzielczości częstotliwościowej składają się z dwa razy mniejszej liczby podpasm QMF w stosunku do wysokiej rozdzielczości częstotliwościowej.

Aby uniknąć chwilowego wzrostu prędkości bitowej, koder SBR stosuje dwa schematy grupowania podpasm w dziedzinie częstotliwości. Następnie, użyty jest schemat, którego zastosowanie nie spowoduje przekroczenia limitu dostępnych bitów. Siatka czasowo-częstotliwościowa przedstawiona na rysunku 2.9 została nałożona na spektrogram sygnału. Można zaobserwować, że wokół transjentów jest większa rozdzielczość czasowa, kosztem częstotliwościowej. Odwrotna sytuacja występuje dla fragmentów sygnału o stacjonarnym charakterze.

Pojedynczy segment energetyczny może być stosowany dla szerokiego zakresu czasu i częstotliwości. Jednak rozdzielczość czasowo-częstotliwościowa, z jaką można określić granice segmentu energetycznego, zależy od rozdzielczości zespołu filtrów analizy QMF. Konstrukcja zespołu filtrów umożliwia zadawalającą rozdzielczość czasowo-częstotliwościową w celu poprawnej estymacji obwiedni częstotliwościowej dla szerokiego zakresu sygnałów.

W koderach kształtu fali, tj. MP3 lub AAC stosowany jest adaptacyjny mechanizm zmiany rozdzielczości zespołu filtrów, w celu dostosowania kodera do charakteru zmian sygnału. W technice SBR stosowany jest zespół filtrów o stałej rozdzielczości. Możliwa jest jednak poprawna estymacja obwiedni widma poprzez adaptacyjne grupowanie i uśrednianie energii w podpasmach.

Estymacja współczynnika tonalności

Wyznaczanie dodatkowych parametrów opisujących obwiednię energetyczną sygnału odbywa się na zasadzie analizy przez syntezę. Koder rekonstruuje składowe HF poprzez replikację widma, a następnie porównuje tak otrzymany sygnał z oryginałem. Jednym z problemów jest poprawne oszacowanie stosunku energii tonalnej do szumowej TNR (ang. *tonal to noise ratio*) w sygnale oryginalnym oraz zrekonstruowanym w procesie replikacji widma. Na tej podstawie koder decyduje ile szumu ma być dodane po stronie dekodera, aby otrzymać poprawny stosunek TNR dla składowych HF.

Standard MPEG-4 AAC HE nie precyzuje w jaki sposób ma być wyznaczony stosunek energii TNR. Powstały jednak prace, które rozwiązują ten problem poprzez wyznaczenie współczynnika tonalności dla każdego z podpasm QMF. Współczynnik ten jest następnie przeliczany na stosunek TNR. Jednym z podejść do tego problemu jest zastosowanie modelowania sinusoidalno-szumowego dla składowych HF [134]. Na podstawie proporcji

wyznaczonych składników tonalnych do szumowych w danym podpaśmie QMF wyznaczany jest współczynnik tonalności.

Inne metody bazują na zastosowaniu predykcji liniowej niskiego rzędu dla każdego podpasma QMF i na tej podstawie wyznaczenia energii składowych tonalnych [58]. Podobna metoda została zaimplementowana w koderze MPEG-4 AAC HE v2 zaproponowanym przez konsorcjum 3GPP [2]. W tej implementacji zastosowano predykcję liniową drugiego rzędu. Koder autorstwa 3GPP został wykorzystany podczas badań eksperymentalnych przeprowadzonych przez autora rozprawy.

Na podstawie współczynnika tonalności koder SBR oblicza dodatkowe parametry w podpasmach QMF, t.j. poziom energii szumów, współczynniki filtracji odwrotnej LPC, oraz położenie i energię składników tonalnych.

Kwantyzacja i kodowanie

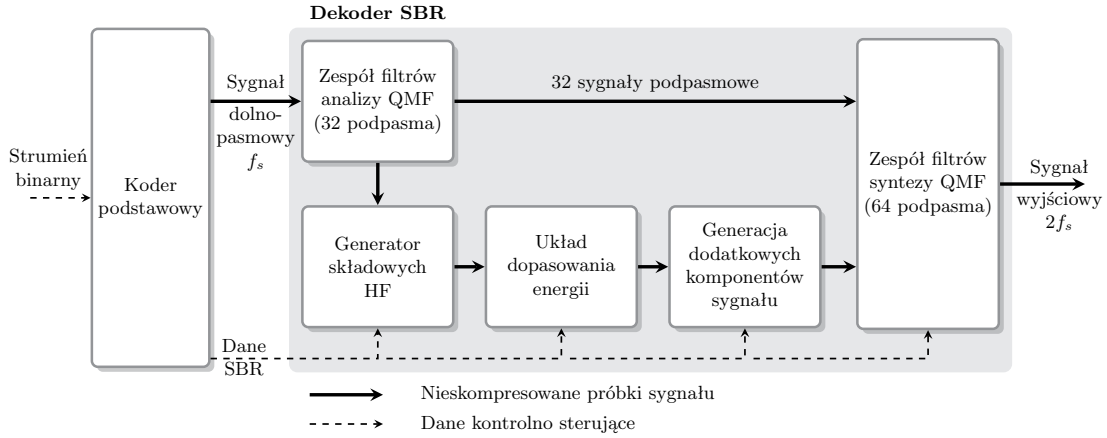
Współczynniki opisujące obwiednię energetyczną, składowe tonalne oraz szumowe, poddawane są kwantyzacji oraz kodowaniu różnicowemu. Kodowanie odbywa się w kierunku osi czasu lub częstotliwości tak aby otrzymać najmniejszy strumień binarny. Następnie dane kodowane są entropijnie wykorzystując kodowanie Huffmana.

2.5.3. Proces dekodowania

Schemat blokowy dekodera SBR został przedstawiony na rysunku 2.10. Strumień binarny podany jest na wejście dekodera podstawowego, który dekoduje sygnał dolnopasmowy. Następnie sygnał dolnopasmowy wraz z danymi opisującymi składowe HF podany zostaje na wejście dekodera SBR. Kolejnym etapem jest analiza sygnału dolnopasmowego w zespole filtrów QMF dekomponującego sygnał na 32 podpasma. Grupa sygnałów podpasmowych wraz z danymi sterującymi zawartymi w strumieniu binarnym umożliwia odtworzenie składowych HF w generatorze dużych częstotliwości. Otrzymany w ten sposób sygnał składowych HF poddany jest operacji korekcji obwiedni widma w układzie dopasowania energii. Następnie w zależności od charakteru sygnału dodawane są składniki szumowe lub tonalne (sinusoidy). Duże znaczenie dla poprawnej rekonstrukcji składowych HF ma również stopień korelacji składowych o małych i dużych częstotliwościach. Zależność ta zmienia się dla różnych sygnałów fonicznych. Przykładowo jeśli w paśmie małych częstotliwości jest

więcej składowych tonalnych niż w paśmie dużych częstotliwości, wówczas stosowany jest proces odwrotnej filtracji na sygnale podpasmowym HF, filtrem FIR niskiego rzędu.

Ostatecznie otrzymujemy sygnał dolnopasmowy zdekodowany w dekoderze podstawowym (AAC) oraz sygnał składowych HF zrekonstruowany w dekoderze SBR. W zespole 64 filtrów syntezy QMF łączymy otrzymane sygnały podpasmowe otrzymując sygnał pełnopasmowy, który ma dwukrotnie większą częstotliwość próbkowania w porównaniu do sygnału zdekodowanego w dekoderze podstawowym.



Rysunek 2.10: Schemat blokowy dekodera SBR.

Generator składowych HF

Zespolone sygnały podpasmowe otrzymane w wyniku filtracji w zespole filtrów analizy QMF poddane są następnie, przetwarzaniu w bloku generatora składowych HF, którego zadaniem jest utworzenie grupy sygnałów podpasmowych reprezentujących duże częstotliwości. Algorytm rekonstrukcji wybiera sygnały podpasmowe reprezentujące składowe LF, a następnie na podstawie określonych reguł kopiowania tworzy sygnały podpasmowe odpowiadające składowym HF. Kopiowanie w dziedzinie częstotliwości odbywa się na pewnej grupie sygnałów podpasmowych zwanej łatą (ang. *patch*). Zakres kanałów QMF wchodzących w skład łaty oraz położenie źródłowych i docelowych sygnałów podpasmowych dekoder SBR odczytuje ze strumienia binarnego. Poniżej przedstawiono warunki jakie musi wziąć pod uwagę algorytm generujący strukturę łat:

- Łat SBR powinno być jak najmniej i powinny pokrywać zakres częstotliwości do 16 kHz. Standard SBR przewiduje maksymalnie 5 łat. W skład pojedynczych łat

nie powinny wchodzić podpasma QMF o najniższych indeksach (małych częstotliwościach), ponieważ zawierają one składową stałą sygnału.

- Należy unikać sytuacji, w której powstaje nieciągłość struktury harmonicznej sygnału. W praktyce warunek ten jest trudny do spełnienia dla sygnałów o bogatym widmie harmonicznym, ze względu na niewielką liczbę łat możliwych do zastosowania.

- Zakres częstotliwości podlegających kopiowaniu powinien zaczynać się dla czętotliwości podziału f_{SBR}.

Tego typu podejście do rekonstrukcji składowych HF bazuje na założeniu, że składowe o małych i dużych częstotliwościach są ze sobą silnie skorelowane. W rzeczywistych sygnałach taka sytuacja nie musi jednak wystąpić. Składowe LF i HF w wielu sygnałach mogą mieć zupełnie inny charakter, a prosty algorytm replikacji widma nie jest w stanie poprawnie odtworzyć ich struktury. Szerzej ten problem zostanie omówiony w rozdziale 2.6.

Mając na uwadze problemy wynikające z niskiej korelacji składowych LF i HF w technice SBR stosuje się filtrację odwrotną z zastosowaniem współczynników LPC. Tego typu operacja powoduje wygładzanie widma lub wybielanie (ang. *whitening*), którego celem jest rozszerzenie serii harmonicznych sygnałów dużych częstotliwości, co jednocześnie nadaje sygnałowi szumowy charakter. Filtracja jest realizowana na każdym z zespolonych sygnałów podpasmowych z osobna przy wykorzystaniu filtrów FIR niskiego rzędu. Współczynniki filtru wyznaczane są na podstawie analizy sygnału pochodzącego z dekodera podstawowego oraz danych ze strumienia binarnego SBR. Transmitancję $H(z)$ omawianego filtru przedstawia wzór:

$$H(z) = 1 + \alpha_0 z^{-1} + \alpha_1 z^{-2}. \tag{2.6}$$

Współczynniki α_1 i α_2 wyznaczane są metodą kowariancyjną. Dla zastosowanej filtracji LPC uwzględnia się również współczynnik rozszerzenia pasma bw (ang. *bandwidth*), który w normie MPEG-4 AAC HE określany jest jako „chirp factor". Współczynnik ten określa w jakim stopniu sygnał ma zbliżony charakter do szumu białego. Wartość tego parametru wyznaczana jest w koderze i przesyłana jest do dekodera. Ostatecznie proces rekonstrukcji pojedynczego pasma QMF o numerze k opisany jest wzorem:

$$\begin{aligned} X_{High}(k,l) \;&=\; X_{Low}(p,l) + \\ &+ \alpha_0(p) \cdot bw \cdot X_{Low}(p,l-1) \;+\; \alpha_1(p) \cdot bw^2 \cdot X_{Low}(p,l-2), \end{aligned} \tag{2.7}$$

gdzie:

$X_{Low}(p,l)$ - jest próbką sygnału podpasma małych częstotliwości, wyznaczaną dla p-tego subpasma i l-tej próbki sygnału w tym paśmie,

$X_{High}(k,l)$ - jest próbką sygnału podpasma dużych częstotliwości, wyznaczaną dla k-tego subpasma i l-tej próbki sygnału w tym paśmie.

Układ dopasowania energii widma

Sygnał otrzymany na wyjściu generatora HF w dalszym ciągu znacząco różni się od oryginału. Generator HF rekonstruuje sygnały podpasmowe starając się zachować właściwości tonalne i szumowe każdego z nich, nie oddaje jednak poprawnie właściwości energetycznych. Zadaniem układu dopasowania energii jest poprawne ukształtowanie obwiedni widma w danych segmentach czasowo-częstotliwościowych.

Wartości próbekk sygnału w podpasmach są skalowane współczynnikiem wzmocnienia $G(k)$, tak aby poziom energii zrekonstruowanego sygnału HF odpowiadał oryginałowi. Dodatkowo sygnał HF uzupełniany jest o składowe tonalne i szumowe, których energia zależy proporcjonalnie od współczynnika TNR. Ostatecznie zrekonstruowany sygnał $Y(k,l)$ dla k-tego pasma QMF, opisany jest wzorem:

$$Y(k,l) = G(k) \cdot X_{High}(k,l) + C(k)e^{j\theta(k,l)}, \tag{2.8}$$

gdzie ostatni składnik sumy reprezentuje składnik korekcji tonalno-szumowej, $C(k)$ jego energię i $\theta(k,l)$ jest szumem o losowej fazie lub sinusoidą o ustalonej fazie. Współczynniki $G(k)$ oraz $C(k)$ wyznaczane są po stronie dekodera na podstawie współczynników skalujących energię sygnału E_o oraz współczynnika Q, określającego poziom energii składowych szumowych w sygnale. Współczynnik E i Q obliczane są po stronie kodera SBR.

Przed przystąpieniem do procesu skalowania sygnałów podpasmowych niezbędne jest wyznaczenie energii dla każdego segmentu energetycznego. Jeśli oś czasu podzielimy na T odstępów o granicach $\{l_t\}_{t=0}^{T}$ oraz oś częstotliwości podzielimy na P podpasm o granicach $\{k_p\}_{p=0}^{P}$, wówczas średnia energia w segmencie (p,t) jest określona wzorem:

$$\bar{E}(p,t) = \frac{1}{\Delta} \sum_{k=k_{p-1}}^{k_p-1} \sum_{l=l_{t-1}}^{l_t-1} |X(k,l)|^2, \tag{2.9}$$

gdzie $\Delta = (k_p - k_{p-1})(l_t - l_{t-1})$. Na podstawie wzoru (2.9) po stronie kodera wyznaczana jest energia $E_o(k)$ reprezentująca energię składowych HF w sygnale oryginalnym.

Dodatkowo dekoder wyznacza energię składowych HF utworzonych w generatorze dużych częstotliwości $E_r(k)$. Energie $E_o(k)$ oraz $E_r(k)$ dotyczą tego samego przedziału czasu $l_t < l < l_{t+1}$.

Układ dopasowania energii dąży do jak najlepszej percepcyjnej aproksymacji sygnału oryginalnego. Na podstawie współczynników skalujących można wyznaczyć odpowiednie wzmocnienie podpasm QMF oraz energię dodatkowych składowych. Należy zauważyć, że jeśli w dekoderze SBR dla danego podpasma QMF dodawana jest składowa szumowa, wówczas nie można dodać składowej tonalnej i na odwrót.

Jeśli do sygnału podpasmowego dodawany jest szum, wówczas standard MPEG-4 AAC HE [67, 68] w następujący sposób definiuje parametry $G(k)$ oraz $C(k)$:

$$G(k) = \sqrt{\frac{E_o(k)}{E_r(k)(1+Q(k))}}, \qquad C(k) = \sqrt{\frac{E_o(k)Q(k)}{1+Q(k)}}. \tag{2.10}$$

Jeśli dodawany jest składnik tonalny wówczas parametry $G(k)$ oraz $C(k)$ opisane są wzorem:

$$G(k) = \sqrt{\frac{E_o(k)Q(k)}{E_r(k)(1+Q(k))}}, \qquad C(k) = \sqrt{\frac{E_o(k)}{1+Q(k)}}. \tag{2.11}$$

Dodanie składnika tonalnego oznacza wygenerowanie pojedynczego tonu sinusoidalnego o energii określonej przez wartość obwiedni widma dla danego podpasma QMF. Operacja ta nie jest precyzyjna, ponieważ koder nie przesyła wartości częstotliwości sinusoidy. Dodatkowa informacja zawiera numer podpasma QMF, w którym ma być wygenerowana sinusoida. Według standardu MPEG-4 AAC HE częstotliwość dodatkowej sinusoidy równa się połowie szerokości aktualnego podpasma QMF. Ponieważ całe pasmo składa się z 64 podpasm QMF, maksymalna rozdzielczość częstotliwościowa dla pojedynczej sinusoidy wynosi:

$$\Delta f = \frac{f_s/2}{M}, \tag{2.12}$$

gdzie $M = 64$ to liczba podpasm QMF. Jeśli $f_s = 44100$, $M = 64$, wówczas $\Delta f \approx 344,5$ Hz. Wygenerowany ton sinusoidalny ma stałą częstotliwość na odcinku danego segmentu energetycznego. Oznacza to, że technika SBR nie jest w stanie odtworzyć składowych sinusoidalnych o szybkich zmianach chwilowej częstotliwości.

Dla wyznaczonego poziomu wzmocnienia $G(k)$ dekoder SBR stosuje procedurę progowania. Jej celem jest ograniczenie poziomu energii w danym podpaśmie QMF, w stosunku do energii podpasma sygnału oryginalnego. Dodatkowo funkcja progująca nie pozwala na powstawanie zbyt dużych różnic energii w sąsiednich podpasmach QMF.

Synteza QMF

Zrekonstruowany sygnał HF oraz sygnał dolnopasmowy pochodzący z kodera podstawowego zostają poddane przetwarzaniu w wyjściowym zespole filtrów syntezy QMF. Zespół filtrów pracuje z tą samą częstotliwością, co sygnał oryginalny. Na wyjściu zespołu filtrów otrzymujemy sygnał pełnopasmowy o wartościach rzeczywistych. Część urojona sygnału jest odrzucona.

2.6. Zniekształcenia wprowadzanie przez techniki rozszerzania widma

Przedstawione w poprzednich rozdziałach techniki rozszerzania widma mają szerokie zastosowanie w dziedzinie kompresji sygnałów cyfrowych. Ich główną zaletą jest mała prędkość bitowa przy stosunkowo wysokiej jakości zdekodowanego sygnału.

Techniki rozszerzania widma rekonstruują składowe HF sygnału w zakresie od około 4 kHz do częstotliwości Nyquista f_N. Oznacza to, że w skrajnych przypadkach i częstotliwości próbkowania $f_s = 44100Hz$ rekonstrukcja obejmuje około 80 % składowych widma.

Zakres działania technik rozszerzania widma przy jednoczesnym wysokim stopniu kompresji sygnału składowych hf, nie pozostaje jednak bez wpływu na jakość rekonstruowanego sygnału fonicznego. W zdekodowanym sygnale powstaje szereg dokuczliwych artefaktów [107]. Problem staje się istotny zwłaszcza dla sygnałów, które zawierają nagrania instrumentów o bogatym widmie harmonicznym.

Dodatkowo dla bardzo małych prędkości bitowych poniżej 30 kb/s dekoder podstawowy generuje sygnał składowych LF o niskiej jakości. Często w sygnale LF zniekształceniu na skutek kwantyzacji ulegają podstawowe składowe źródeł dźwięku występujących w sygnale. W takiej sytuacji techniki BWE, a zwłaszcza bazujące na replikacji widma [68], już na wejściu otrzymują niepoprawne dane.

W rozdziale tym zostaną zaprezentowane jedne z najistotniejszych zniekształceń, z którymi mamy do czynienia przy stosowaniu technik rozszerzania widma. Skupiono się głównie na zniekształceniach powodowanych w tonalnej części sygnału, istotnych z punktu widzenia proponowanych technik BWE. Pozostałe zniekształcenia technik rozszerzania widma zostały zaprezentowane w celach przeglądowych i porównawczych.

2.6.1. Szorstkość dźwięku

Jednym z problemów który dotyczy metod BWE jest szorstkość dźwięku (ang. *auditory roughness*) [161], która określana jest jako postrzeganie gwałtownych zmian amplitudy tonu sinusoidalnego. Sygnał zawierający dwie sinusoidy jest odbierany przez słuchaczy jako „szorstki", jeśli różnica częstotliwości pomiędzy tymi tonami zawiera się w przedziale od 30 do 600 Hz. Odczuwalna dokuczliwość efektu szorstkości dźwięku zależy również od położenia tonów w widmie amplitudowym i ma swoje maksimum dla 1 kHz.

Źródłem efektu szorstkości dźwięku, w technikach rozszerzania widma, jest operacja kopiowania widma opisana w rozdziale 2.4.1. Proste kopiowanie, które odpowiada przesunięciu widmowemu uniemożliwia prawidłowe odtworzenie proporcjonalnych zmian częstotliwości składowych tonalnych i ich harmonicznych relacji.

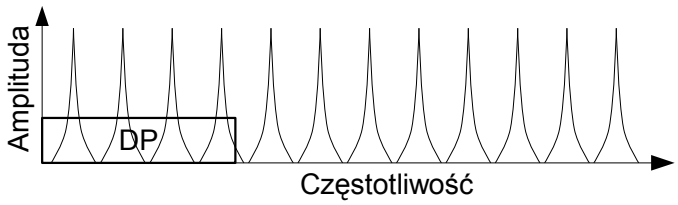
2.6.2. Dudnienia

Dodatkowo, na granicy połączenia składowych małych i dużych częstotliwości może powstać efekt dudnienia. Efekt ten powstaje, gdy składowe tonalne wygenerowane w wyniku kopiowania, znajdą się w zbyt małej odległości od siebie. Powoduje to powstanie okresowych drgań amplitudy, co w efekcie może również objawiać się jako efekt szorstkości dźwięku. Rysunek 2.11b obrazuje powstanie tego zjawiska na przykładzie techniki SBR. Jeśli w tej samej chwili czasu mamy do czynienia z dwoma tonami sinusoidalnymi o takiej samej amplitudzie, wówczas otrzymujemy sygnał opisany wzorem:

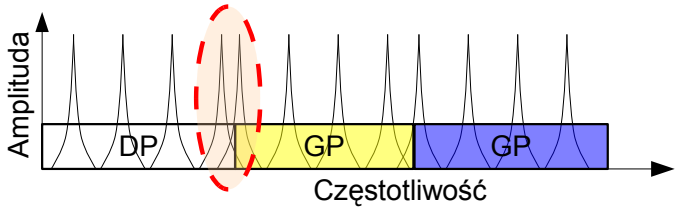
$$\begin{aligned} x(t) &= \sin(\omega_1 t) + \sin(\omega_2 + \varphi) \\ &= 2\cos\left(\Delta\omega t + \frac{\varphi}{2}\right)\sin\left(\bar{\omega} t + \frac{\varphi}{2}\right), \end{aligned} \tag{2.13}$$

gdzie $\Delta\omega = (\omega_2 - \omega_1)/2$ oraz $\bar{\omega} = (\omega_2 + \omega_1)/2$. Jeśli, dwa tony sinusoidalne znajdują się blisko siebie, czyli $\Delta\omega$ jest mała, wówczas wolno zmienna funkcja kosinusoidalna moduluje szybkozmienną składową sinusoidalną powodując powstawanie okresowych dudnień. Na tego typu efekt szczególnie narażona jest technika SBR, w której syntetycznie generowane składowe tonalne mają niepoprawną częstotliwość, co wynika z ograniczeń standardu [68].

Dodatkowo zastosowanie kopiowania widma w technice rozszerzania widma może spowodować przesunięcie do tego samego pasma krytycznego [136] składowych tonalnych,



(a) Oryginał



(b) SBR

Rysunek 2.11: Powstawanie efektu dudnienia oraz nieharmoniczności podczas procesu kopiowania widmowego.

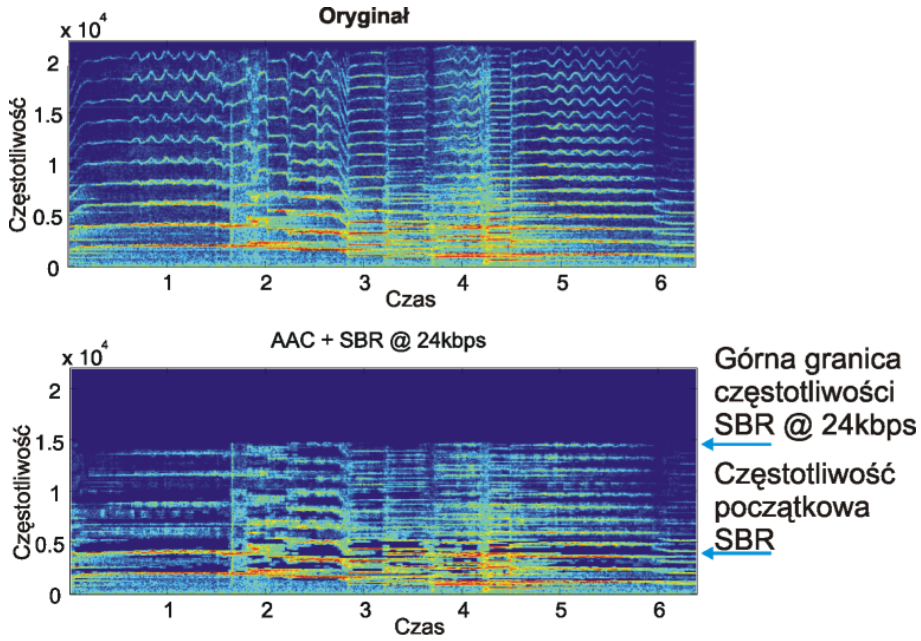
które naturalnie znajdowały się w różnych pasmach krytycznych [161]. Dzieje się tak, ponieważ szerokość pasm krytycznych wzrasta wraz z częstotliwością, natomiast kopiowanie widma zmienia relacje częstotliwościowe wygenerowanych składowych w stosunku do synalu naturalnego. W takim przypadku, składowe tonalne będą postrzegane jako pojedynczy składnik tonalny o zmiennej w czasie modulacji amplitudy.

2.6.3. Składniki tonalne o zmiennej częstotliwości

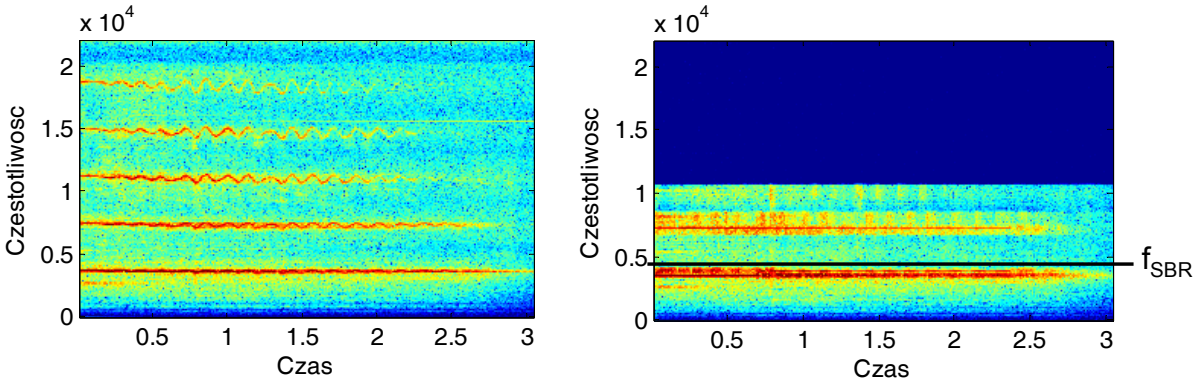
Na rysunkach 2.12 oraz 2.13 zaprezentowano sposób w jaki klasyczna technika SBR rekonstruuje składniki tonalne. Możemy zauważyć, że istotnym problemem jest rekonstrukcja sygnałów zawierających składowe tonalne o stosunkowo szybko zmieniającej się częstotliwości chwilowej (np. efekt vibrato).

Efekt ten jest szczególnie widoczny na rysunku 2.13, na którym przedstawiony jest spektrogram krótkiego fragment nagrania skrzypiec. Dekoder SBR skopiował fragmenty widma poniżej częstotliwość f_{SBR} i przeniósł je w zakres składowych dużych częstotliwość. Ponieważ mamy do czynienia ze zwykłym przesunięciem w dziedzinie częstotliwości, więc kształt zrekonstruowanej wyższej harmonicznej jest dokładnie taki sam, jak jej dolnopasmowy odpowiednik. W naturalnym sygnale kolejne składowe szeregu harmonicznego są wielokrotnością częstotliwości podstawowej f_0, a więc wyższe harmoniczne mają większą modulację częstotliwości. Tego typu błędy wpływają znacząco na barwę dźwięku, mogą również zupełnie zmienić charakter sygnału.

Szczególnie wyraźnie ten problem widoczny jest na rysunku 2.14, który przedstawia sygnał syntetyczny szybkozmiennych składowych. Zrekonstruowany w dekoderze MPEG-4 AAC HE zawiera znaczące zniekształcenia sygnału powstałe w wyniku replikacji widma.



Rysunek 2.12: Przykład rekonstrukcji składowych tonalnych z wykorzystaniem MPEG-4 AAC HE (SBR).

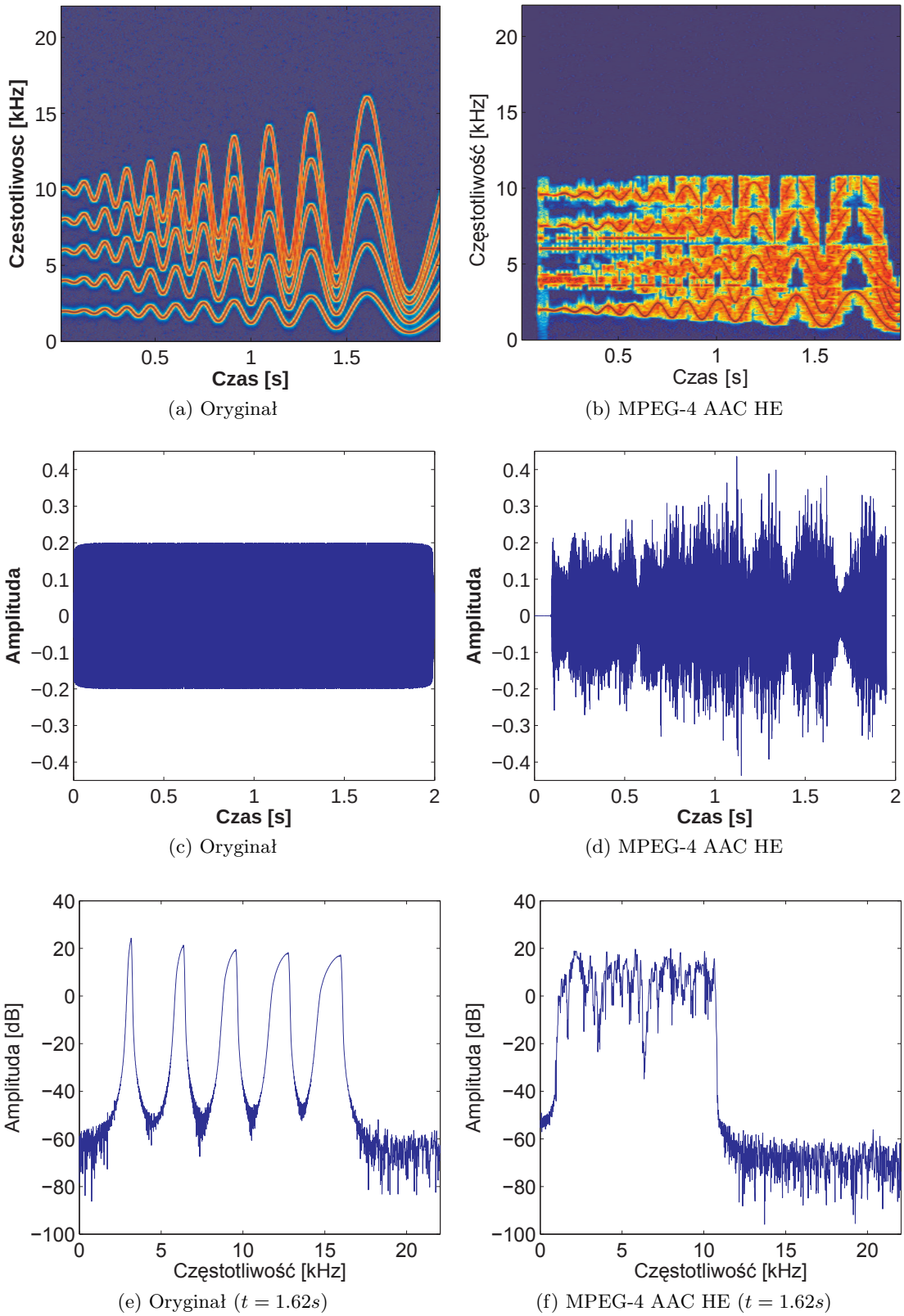


(a) Oryginał (b) MPEG-4 AAC HE

Rysunek 2.13: Przykład rekonstrukcji składowych tonalnych z wykorzystaniem MPEG-4 AAC HE (SBR).

2.6.4. Problem klasyfikacji składowych tonalnych i szumowych

Składowe tonalne, których częstotliwość chwilowa bardzo gwałtownie ulega zmianie, widoczne są w krótkookresowym widmie nie jako pojedyncze sinusoidy, lecz jako sygnały



(a) Oryginał (b) MPEG-4 AAC HE

(c) Oryginał (d) MPEG-4 AAC HE

(e) Oryginał ($t = 1.62s$) (f) MPEG-4 AAC HE ($t = 1.62s$)

Rysunek 2.14: Przykład rekonstrukcji składowych tonalnych z wykorzystaniem MPEG-4 AAC HE (SBR).

wąskopasmowe. Dla tego typu komponentów dźwięku bardzo trudno jest dokonać klasyfikacji tonalnej, a ich widmo amplitudowe może przypominać składnik szumowy.

Algorytm wyznaczania tonalności stosowany w technice SBR operuje w ramach danych z jednej ramki analizy. Powoduje to, że wiele składników o szybkich zmianach lub głębokiej modulacji częstotliwości klasyfikowanych jest jako szum i nie są poprawnie rekonstruowane. Przykład tego efektu został zaprezentowany na rysunku 2.15, gdzie zamiast składowej sinusoidalne został wygenerowany szum pasmowy. Widać to szczególnie wyraźnie dla fragmentu sygnału powyżej chwili czasu 1,5 s.

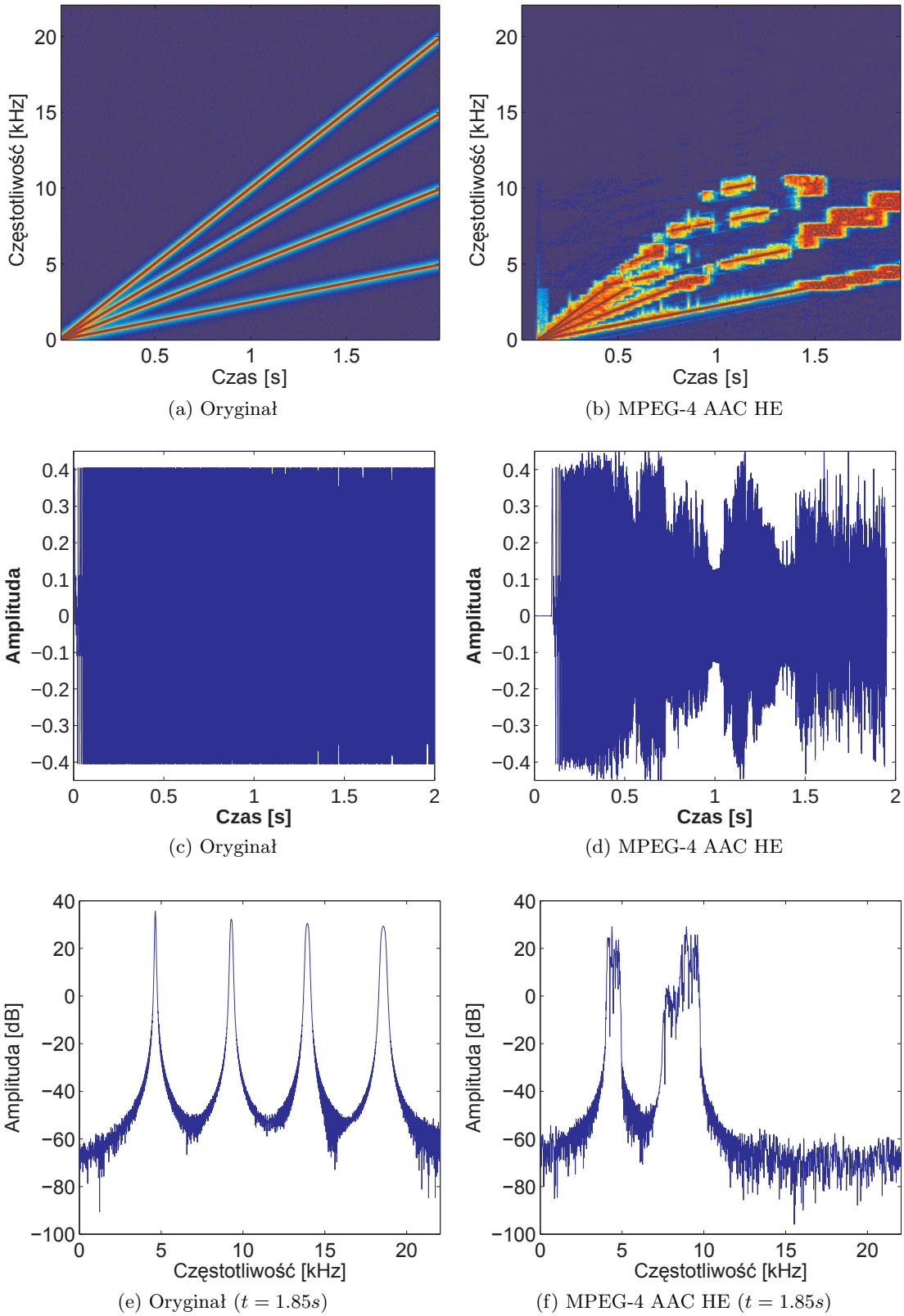
2.6.5. Rekonstrukcja brakujących składowych tonalnych w dekoderze SBR

W koderze SBR wyznaczany jest współczynnik tonalności, którego zadaniem jest określenie czy dla danego pasma skalującego należy wygenerować brakujące składniki sinusoidalne po stronie dekodera.

Dodatkowe sinusoidy nie są jednak precyzyjnie umieszczane w widmie sygnału. Dekoder generuje brakujące składniki tonalne jako sinusoidy o stałej częstotliwości, wyznaczanej jako środek aktualnie rozpatrywanego przedziału skalowania, co obrazuje wzór (2.12). Jeśli sygnał próbkowany jest z częstotliwością $f_s = 44100$ Hz, wówczas częstotliwość dodatkowych sinusoid ustalana jest z maksymalną dokładnością wynoszącą jedno podpasmo QMF czyli około 344 Hz.

Tego typu rekonstrukcja składników tonalnych powoduje powstawanie artefaktów związanych z szorstkością widma, dudnieniami oraz niepoprawną rekonstrukcją barwy dźwięku [107]. Częstotliwość chwilowa zrekonstruowanego składnika obarczona jest błędem, który rośnie wraz z pasmem skalowania (pasma skalowania mają większy zakres wraz z częstotliwością).

Niepoprawna wartość częstotliwości sinusoidy nie jest jedynym problemem rekonstrukcji składowych tonalnych w technice SBR. Syntezowane po stronie dekodera sinusoidy mają stałą wartość częstotliwości w przedziale czasu okreslonym przez pojedynczy segment siatki czasowo-częstotliwościowej. Powoduje to, że nie jest możliwe poprawne zrekonstruowanie składowych szybkozmiennych w dziedzinie czasu, które występują w sygnałach muzycznych, np. efekt vibrato.



(a) Oryginał (b) MPEG-4 AAC HE

(c) Oryginał (d) MPEG-4 AAC HE

(e) Oryginał ($t = 1.85s$) (f) MPEG-4 AAC HE ($t = 1.85s$)

Rysunek 2.15: Rysunek przedstawiający spktrogram funkcji sweep oraz pojedynczą ramkę analizy FFT.

2.7. Poprawna rekonstrukcja składowych tonalnych w technice SBR

Podstawowym mankamentem techniki SBR jest niepoprawna rekonstrukcja szeregów harmonicznych. Algorytm kopiowania pasm QMF powoduje, że zrekonstruowane szeregi harmoniczne są niekompletne lub poszczególne harmoniczne nie są wielokrotnością odpowiedniej częstotliwości f_0.

W algorytmie SBR przewidziano mechanizm, którego celem jest generowanie dodatkowych sinusoid, jeśli nie zostaną one zrekonstruowane poprzez kopiowanie widma.

Mechanizm ten okazuje się wysoce nieprecyzyjny, ponieważ generowane sinusoidy umieszczane są na środku danego pasma skalowania. Jednocześnie, decyzja o generowaniu dodatkowych sinusoid uzależniona jest od współczynnika tonalności wyznaczonego dla danego pasma skalowania. Jeśli pasmo zostanie źle sklasyfikowane, wówczas zamiast sinusoidy wzmocnieniu ulegnie szum pasmowy. W efekcie doprowadzi to do słyszalnych artefaktów.

Z tego powodu, istotnym wydaje się mechanizm poprawnej ekstrakcji parametrów techniki SBR, a zwłaszcza poprawna klasyfikacja tonalności. W pracach [58, 134] przedstawione zostały metody estymacji współczynnika tonalności oraz wpływ na jakość zdekodowanego sygnału.

Alternatywną metodę poprawnej rekonstrukcji składowych tonalnych w dekoderze MPEG-4 AAC HE zaprezentowano w pracy [135]. Główna idea polega na zastąpieniu dekodera SBR przez alternatywny dekoder bazujący na modelowaniu sinusoidalno-szumowym oraz danych ze strumienia MPEG-4 AAC HE.

Dla techniki SBR opracowano również szereg miar tonalności [58, 134, 135], których zadaniem jest określenie dla którego subpasma QMF należy wygenerować brakujące sinusoidy. Zadaniem miary tonalności jest również określenie energii dodatkowej sinusoidy. Jednakże, metody te nie rozwiązują problemów związanych z niewłaściwą reprezentacją częstotliwościową składowej tonalnej (punkt 2.6). Pomagają jednak, precyzyjnie określić energie składowej tonalnej.

2.8. Accurate Spectral Replacement - przykład techniki opartej na modelowaniu sinusoidalno-szumowym

Odmienną grupę stanowią techniki, które zamiast replikacji widma wykorzystują modelowanie sinusoidalno - szumowe. W metodach tych składowe dużych częstotliwości rekonstruowane są przy użyciu techniki LPC [97] lub dekomponowane są na część tonalną i szumową [44, 143].

Ferreira i Sinha [44] zaproponowali technikę rozszerzania widma nazwaną *Accurate Spectral Replacement* (ASR), która wykorzystuje zespół filtrów MDCT [110, 127] oraz modelowanie sinusoidalne. Technika ASR dokonuje segmentacji komponentów sygnału na składowe sinusoidalne i szumowe, a następnie przetwarza je niezależnie. Koder analizuje sygnał w nakładających się na siebie oknach analizy. Sygnał dla pojedynczego okna analizy przekształcany jest do postaci zespolonej (OFDM). W każdym oknie analizy wykrywane, a następnie usuwane są składowe sinusoidalne. Otrzymane, w ten sposób, składowe sinusoidalne przesyłane są do dekodera jako grupa parametrów opisująca ich częstotliwość, amplitudę i fazę. Pozwala to uzyskać wyższą dokładność reprezentacji sinusoid w stosunku do rozdzielczości zespołu filtrów. Residuum szumowe przekształcane jest do postaci rzeczywistych współczynników MDCT, które podlegają następnie kwantyzacji i są przesyłane do dekodera.

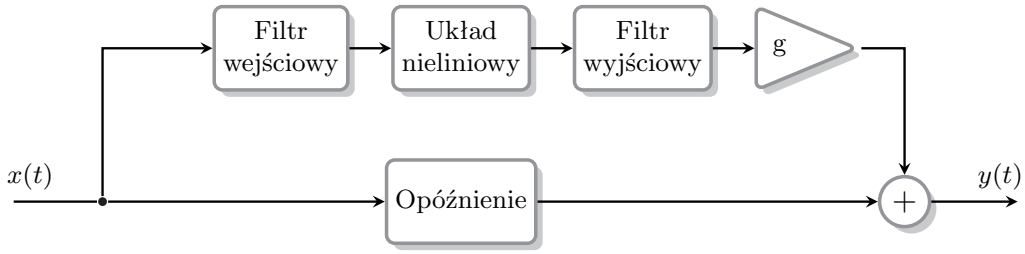
Dekoder ASR rekonstruuje składową szumową ze współczynników MDCT zawartych w strumieniu, a następnie powiela składowe szumowe w sposób podobny do techniki generacji składowych HF w dekoderze SBR. Osobno syntezowany jest sygnał opisujący trajektorie sinusoidalne. Suma sygnałów reprezentujących szum i sinusoidy daje zrekonstruowany sygnał HF.

Prędkość bitowa strumienia binarnego uzyskana przez opisaną technikę wynosi 2 – 3 kbit/s.

Technika ASR została stworzona przez firmę ATC Labs i w chwili powstawania rozprawy opisywany kodek nie był dostępny w sprzedaży. Artykuły opisujące technikę ASR nie przedstawiają również badań eksperymentalnych dla opisywanej techniki. Na tej podstawie trudno jest ustalić jakiej jakości jest rekonstruowany sygnał składowych HF.

2.9. Ślepe techniki rekonstrukcji składowych dużych częstotliwości

Przykładem techniki ślepej rekonstrukcji składowych HF jest metoda zaprezentowana na rysunku 2.16, której autorami są Larsen i Aarts [104].



Rysunek 2.16: Schemat rekonstrukcji składowych HF wykorzystujących układ nieliniowy.

Rekonstrukcja sygnału $y(t)$ składa się z następujących etapów:

- Filtr wejściowy jest filtrem pasmowo przepustowym, którego zadaniem jest ekstrakcja najwyższej oktawy sygnału $x(t)$. Jeśli przyjmiemy, że f_h jest najwyższą częstotliwością składowej w sygnale wejściowym, wówczas filtr wejściowy ma częstotliwości graniczne odpowiednio $f_h/2$ i f_h.
- Sygnał ograniczony do jednej oktawy poddawany jest przetwarzaniu w układzie nieliniowym (np. prostownik). W efekcie generowane są kolejne harmoniczne sygnału w przedziałach częstotliwości: $f_h - 2f_h$, $2f_h - 3f_h$, itd.
- Zadaniem filtru wyjściowego jest wyodrębnienie sygnału HF powyżej częstotliwości f_h.
- Następnym etapem jest skalowanie energii składowych HF. W tym miejscu używane są układy stosujące narastające tłumienie w paśmie dużych częstotliwości (np. filtr dolnoprzepustowy niskiego rzędu).
- Ostatnim elementem jest zsumowanie zrekonstruowanego sygnału dużych częstotliwości oraz $x(t)$.

Jednym z problemów podczas stosowania tego typu technik jest określenie szerokości pasma wejściowego sygnału, a na tej podstawie częstotliwości f_h oraz parametrów filtrów. Rozwiązanie tego problemu zostało zaprezentowane w pracy [5]. Detekcja częstotliwości podziału f_h polega na wyznaczeniu maksimum funkcji korelacji pomiędzy sygnałem

wejściowym $x(t)$ oraz zestawem sygnałów reprezentujących szum wąskopasmowy o zmieniającym się zakresie częstotliwości. Problem związany z poprawnym przeskalowaniem składowych HF można rozwiązać poprzez zastosowanie liniowej ekstrapolacji obwiedni widma sygnału [25, 59, 108].

Zaletą ślepych technik rekonstrukcji składowych HF jest niska złożoność algorytmu, niezależność od zastosowanego formatu sygnału oraz szerokie zastosowanie dla sygnałów mowy oraz muzyki. Stosując opisane powyżej techniki nie jesteśmy w stanie uzyskać wysokiej jakości sygnału porównywalnej z oryginałem. Układ nieliniowy może generować zniekształcenia intermodulacyjne związane z powstawaniem szeregów harmonicznych składowych HF, które nie mają swoich odpowiedników w paśmie LF. Nie jesteśmy również w stanie poprawnie odtworzyć złożonych sygnałów muzycznych bez wcześniejszej wiedzy o sygnale oryginalnym.

2.10. Podsumowanie i uzasadnienie podjęcia tematu pracy

W rozdziale 2 przedstawione zostały techniki wykorzystujące rozszerzanie widma, które umożliwiają znaczącą redukcję strumienia bitowego – nawet do 1 kb/s dla składowych HF - uzyskując dobrą jakość zdekodowanego dźwięku. Autor rozprawy zaprezentował korzyści wynikające ze stosowania technik BWE (punkt 2.5) oraz wskazał potencjalne ograniczenia aktualnie stosowanych rozwiązań (punkt 2.6).

Zastosowanie rozszerzania widma stanowi znaczący wkład w rozwój metod kodowania sygnałów fonicznych. Rozwiązania te posiadają jednak wady, które są wspólne dla większości obecnie stosowanych technik BWE, w szczególności są to:

- niska jakość sygnałów zdekodowanych dla prędkości bitowych niższych niż 40 kb/s,
- zniekształcenie wprowadzane przy rekonstrukcji składowych tonalnych sygnału.

Mając na uwadze powyższe problemy, autor rozprawy zdefiniował następujące cele badawcze:

cel 1: Znalezienie reprezentacji sygnału dla tonalnych składowych dużych częstotliwości, która pozwoli kodować sygnały dla prędkości bitowych poniżej 30 kb/s, uzyskując jakość sygnału wyższą od osiąganej w obecnie stosowanych metodach.

cel 2: Opracowanie techniki rekonstrukcji składowych dużych częstotliwości jako rozszerzenie dla już istniejących standardów kodowania BWE.

cel 3: Opracowanie nowych technik rekonstrukcji składowych dużych częstotliwości.

Analiza efektywności technik BWE oraz otrzymane wyniki eksperymentalne pozwoliły zaproponować nowe, autorskie rozwiązania w zakresie kodowania z rozszerzaniem widma, które zostaną przedstawione w następnych rozdziałach.

Rozdział 3

Metoda rozszerzania widma wykorzystująca modelowanie sinusoidalne

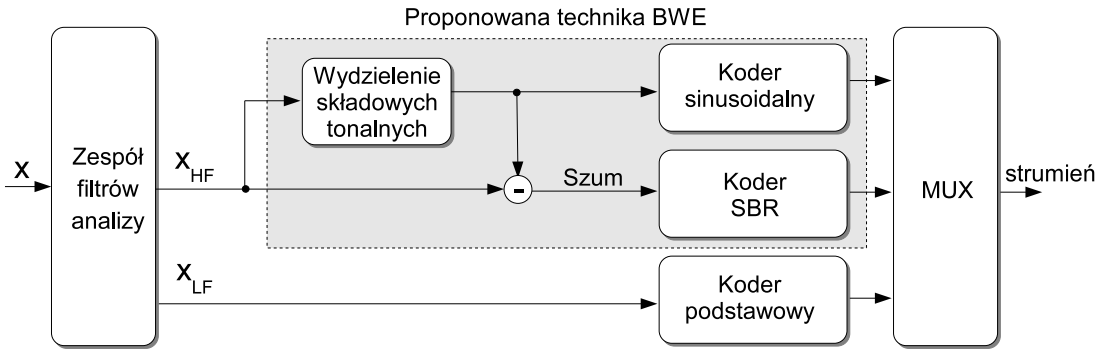
3.1. Wprowadzenie

Zastosowany w narzędziu SBR standardu MPEG-4 AAC HE algorytm replikacji widma (opisany w punkcie 2.5) powoduje powstawanie dodatkowych zniekształceń w rekonstruowanym sygnale, co zostało opisane w rozdziale 2.6. Mając na uwadze niedoskonałości techniki SBR, w rozprawie zaproponowano technikę rozszerzania widma umożliwiającą redukcję wspomnianych zniekształceń oraz poprawną rekonstrukcję tonalnej części sygnału HF.

W tym rozdziale zostanie omówiona technika rozszerzania widma wykorzystująca łącznie metody replikacji widma oraz modelowania sinusoidalnego. W rezultacie uzyskano technikę efektywnej kompresji, umożliwiającą dokładniejszą reprezentację składowych tonalnych w stosunku do MPEG-4 AAC HE. Wyniki zostaną zaprezentowane w dalszej części pracy. Technika ta stanowi oryginalny wkład autora rozprawy w rozwój metod rozszerzania widma. Część rezultatów wstępnych badań zaprezentowanych w rozdziale, została zawarta we wcześniejszych publikacjach autora rozprawy [164, 165].

3.2. Koncepcja metody

Podstawową ideą proponowanego rozwiązania jest zastosowanie odrębnych narzędzi do rekonstrukcji składowych tonalnych oraz szumu. Na tej podstawie, schemat kompresji z rozszerzaniem widma (rysunek 1.3) został uzupełniony o dodatkowy blok kodera sinusoidalnego, zgodnie z rysunkiem 3.1. Tego typu podejście jest uniwersalne i możliwe w połączeniu z dowolną techniką BWE. Dla potrzeb badań eksperymentalnych, proponowana technika została zaimplementowana jako opcjonalne rozszerzenie standardu MPEG-4 AAC HE.

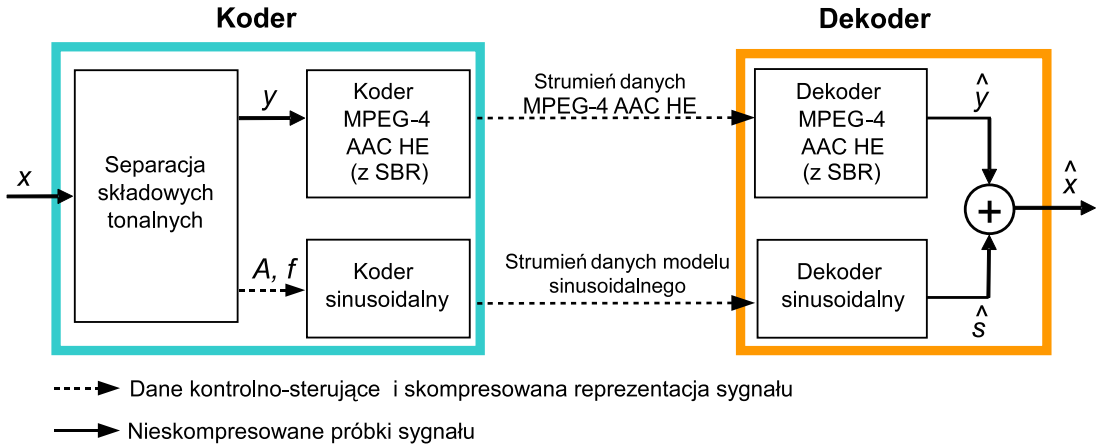


Rysunek 3.1: Rozszerzenie techniki SBR o dodatkowy blok modelowania sinusoidalnego.

W tej części rozprawy przedstawiona zostanie specjalistyczna metoda służąca kodowaniu składowych dużych częstotliwości. Celem proponowanej metody jest zachowanie poprawnych zależności pomiędzy harmonicznymi w sygnale oraz zachowanie szybkozmienne składowe sinusoidalne dużych częstotliwości. Granicę pomiędzy pasmem małych i dużych częstotliwości wyznacza częstotliwość podziału f_{SBR}, określająca równocześnie zakres częstotliwości, powyżej których wykorzystywana jest technika SBR. Sposób wykorzystania zaproponowanej techniki wraz kodekiem MPEG-4 AAC HE pokazuje rysunek 3.2.

Proponowana technika wykorzystuje proces wyboru oraz separacji składowych tonalnych. Przez separację, w tym wypadku, rozumie się wyodrębnienie oraz usunięcie składowych tonalnych HF z sygnału oryginalnego. W efekcie, składowe tonalne podlegają kompresji z zastosowaniem wyspecjalizowanego kodeka sinusoidalnego. Natomiast kodek SBR koduje sygnał y, z którego usunięto składowe tonalne HF.

W proponowanym rozwiązaniu koder wydziela z sygnału wejściowego x składowe tonalne HF. Są to składowe, dla których technika SBR często wprowadza zniekształcenia. Powstały sygnał y poddany zostaje kodowaniu techniką MPEG-4 AAC HE. Składowe LF



Rysunek 3.2: Ogólny schemat proponowanej techniki.

sygnału y kodowane są w koderze podstawowym (AAC), natomiast mające charakter szumu składowe HF kodowane są za pomocą techniki replikacji widma SBR w standardowy sposób. Składowe HF sygnału y w dalszym ciągu mogą zawierać składowe tonalne.

Odseparowane składowe tonalne przesyłane są do kodera sinusoidalnego, który koduje grupy parametrów reprezentujących chwilowe częstotliwości i amplitudy sinusoid. Formowany jest strumień binarny zawierający zakodowane parametry modelu sinusoidalnego. Ten strumień binarny stanowi poboczną informację przesyłaną do dekodera wraz ze standardowym strumieniem binarnym MPEG-4 AAC HE. Ten dodatkowy strumień binarny może być traktowany jako opcjonalne rozszerzenie, pozwalające na uzyskanie wyższej jakości zdekodowanego sygnału w stosunku do techniki MPEG-4 AAC HE.

Należy jednak zauważyć, że jeśli dekoder nie będzie przystosowany do odbioru parametrów modelu sinusoidalnego, wówczas może nastąpić obniżenie jakości zdekodowanego sygnału. Stanie się tak, ponieważ zdekodowany sygnał będzie pozbawiony składowych tonalnych HF. Natomiast sygnały o charakterze szumu będą odtwarzane prawidłowo.

3.3. Opis proponowanej techniki

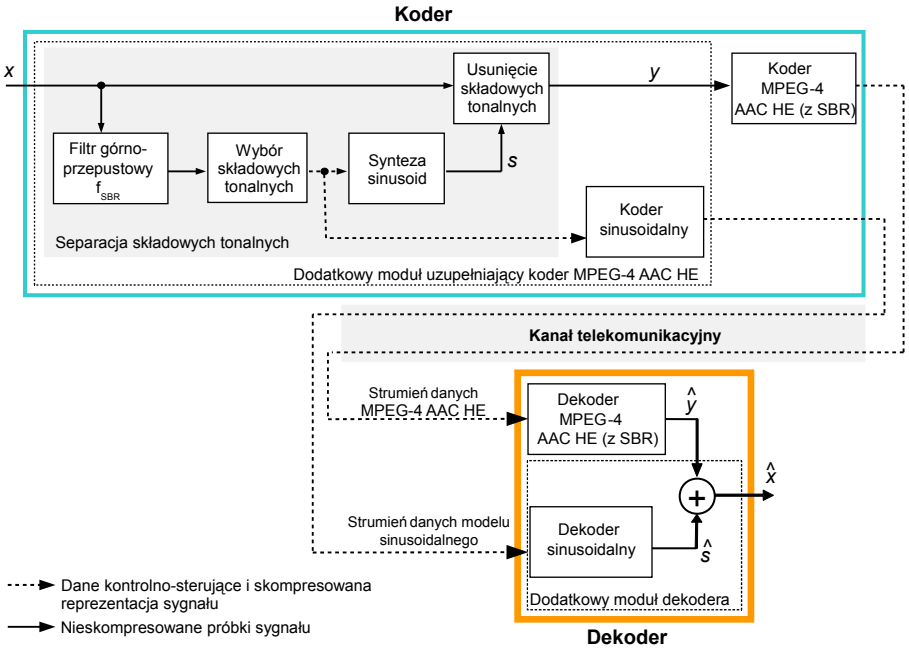
Bazując na idei rozszerzania widma oraz modelowania sinusoidalnego pokazanego na rysunkach 3.1 i 3.2, autor rozprawy skonstruował dodatkowe bloki kodeka, które następnie zostały dodane do kodeka MPEG-4 AAC HE. W kolejnych rozdziałach zostaną omówione poszczególne elementy proponowanej techniki rozszerzania widma. Szczegółowy schemat kodera i dekodera znajduje się na rysunku 3.3.

Pierwszym krokiem realizowanym poprzez proponowaną technikę rozszerzania widma (BWE) jest separacja składowych tonalnych powyżej częstotliwości granicznej f_{SBR}. Głównymi elementami separacji składowych jest wybór oraz usunięcie składowych tonalnych sygnału HF. Proces ten realizowany jest przez algorytm modelowania sinusoidalnego. Podstawy modelowania sinusoidalnego opisano w punkcie 2.2.1.

Zidentyfikowane parametry składowych sinusoidalnych, czyli chwilowe amplitudy (A) i częstotliwości (f), poddawane są kompresji i przesyłane do dekodera. Równocześnie na podstawie parametrów syntezowane są sinusoidy. Otrzymany w ten sposób sygnał s reprezentuje składowe tonalne HF. Dokładność reprezentacji wynika ze szczegółowości zastosowanej analizy sinusoidalnej.

Wyjściem bloku separacji składowych tonalnych jest sygnał y, który powstaje w wyniku usunięcia składowych sinusoidalnych dużych częstotliwości z sygnału wejściowego x. Sygnał y poddawany jest kodowaniu w bloku kodera MPEG-4 AAC HE.

Uzyskane poprzez analizę sinusoidalną (wybór składowych tonalnych) wartości amplitudy i częstotliwości wykorzystane są w dalszym procesie przetwarzania w celu:



Rysunek 3.3: Schemat zmodyfikowanego kodeka MPEG-4 AAC HE uwzględniający dodatkowe bloki modelowania sinusoidalnego. Dodatkowe bloki kodera i dekodera służą kodowaniu tonalnych składowych HF.

- resyntezy w koderze sygnału s, na podstawie którego usuwane są składowe tonalne z sygnału wejściowego x,
- kompresji w koderze sinusoidalnym, zgodnym z algorytmem przedstawionym w dalszej części pracy (dodatek B).

Z kodera do dekodera przesyłany jest klasyczny strumień danych MPEG-4 AAC HE oraz dodatkowy strumień bitowy parametrów modelu sinusoidalnego. Sygnał wyjściowy dekodera $\hat{x}$ jest sumą sygnału $\hat{m}$ uzyskanego na wyjściu dekodera MPEG-4 AAC HE oraz sygnału $\hat{s}$ składowych tonalnych HF, odtworzonego na podstawie przesłanych parametrów modelu sinusoidalnego.

Powyżej zaprezentowano ogólną koncepcję kodeka, który odtwarza harmoniczne dużych częstotliwości wierniej niż kodek MPEG-4 AAC HE. Eksperymentalne potwierdzenie tej tezy zostało opisane w rozdziale 5. W dalszych punktach rozdziału 3 zostanie opisana konstrukcja kodeka implementująca powyżej opisaną ideę. Prawdopodobnie istnieją inne warianty tej konstrukcji. Wybór elementów konstrukcji przedstawionej w pracy wynika z doświadczenia autora rozprawy zdobytego podczas konstrukcji i implementacji kodeków sygnałów fonicznych oraz studiów literaturowych.

3.4. Separacja składowych tonalnych

Istotnym elementem proponowanej techniki jest proces wydzielenia tonalnych składowych dużych częstotliwości z sygnału wejściowego x. W wyniku separacji otrzymujemy sygnał y, pozbawiony składowych tonalnych HF, oraz zbiór parametrów opisujących składowe sinusoidalne, czyli amplitudy i częstotliwości.

Sygnał y poddawany jest kompresji w koderze MPEG-4 AAC HE. W ten sposób za pomocą techniki SBR koduje się sygnał HF, który po usunięciu sinusoid ma charakter szumu. Dla tego typu sygnałów mechanizm replikacji widma (SBR) pozwala na efektywną kompresję oraz zapewnia wysoką jakość zrekonstruowanego sygnału HF. Dodatkowo, dokładna separacja składowych tonalnych pozwala uniknąć powstawaniu w koderze SBR zniekształceń wyrażających się zwłaszcza niepoprawną reprezentacją struktury harmonicznej sygnału oraz szybkozmiennych składowych. Problem ten został opisany w punkcie 2.6.

W celu osiągnięcia wysokiej skuteczności separacji składowych tonalnych w proponowanej technice zastosowano algorytm, który składa się z następujących etapów:

- Wybór składowych tonalnych - w oparciu o analizę składowych sinusoidalnych;
- Synteza składowych sinusoidalnych;
- Usuwanie składowych tonalnych - celem jest usunięcie składowych tonalnych z sygnału wejściowego x.

W proponowanej technice ciężar poprawnej rekonstrukcji tonalnych składowych HF został przeniesiony na stronę modelowania i kodowania sinusoidalnego.

3.4.1. Wejściowy filtr górnoprzepustowy

Pierwszym elementem bloku separacji składowych tonalnych jest filtr górnoprzepustowy o częstotliwości granicznej równej f_{SBR}. Służy on wyodrębnieniu z sygnału wejściowego x składowych dużych częstotliwości, które w standardowym kodeku MPEG-4 AAC HE przetwarzane są przez algorytm SBR. Częstotliwość podziału pasma f_{SBR} jest przesyłana w strumieniu binarnym generowanym przez koder MPEG-4 AAC HE i jest równa wielokrotności szerokości subpasm uzyskiwanych w zespołach QMF (rozdział 2.5). Wartość częstotliwości f_{SBR} nie jest ujęta w opisie standardu, jednak większość obecnych implementacji [2, 116] kodera MPEG-4 AAC HE stosuje wartości podane w tabeli 2.1. Dodatkowo częstotliwość f_{SBR} we wspomnianych implementacjach jest stała dla założonej prędkości bitowej strumienia binarnego.

Zakres częstotliwości rekonstruowanych przez technikę SBR zawiera się w przedziale od f_{SBR} do $f_{SBR_{max}}$. Sposób replikacji widma w algorytmie SBR, został opisany w normie MPEG-4 AAC HE [68] oraz rozdziale 2.5. Podczas badań eksperymentalnych zastosowano wartości częstotliwości granicznej filtru górnoprzepustowego zgodnie z tabelą 2.1.

Proponowana technika rozszerzania widma jest w stanie generować szersze pasmo w stosunku do MPEG-4 AAC HE, nawet powyżej częstotliwości $f_{SBR_{max}}$. Potwierdzają to wyniki eksperymentów przedstawione w dalszej części pracy.

3.4.2. Wybór składowych tonalnych do separacji

Poprawna detekcja amplitudy i częstotliwości sinusoid jest istotna dla rekonstrukcji struktury harmonicznej sygnału, co zostało opisane w rozdziale 2.3. Wejściem bloku wyboru składowych tonalnych jest sygnał dużych częstotliwości x_{HF} (rysunek 3.2). Natomiast

na wyjściu otrzymujemy zbiór parametrów reprezentujących zmienne w czasie składowe tonalne.

W literaturze opisano niewiele rozwiązań umożliwiających poprawną klasyfikację składowych tonalnych. Bazują one głównie na analizie chwilowego widma amplitudowego i fazowego [53, 133, 150]. Stosowane są również rozwiązania analizujące przebieg trajektorii sinusoidalnych w dziedzinie czasu i częstotliwości [96].

W opisywanej technice rozszerzania widma wykorzystano metodę wyboru składowych tonalnych dla której można wyodrębnić następujące etapy:

- **I etap** – detekcja wewnątrz pojedynczego okna analizy. Wykorzystanie algorytmu podziału lokalnych szczytów widma, na składowe tonalne i szumowe [133]. Detekcja bazuje na deskryptorze sygnału analizującego lokalny rozkład widma. W efekcie otrzymujemy parametry tonalności dla każdego szczytu widma amplitudowego. Szczyty widma amplitudowego, które nie zostaną zakwalifikowane jako sinusoidy zostają odrzucone z dalszego procesu.

- **II etap** – analiza trajektorii sinusoidalnych. W tym celu wykorzystano algorytm śledzenia trajektorii sinusoidalnych z parametrami opisanymi w dalszej części rozdziału.

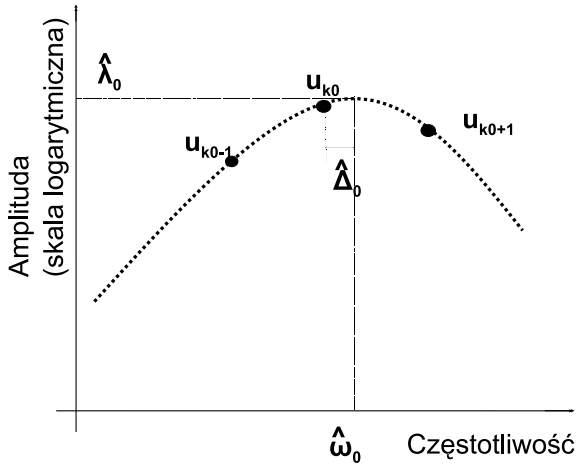
Dla celów kodowania, jako składowe tonalne zostają sklasyfikowane te trajektorie sinusoidalne, które mają największą energię całkowitą oraz najdłuższy czas trwania. Sposób doboru kryteriów zostanie opisany w kolejnych punktach.

Etap I – Detekcja składowych tonalnych w pojedynczym oknie analizy

W rozdziale 2.2.1 pokazano, że można wyznaczyć parametry składowych tonalnych przypisując im wartości amplitudy i częstotliwości lokalnych maksimów w widmie amplitudowym. Tego typu metoda nie jest jednak precyzyjna ze względu na skończoną rozdzielczość przekształcenia DFT. Dlatego w pracy zastosowano algorytm QIFFT (ang. *Quadratically interpolated FFT*) [4, 13, 146], który za pomocą interpolacji parabolicznej sąsiednich prążków FFT wyznacza parametry tonu sinusoidalnego (rysunek 3.4).

Opis wyznaczania parametrów sinusoid w chwilowym widmie amplitudowym na podstawie algorytmu QIFFT:

1. Wyznaczenie widma amplitudowego (punkty na rysunku 3.4).



Rysunek 3.4: Paraboliczna interpolacja prążków widma – QIFFT.

2. Wyznaczenie lokalnych maksimów amplitudy (u_{k0}).

3. Interpolacja paraboliczna amplitudy w dziedzinie logarytmicznej, wykorzystując lokalne maksimum oraz sąsiednie prążki widma (przerywana linia).

4. Wyznaczenie częstotliwości i amplitudy na podstawie interpolacji parabolicznej ($\hat{\omega}_0$ i $\hat{\lambda}_0$).

5. Usunięcie prążka DFT reprezentującego lokalne maksimum widma.

6. Powtórzenie algorytmu dla każdego znalezionego maksimum widma amplitudowego.

Dokładność estymacji parametrów sinusoid w pewnym stopniu zależy również od rozdzielczości analizy DFT. Zwiększając długość okna uzyskamy wyższą rozdzielczość DFT oraz wyższą dokładność estymowanych parametrów. Wysoka rozdzielczość analizy widma ma znaczenie dla estymacji parametrów składowych małych częstotliwości. W przypadku proponowanego rozwiązania składowe LF nie są analizowane. Natomiast zastosowanie algorytmu QIFFT w wystarczającym stopniu niweluje błędy estymacji częstotliwości i amplitudy sinusoid wynikające ze skończonej rozdzielczości DFT.

W zastosowanym algorytmie modelowania sinusoidalnego wykorzystano algorytm klasyfikacji składowych tonalnych i szumowych na podstawie rozkładu prążków widma amplitudowego [133]. Zgodnie z zaleceniami autorów artykułu zastosowano deskryptor znormalizowanego pasma (ang. *Normalized Bandwidth Descriptor*) (NBD) [133]. Deskryptor NBD jest miarą energii szumu w okolicy lokalnego maksimum widma. Jeśli przez $\bar{\omega}$ oznaczymy średnią częstotliwość, przez B szerokość pasma dla pojedynczego szczytu widma

amplitudowego, przez L liczbę próbek dla pasma B, wówczas deskryptor NBD można zapisać jako:

$$\bar{\omega} = \frac{\sum_k k\,|X(k)|^2}{\sum_k |X(k)|^2}, \tag{3.1}$$

$$NBD = \frac{B}{L} = \frac{\sum_k (k-\bar{\omega})^2 |X(k)^2|}{L \sum_k |X(k)^2|}, \tag{3.2}$$

gdzie operacja sumowania przeprowadzana jest po wszystkich współczynnikach DFT, k jest indeksem prążka DFT, $X(k)$ jest wartością próbki DFT dla k-tego prążka analizowanego sygnału x. Autorzy pracy[133] zbadali rozkład deskryptora NBD dla sinusoid oraz szumu. Na tej podstawie można skonstruować miarę tonalności bazującą na NBD, przy czym $NBD \in < 0, 1 >$ - wartość 0 oznacza szum, natomiast 1 składnik tonalny.

W proponowanej technice zastosowano mechanizm progowania wartości deskryptora NBD w celu odrzucenia części wykrytych maksimów widma amplitudowego. Wybrany próg tonalności dla lokalnych szczytów widma T_{thr} wynosi $0,5$. Wartości T_{thr} nie można progować w bliskim zakresie 1 ($T_{thr} = 1$ oznacza składnik tonalny/sinusoidę), ponieważ należy zachować pewien margines błędu dla detekcji. Zwłaszcza składowe, które gwałtownie zmieniają swoją częstotliwość, mogą być na podstawie takiego badania zaklasyfikowane jako szum (ich lokalne wimo amplitudowe nie ma wyraźnego maksimum). Wartość parametru T_{thr} została dobrana na podstawie rozkładu wartości deskryptora NBD dla sinusoid i szumu (opisanego przez autorów pracy[133]). Dobór progu tonalności stanowi kompromis pomiędzy dokładnością wyznaczonego rozkładu NBD, a liczbą wykrywanych sinusoid. Dodatkowo wartość progu T_{thr} została zweryfikowana na podstawie nieformalnych testów.

Etap II – selekcja składowych tonalnych na podstawie długości trajektorii sinusoidalnej

Metoda opisana w poprzednim punkcie jest niewystarczająca dla tonów sinusoidalnych, które zmieniają swoje parametry w obrębie sąsiadujących ze sobą okien analizy. Z tego powodu zastosowano algorytm śledzenia trajektorii sinusoidalnych, którego celem jest łączenie parametrów lokalnych maksimów widma DFT w sąsiednich oknach analizy. Połączone w ten sposób parametry tworzą trajektorię sinusoidalną (dodatek A).

Istotnym etapem wyboru składowych tonalnych jest analiza długości wyznaczonych trajektorii sinusoidalnych. Trajektorie, których łączna długość nie przekracza $L_{trj,thr} <$ 5 ramek (około 46 ms) usuwane są z dalszego procesu analizy. Trajektorie o tej długości mogą być częścią transjentu lub składowej szumowej, która zakłóca działanie algorytmu

śledzącego. Krótkie trajektorie (do kilkudziesięci ms), które gwałtownie pojawiają się i gasną, dodatkowo zmieniające w sposób przypadkowy swoją częstotliwość chwilową mogą generować efekt, który w literaturze opisywany jest jako *szum muzyczny* [38]. Dlatego istotnym kryterium wyboru składowych tonalnych jest długość trajektorii sinusoidalnej, która w proponowanej technice BWE nie powinna być mniejsza niż $L_{trj,thr}$. Wartość ta została ustalona podczas nieformalnych badań eksperymentalnych.

Na wyjściu bloku wyboru składowych tonalnych otrzymujemy zbiór parametrów reprezentujących składowe tonalne P.

3.4.3. Usuwanie składowych tonalnych

Blok usuwania składowych tonalnych ma za zadanie wydzielić składowe tonalne dużych częstotliwości z sygnału wejściowego x. Element ten może zostać zrealizowany na wiele sposobów, np. jako różnica sygnałów x i s, odejmowanie widma (ang. *spectral subtraction*) [19] lub zastosowane w pracy tłumienie widma (ang. *spectral attenuation*) [26, 106]. Na wyjściu układu usuwania składowych tonalnych otrzymujemy sygnał y, którego widmo składowych HF ma charakter szumu.

Zastosowany podczas implementacji proponowanej techniki BWE, model sinusoidalny, nie zawiera informacji o parametrze fazy sinusoid (punkt 2.3.3). Z tego powodu autor rozprawy zdecydował się na zastosowanie tłumienia widmowego. W algorytmie tym przekształceniom podlega moduł widma amplitudowego, natomiast informacja o fazie sygnałów x i s jest ignorowana. Dokładny opis algorytmu znajduje się w dalszej części bieżącego rozdziału.

Proces usuwania składowych tonalnych realizowany jest w kolejno następujących po sobie oknach analizy, których długość jest taka sama jak w analizie sinusoidalnej i wynosi $N = 2048$ próbek. Zastosowane jest przesunięcie okien analizy (skok) o wartości $H = 512$ próbek. Do wyznaczenia wartości próbek w kolejnych oknach posłużono się dyskretną analizą krótkookresową STFT [149] opisaną wzorem (3.3).

$$X_m(k) = \sum_{n=0}^{N-1} x\left(n - mH\right) \cdot w(n) \cdot \exp\left(-j2\pi \frac{kn}{N}\right), \tag{3.3}$$

gdzie:

$X_m(k)$ – widmo STFT sygnału,

m – indeks okna analizy,

N – liczba próbek okna analizy,

H – przesunięcie pomiędzy oknami analizy,

n – numer próbki w oknie, $n = 0, \ldots, N-1$,

k – indeks prążka DFT, $n = 0, \ldots, N-1$,

w – funkcja okna czasowego,

x – analizowany sygnał wejściowy.

W procesie analizy STFT zastosowano kosinusoidalne okno czasowe, które pełni rolę ograniczającą skoki wartości sygnału na granicach bloku poprzez łagodne zmniejszenie ich amplitudy do zera. Funkcja okna zdefiniowana jest jako:

$$w(n) = \cos\left(\frac{\pi n}{N-1} - \frac{\pi}{2}\right) = \sin\left(\frac{\pi n}{N-1}\right) \tag{3.4}$$

W celu usunięcia składowych tonalnych w pracy zastosowano zmodyfikowaną metodę krótkookresowego tłumienia widma (ang. *Short-Time Spectral Attenuation*) (STSA), którego idea pochodzi z prac Ephraima i Malaha dotyczących redukowania szumu dla sygnałów mowy [38, 39].

Tłumienie widma

W ogólnym założeniu algorytm STSA operuje na danych wyznaczonych w drodze transformacji STFT. W kolejnych oknach analizy wyznaczana jest charakterystyka sygnału zakłócającego z (szumu), a następnie sygnał wynikowy x, poddany zakłóceniu zostaje stłumiony w miejscu występowania zakłócenia. Jeśli przez y oznaczymy sygnał użyteczny, wówczas otrzymamy:

$$x = y + z. \tag{3.5}$$

Stosując przekształcenie STFT otrzymamy:

$$X_m(k) = Y_m(k) + Z_m(k), \tag{3.6}$$

gdzie $Y_m(k)$ jest widmem sygnału użytecznego dla chwili czasu m oraz indeksu częstotliwości k.

Aby uzyskać sygnał użyteczny $Y_m(k)$ należy przekształcić wzór (3.6) do postaci:

$$Y_m(k) = X_m(k) - Z_m(k) = \left(1 - \frac{Z_m(k)}{X_m(k)}\right) \cdot X_m(k), \tag{3.7}$$

co można zapisać jako:

$$Y_m(k) = G_m(k) \cdot X_m(k), \tag{3.8}$$

gdzie $G(k)$ jest transmitancją układu redukcji zakłócenia (szumu). Transmitancję $G(k)$ będziemy również określać jako funkcję tłumiącą.

Na podstawie wzoru (3.8) można zrealizować operację tłumienia składowych tonalnych w wejściowym sygnale x. W literaturze dotyczącej układów redukcji szumów opisane są różne techniki wyznaczania funkcji tłumiącej $G(k)$, np. tzw. reguła energetyczna [106] lub reguła Wienera [112]. W tym przypadku sygnałem, który traktujemy jak zakłócenie jest sygnał s utworzony na etapie syntezy parametrów modelu sinusoidalnego. $S_m(k)$ reprezentuje widmo składowych tonalnych dużych częstotliwości sygnału wejściowego x.

Mając dane widmo amplitudowe sygnału s można skonstruować funkcję tłumiącą:

$$G_m(k) = \frac{1}{|S_m(k)|}, \tag{3.9}$$

gdzie k to indeks współczynnika DFT, $S_m(k)$ to widmo sygnału zakłócenia (składowe tonalne HF). Parametry filtru $G_m(k)$ wyznaczane są osobno dla każdego okna analizy.

Na podstawie wzorów (3.8) i (3.8) można wyznaczyć sygnał y, który jednocześnie stanowi wynik operacji w bloku usuwania składowych tonalnych.

Sygnał tłumiący s tworzą pojedyncze sinusoidy, które w krótkookresowym widmie amplitudowym $|S_m(k)|$ reprezentowane są przez prążki DFT o dużej amplitudzie w stosunku do otoczenia. Bezpośrednie zastosowanie widma sygnału $S_m(k)$ jako funkcji tłumiącej spowoduje silne stłumienie bardzo wąskiego zakresu pasma częstotliwości. Jest to zabieg tożsamy z zastosowaniem filtru pasmowo-zaporowego o bardzo wąskim paśmie i stromych zboczach. Tego typu filtracja charakteryzuje się niepożądanym bardzo długim czasem odpowiedzi impulsowej. Aby zapobiec powstawaniu tego typu artefaktom, w rozprawie zaproponowano rozwiązanie ze zmodyfikowaną funkcją tłumiącą o następującej postaci:

$$G_m(k) = \frac{1}{\min\left(\frac{|S_m(k)|}{\epsilon}, 1\right) * h(k)}, \tag{3.10}$$

gdzie:

$Y(k)$ – widmo sygnału ze stłumionymi składowymi,

$S(k)$ – widmo sygnału tłumiącego,

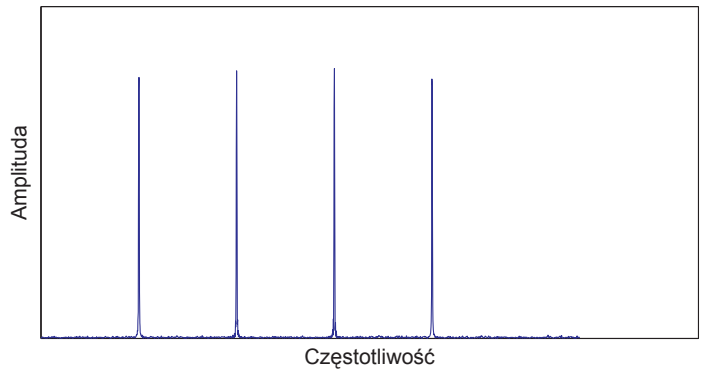
ϵ – współczynnik tłumienia,

$h(k)$ – filtr o skończonej odpowiedzi impulsowej (FIR) operujący na prążkach DFT, wygładza widmo sygnału s, opisany wzorem:

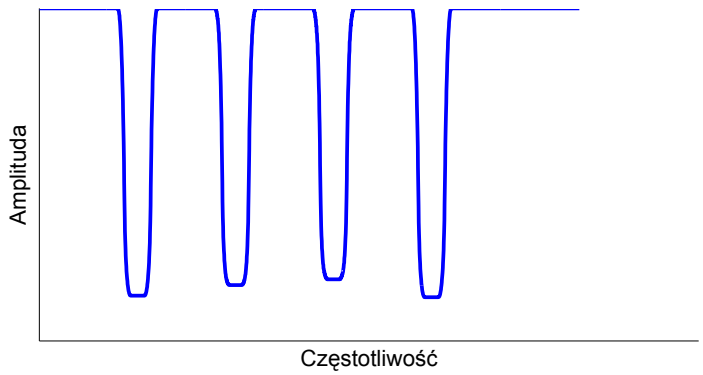
$$h(k) = \begin{cases} 1, & |k| \leqslant K \\ 0, & |k| > K \end{cases}, \tag{3.11}$$

gdzie $K = 4$, został dobrany na podstawie nieformalnych testów. Parametr K jest wynikiem kompromisu pomiędzy rozdzielczością DFT (2048 punktów), a szerokością pasma reprezentowaną przez składową tonalną.

Przykład tłumienia widma dla pojedynczego okna analizy przedstawiono na rysunku 3.5. Widmo sygnału tłumiącego $S(k)$ (rysunek 3.5a), poddano operacji wygładzania oraz przekształcono według wzoru (3.10), a wynik zaprezentowano na rysunku 3.5b. Należy podkreślić, że tłumienie sygnału x następuje tylko w miejscach występowania składowych tonalnych. Pozostała część widma nie jest modyfikowana.



(a) Widmo amplitudowe sygnału tłumiącego



(b) Charakterystyka częstotliwościowa filtru $H(k)$

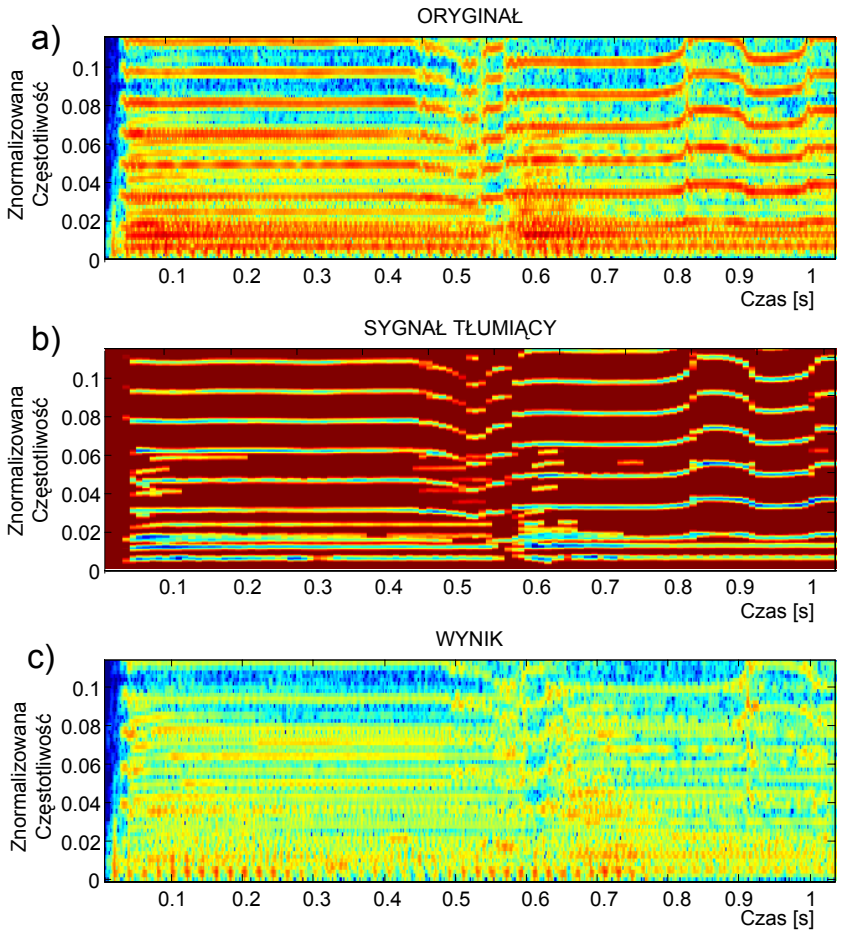
Rysunek 3.5: Rysunek poglądowy przedstawiający zastosowanie tłumienia widma.

W pracy zastosowano współczynnik $\epsilon = 0.01$. Parametr ϵ dobrano jako kompromis pomiędzy skutecznością tłumienia, a ewentualnymi zniekształceniami spowodowanymi usunięciem fragmentu składowych widma. Natomiast funkcja min ze wzoru (3.10) zabezpiecza przed nadmiernym przypadkowym wzmocnieniem sygnału oraz dzieleniem przez zero.

Wynikiem operacji tłumienia składowych tonalnych jest sygnał y, który następnie poddany zostanie kodowaniu w MPEG-4 AAC HE. Należy zauważyć, że w widmie składowych HF sygnału y pozostaną głównie składniki szumowe, które z percepcyjnego punktu widzenia są mniej istotne, niż składniki tonalne. Składniki szumowe zostaną również bardzo efektywnie zakodowane przez koder SBR poprzez replikację widma. Spowoduje to, że koder SBR nie prześle w strumieniu dodatkowych informacji związanych z kodowaniem składowych tonalnych (współczynniki skalujące widmo oraz parametry opisujące generację syntetycznych sinusoid). W ten sposób zmniejsza się prędkość bitowa strumienia binarnego kodera SBR, a zaoszczędzone bity są przeznaczone na kompresję parametrów modelu sinusoidalnego. Przykład tłumienia widma dla krótkiego fragmentu nagrania przedstawiony jest na rysunku 3.6.

Sygnał s jest resyntezowany na podstawie wszystkich składowych tonalnych wykrytych w drodze modelowania sinusoidalnego. Natomiast, w dalszym procesie kodowana uwzględnione zostaną tylko trajektorie sinusoidalne percepcyjnie najistotniejsze. Kryterium wyboru trajektorii sinusoidalnych stanowi część strategii kodowania i zostanie zaprezentowane w rozdział 3.9.4.

Tłumienie widma stanowi bardzo ważny element proponowanej techniki BWE. Razem z blokiem wyboru składowych tonalnych, decyduje o dokładności separacji składowych sygnału, a tym samym o skuteczności działania proponowanej techniki BWE.

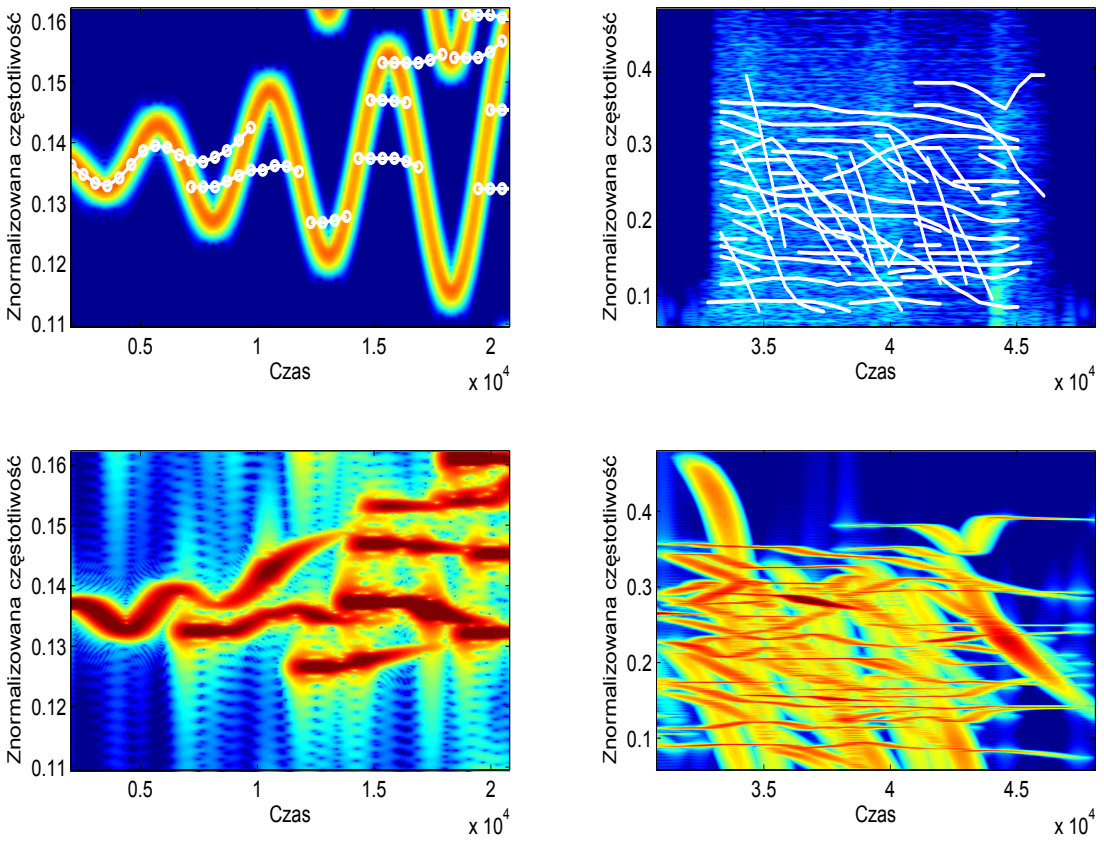


Rysunek 3.6: Przykład zastosowania tłumienia widma. a) spektrogram sygnału oryginalnego x, b) spektrogram sygnału tłumiącego s - tylko składowe tonalne, c) wynik tłumienia widma y - sygnał ma charakter szumowy.

3.5. Ograniczenia separacji składowych tonalnych

Jednym z podstawowych problemów separacji składników tonalnych jest ich niezawodna klasyfikacja. Badania eksperymentalne przeprowadzone przez autora rozprawy wykazały, że trudno jest znaleźć algorytm, który jednoznacznie wskaże składowe istotne dla syntezy i rekonstrukcji sygnału $\hat{x}$. Jeśli niektóre składniki szumowe zostaną sklasyfikowane jako tonalne, to wówczas zostaną one usunięte w procesie tłumienia widma. W tej sytuacji dekoder SBR nie będzie w stanie odtworzyć składowych szumowych. Efekt ten jest szczególnie dokuczliwy i słyszalny dla fragmentów muzyki, w której występują instrumenty perkusyjne - rysunek 3.7b.

Również dźwięki, które gwałtownie zmieniają swoją częstotliwość mogą nie zostać poprawnie wykryte. Spowodowane to jest ograniczoną rozdzielczością czasowo-częstotliwościową analizy STFT. W efekcie prowadzi to do błędnej rekonstrukcji sygnału - rysunek 3.7a.



(a) (b)

Rysunek 3.7: Ilustracja problemu błędnej rekonstrukcji składowych tonalnych. U góry przedstawione są spektrogramy sygnałów oryginalnych oraz trajektorie sinusoidalne (biała linia). Na dole przedstawione są spektrogramy sygnałów zsyntezowanych na podstawie parametrów modelu sinusoidalnego. a) szybkozmienne składniki tonalne nie zostały poprawnie zidentyfikowane przez model sinusoidalny, b) spektrogramy przedstawiają transjent szumowy - model sinusoidalny zaklasyfikował szum jako składowe tonalne.

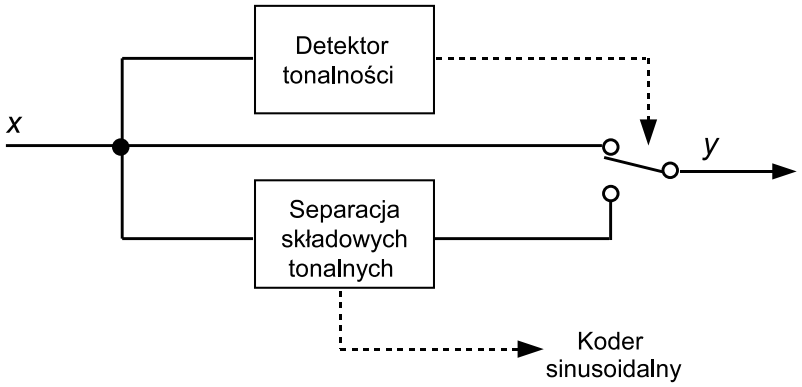
Nie wykrycie składnika tonalnego, a w konsekwencji pozostawienie go w sygnale, może spowodować jego niepoprawną rekonstrukcję przez dekoder SBR. Efekt ten został opisany w poprzednim rozdziale, został również pokazany na rysunku 3.7.

W celu zminimalizowania efektu błędnej decyzji w procesie wyboru składowych tonalnych, w pracy zaproponowano zastosowanie detektora tonalności, opisanego w rozdziale 3.6.

3.6. Detektor tonalności

Modelowanie sinusoidalne w połączeniu z algorytmem śledzenia trajektorii sinusoidalnych może stanowić efektywną metodę wyboru składowych tonalnych. Jednak w pewnych sytuacjach, tego typu algorytm nie jest w stanie poprawnie sklasyfikować składowych tonalnych i szumowych. (Problem został dokładniej opisany w rozdziale 3.5). Dodatkowo modelowanie sinusoidalne jest procesem czasochłonnym, a złożoność algorytmu rośnie wprost proporcjonalnie do liczby składowych sygnału.

W celu zniwelowania problemów związanych z błędną klasyfikacją składowych sygnału, w rozprawie zaproponowano układ detektora tonalności. Detektor tonalności podejmuje decyzję, czy dla danego bloku danych należy przeprowadzić proces wyboru składowych tonalnych sygnału. Jeśli sygnał ma charakter szumowy wówczas blok separacji składowych tonalnych nie jest wykorzystywany. Sposób włączenia detektora tonalności do proponowanej techniki został przedstawiony na rysunku 3.8.



Rysunek 3.8: Schemat umieszczenia dodatkowego modułu detektora tonalności oraz przełącznika wyboru pracy proponowanej techniki BWE.

Detektor tonalności ma za zadanie określić, jaki jest udział składowych tonalnych w danej części widma amplitudowego. Jeśli udział składowych tonalnych jest zbyt niski, wówczas ten fragment widma traktowany jest jako szum i nie wykonywana jest analiza sinusoidalna. Podejście to pozwala zmniejszyć błędy powstające w wyniku niewłaściwej klasyfikacji składowych widma. Jednocześnie ograniczona zostaje złożoność obliczeniowa algorytmu, ponieważ rezygnujemy z czasochłonnej analizy sinusoidalnej dla danego fragmentu sygnału.

W rozprawie wykorzystano detektor tonalności oparty na mierze płaskości widma SFM [94, 125] (ang. *spectral flatness measure*). Współczynnik SFM wyznaczono na podstawie poniższego wzoru:

$$SFM = \frac{\sqrt[N]{\prod_{n=0}^{N-1} x(n)}}{\frac{1}{N}\sum_{n=0}^{N-1} x(n)}. \tag{3.12}$$

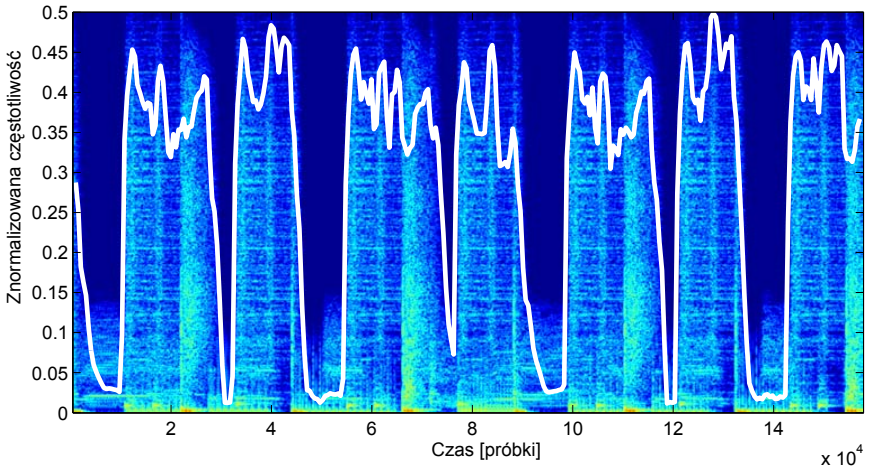
Na podstawie współczynnika SFM wyznacza się współczynnik tonalności T_m:

$$SFM_{dB} = 10 \cdot \log_{10}(SFM), \tag{3.13}$$

$$T_m = \min\left(\frac{SFM_{dB}}{-60}, 1\right), \tag{3.14}$$

gdzie m to indeks okna analizy.

Współczynnik tonalności T_m mieści się w przedziale $< 0, 1 >$, gdzie 0 oznacza szum, a 1 widmo całkowicie tonalne. Na podstawie badań eksperymentalnych w pracy przyjęto próg tonalności $T_{m,thr} = 0.4$, poniżej którego widmo klasyfikowane jest jako szum. Przykład działania detektora tonalności przedstawiono na rysunku 3.9.



Rysunek 3.9: Spektrogram sygnału. Białą linią zaznaczona wartość współczynnika tonalności dla danej chwili czasu. Dla potrzeb rysunku wartość współczynnika tonalności została przeskalowana – podzielona przez 2.

3.7. Strategia kodowania parametrów modelu sinusoidalnego

3.7.1. Wstęp

Dla potrzeb analizy oraz kodowania składowe tonalne są reprezentowane przez trajektorie sinusoidalne. Pojedyncza trajektoria sinusoidalna przedstawia sekwencję amplitud i częstotliwości opisujących zmiany składowej tonalnej, w określonym odcinku czasu, wyznaczonych przez środki sąsiednich okien analizy. Zbiór danych opisujący punkty wszystkich trajektorii sinusoidalnych w tej samej chwili czasu nazywamy *ramką*.

W algorytmie kompresji parametrów modelu sinusoidalnego, możemy wyodrębnić następujące części:

- kodowanie parmaetró trajektorii sinusoidalnych (A, f) z zastosowaniem predykcji liniowej oraz kodowania różnicowego. W pętli predykcji kwantyzacji podlega błąd predykcji,
- kodowanie entropijne skwantowanych wartości.

Dokładny opis kompresji parametrów modelu sinusoidalnego znajduje się w dodatku B.4.

W dalszej części rozdziału omówiona zostanie strategia kodowania parametrów modelu sinusoidalnego, której celem jest uzyskanie określonej, docelowej prędkości bitowej strumienia binarnego proponowanej techniki BWE.

3.7.2. Sterowanie pracą kodera

Wykryte składowe zostają usunięte z sygnału wejściowego x, a następnie w postaci parametrów opisujących trajektorie są przesłane do dekodera (rysunek 3.3). Jednak kodowanie wszystkich trajektorii sinusoidalnych, w paśmie składowych HF jest nieefektywne. Część składowych ma niewielkie percepcyjne znaczenie, np. energię poniżej progu słyszenia. Można więc pominąć je w procesie kodowania, uzyskując mniejszą prędkość bitową strumienia.

Z tego powodu, zaproponowano rozszerzenie „kodera sinusoidalnego” (rysunki 3.2 i 3.3) o pętlę sterowania przepływnością, której zadaniem jest ograniczenie liczby trajektorii sinusoidalnych. Koder sinusoidalny dokonuje selekcji trajektorii podczas iteracyjnej procedury, w kolejnych krokach pętli sterownia przepływnością. Kryterium selekcji i ograniczenia liczby trajektorii zostanie przedstawione w rozdziale 3.9.4. W tym rozdziale zostanie zaprezentowana ogólna idea sterowania pracą kodera sinusoidalnego.

Podstawowy problem badawczy, jaki należało rozwiązać, opracowując koder składowych tonalnych, to znalezienie odpowiedzi na pytanie ile i które trajektorie przesłać do dekodera, tak aby zapewnić poprawną rekonstrukcję sygnału? W tym celu przeprowadzono serię eksperymentów, na podstawie których oszacowano wstępne parametry pętli sterowania przepływnością, tj. kwantyzację parametrów sinusoid oraz liczbę sinusoid. Wyniki badań przedstawiono w rozdziale 3.9.2. Dysponując wartościami początkowymi parametrów oraz mechanizmem selekcji składowych tonalnych, można określić docelową prędkość bitową strumienia.

Zgodnie z rysunkiem 3.3, proponowana technika generuje dwa strumienie binarne, zawierające dane modelu sinusoidalnego oraz kodera MPEG-4 AAC HE. Docelowa prędkość bitowa proponowanej techniki BWE B_C, jest sumą prędkości bitowej B_{AAC-HE} strumienia generowanego przez technikę MPEG-4 AAC HE oraz prędkości bitowej B_{Sin} strumienia binarnego, zawierającego parametry modelu sinusoidalnego.

$$B_C = B_{AAC-HE} + B_{Sin}. \tag{3.15}$$

Koder MPEG-4 AAC HE zgodnie z tabelą 2.1, przełącza zakres stosowania techniki rozszerzania widma w zależności od docelowej prędkości bitowej. Oznacza to, że zwiększając strumień binarny generowany przez parametry modelu sinusoidalnego, zmniejszamy wartość częstotliwości podziału f_{SBR}. W konsekwencji koder podstawowy AAC rekonstruuje węższe pasmo sygnału.

Tworząc mechanizm sterownia przydziałem bitów, zaproponowano dwa warianty pracy kodera, w których:

- nacisk położono na możliwie wierną rekonstrukcję składowych tonalnych HF za pomocą modelu sinusoidalnego. W tym celu przeznacza się większą pulę bitów na reprezentację parametrów modelu sinusoidalnego B_{Sin}. W konsekwencji może dojść

do zmniejszenia wartości f_{SBR}, przez co poszerza się zakres pasma częstotliwości na którym operuje technika BWE.

- początkowa wartość częstotliwości f_{SBR} nie może ulec zmianie. Osiągnąć to można, ograniczając pulę bitów przeznaczonych do zakodowania trajektorii sinusoidalnych, do wartości B_{Sin}. Może to prowadzić do odrzucenia części trajektorii, a w konsekwencji niewłaściwą reprezentację składowych tonalnych HF. Z drugiej strony, koder podstawowy AAC operuje na maksymalnym, dopuszczalnym przez koder MPEG-4 AAC HE, zakresie częstotliwości pasma LF.

Poniżej opisane zostaną dwa warianty układu sterowania pracą kodera. W każdym przypadku przetwarzane są dane pochodzące z bloku analizy sinusoidalnej.

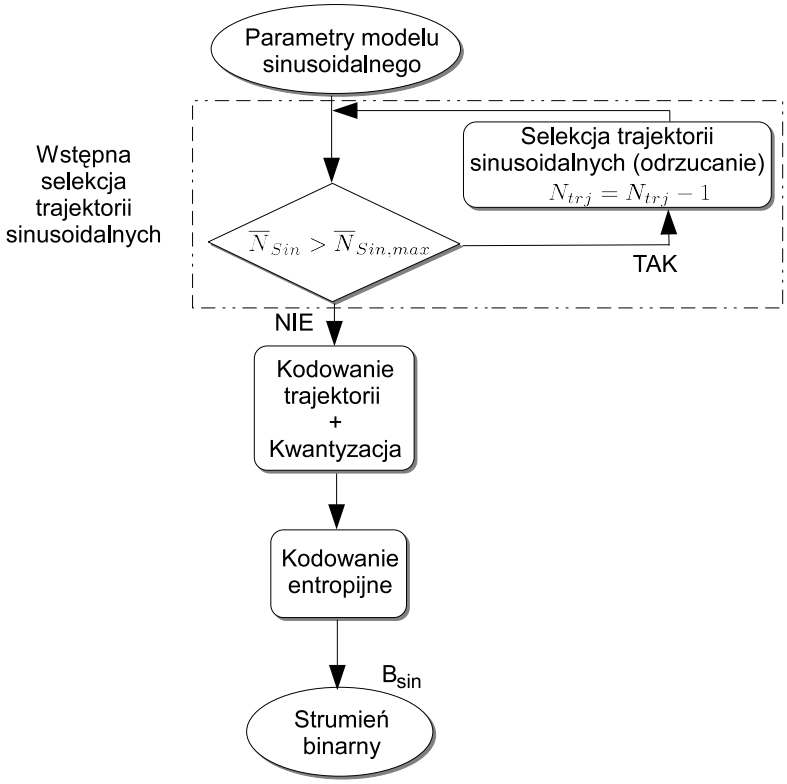
I wariant – bez pętli sterowania przepływnością

W tym wariancie sterowania pracą kodera zakładamy, że percepcyjnie istotniejszą informacją, z punktu widzenie rekonstrukcji sygnału, jest poprawna generacja jak największej liczby składowych tonalnych, kosztem obniżenia częstotliwości podziału f_{SBR}. Schemat blokowy algorytmu przedstawia rysunek 3.10.

Pierwszym etapem kodowania jest blok „wstępna selekcja trajektorii sinusoidalnych". W tym miejscu algorytm dokonuje wyboru N_{trj} trajektorii sinusoidalnych, dla których bieżąca, średnia liczba sinusoid na ramkę $\overline{N}_{Sin}$ nie przekracza $\overline{N}_{Sin,max}$. $\overline{N}_{Sin}$ wyznaczana jest na podstawie wzoru (3.24). Liczba $\overline{N}_{Sin,max}$ została ustalona podczas eksperymentów, których wyniki przedstawiono w rozdziale 3.9.4. $\overline{N}_{Sin,max}$ określa średnią liczbę sinusoid na ramkę, występujących w paśmie HF, przy której subiektywna ocena jakości zrekonstruowanego sygnału jest oceniana jako „dobra" (powyżej 4 na skali MUSHRA). Podczas badań ustalono $\overline{N}_{Sin,max} = 20$.

Jeśli średnia liczba trajektorii na ramkę $\overline{N}_{Sin}$ jest większa od ustalonego progu $\overline{N}_{Sin,max}$, wówczas usuwana jest pojedyncza trajektoria sinusoidalna ze zbioru trajektorii P (ang. *partials*). Powyższy proces odbywa się w bloku „selekcji trajektorii sinusoidalnych" (rysunek 3.10). Dokładny proces selekcji trajektorii sinusoidalnych zostanie omówiony w rozdziale 3.7.3.

Po dokonaniu selekcji trajektorie podlegają kwantyzacji oraz kodowaniu entropijnemu algorytmem Huffmana. Następnie obliczana jest prędkość bitowa strumienia binarnego



Rysunek 3.10: Schemat pracy kodera sinusoidalnego bez petli sterownia przepływnością – wariant I.

wygenerowanego przez parametry modelu sinusoidalnego B_{Sin}. Informacja ta służy do określenia puli bitów, które mogą być wykorzystane przez technikę MPEG-4 AAC HE:

$$B_{AAC-HE} = B_C - B_{Sin}, \tag{3.16}$$

gdzie:

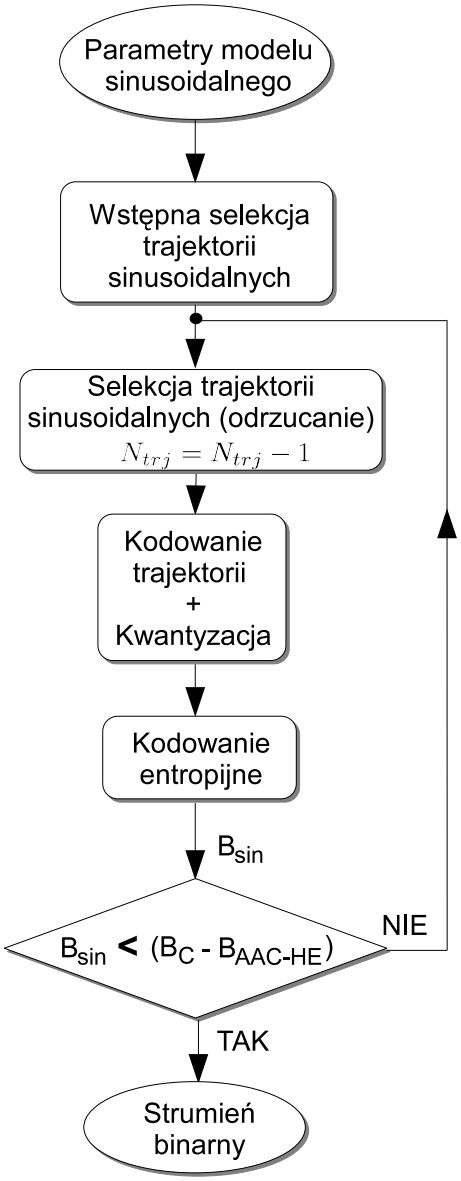
B_{AAC-HE} – maksymalna, dopuszczalna prędkość bitowa strumienia binarnego generowanego przez koder MPEG-4 AAC HE,

B_C – docelowa, całkowita prędkość bitowa,

B_{Sin} – prędkość bitowa strumienia danych określającego zakodowane parametry modelu sinusoidalnego.

II wariant – z pętlą sterowania przepływnością

W wariancie pracy kodera z pętlą sterowania przepływnością obliczany jest zasób bitów, który można przeznaczyć na zakodowanie parametrów modelu sinusoidalnego B_{Sin}. Jednocześnie, nie może nastąpić przełączenie granicy pracy kodera SBR - częstotliwość podziału f_{SBR}. Schemat blokowy algorytmu przedstawia rysunek 3.11.



Rysunek 3.11: Schemat pracy kodera sinusoidalnego z pętlą sterowania przepływnością – wariant II.

Koder podstawowy AAC reprezentuje grupę koderów perceptualnych, które dzięki zastosowaniu modelu psychoakustycznego potrafią utrzymać poziom szumu kwantyzacji poniżej progu słyszenia. Założenie to jest prawdziwe, jeśli koder podstawowy dysponuje odpowiednią pulą bitów do zakodowania składowych LF. Jednym ze sposobów zapewnienia poprawnej rekonstrukcji składowych LF jest ograniczenie pasma sygnału do częstotliwości granicznej f_{SBR}. W przypadku techniki MPEG-4 AAC HE koder podstawowy AAC decyduje o wartości częstotliwości f_{SBR} na podstawie docelowej prędkości bitowej, zgodnie z tabelą 2.1.

Wartości związane z zakresem pracy kodera podstawowego oraz techniki SBR, w zależności od docelowej prędkości bitowej, zostały opracowane przez grupę ekspertów z firmy Coding Technologies. Obecnie implementacja kodera SBR firmy Coding Technologies [2] wykorzystywana jest przez większość popularnych implementacji MPEG-4 AAC HE. Podział pasma w zależności od docelowej prędkości bitowej, na podstawie tabeli 2.1, został również wykorzystany w niniejszej pracy.

Algorytm pracy kodera w wariancie z pętlą sterowania przepływnością przebiega w następujących krokach:

1. Wybór $\overline{N}_{Sin,max}$ percepcyjnie najistotniejszych trajektorii sinusoidalnych, według algorytmu przedstawionego w wariancie I.
2. Odrzucenie pojedynczej trajektorii sinusoidalnej ze zbioru trajektorii P, zgodnie z algorytmem przedstawionym w punkcie 3.7.3.
3. Kwantyzacja i kodowanie entropijne pozostałych trajektorii sinusoidalnych.
4. Jeśli $B_{Sin} > (B_C - B_{AAC-HE})$, wówczas powrót do punktu 2.
5. Koniec algorytmu.

3.7.3. Wybór trajektorii sinusoidalnych do zakodowania

Kodery sinusoidalne, np. MPEG-4 SSC [69] nie kodują wszystkich składowych sinusoidalnych, gdyż byłoby to nieefektywne i wiązałoby się z przesyłaniem dużej liczby danych. Koder musi podjąć decyzję, które składowe są istotne dla poprawnej rekonstrukcji sygnału, i przesłać je do dekodera. Jednym ze sposobów selekcji składowych tonalnych w aktualnie stosowanych kodekach sinusoidalnych jest wykorzystanie współczynnika SMR [28] określającego stosunek sygnał-próg maskowania (ang. *signal-to-mask ratio*). Przez sygnał, w tym wypadku rozumiemy składową tonalną. Levine [105] zaproponował metodę selekcji trajektorii sinusoidalnych bazując na średnim SMR dla całej trajektorii oraz jej czasie trwania. Natomiast, dla skalowalnych koderów sinusoidalnych zaproponowano metody selekcji wykorzystujące jedynie współczynnik SMR, nie uwzględniając czasu trwania sinusoidy [131, 152]. We wspomnianych metodach trajektorie, których współczynnik SMR znajduje się poniżej pewnego przyjętego progu, są usuwane z dalszego przetwarzania.

W proponowanej technice BWE, przedstawiono metodę selekcji trajektorii sinusoidalnych uwzględniającą czas trwania trajektorii, jej całkowitą energię oraz częstotliwości poszczególnych składowych sinusoidalnych.

Pierwszym etapem algorytmu selekcji, jest wyznaczenie parametrów e_{trj}, które w pracy określane będą jako *miara perceptualnej istotności* danej trajektorii trj. Usunięcie trajektorii sinusoidalnej wprowadza zniekształcenie, którego stopień jest wprost proporcjonalny do parametru e_{trj} określonego wzorem:

$$e_{trj} = \sum_{l=1}^{L_{trj}} A_{trj}^2(l) \cdot W_A\left(f_{trj}(l)\right), \tag{3.17}$$

gdzie:

trj – indeks trajektorii,

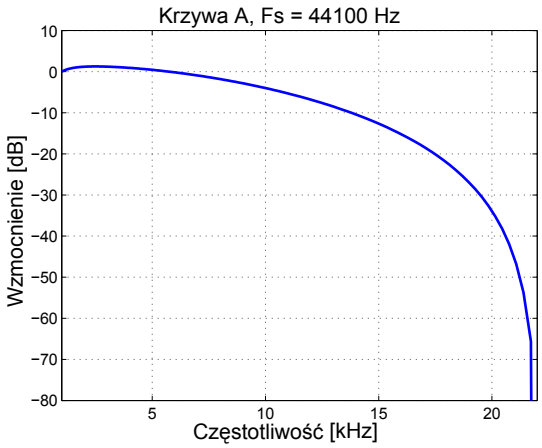
l – indeks sinusoidy,

L_{trj} – długość trajektorii trj,

$A_{trj}(l)$ – amplituda l-tej sinusoidy, dla trajektorii trj,

$W_A(f_{trj}(l))$ – wartość krzywej ważenia A dla częstotliwości f, l-tej sinusoidy w trajektorii o indeksie trj.

Wartość parametru e_{trj} zależy wprost proporcjonalnie od czasu trwania trajektorii sinusoidalnej oraz jej całkowitej energii. Dodatkowo parametr e_{trj} korygowany jest za pomocą krzywej ważenia A (rysunek 3.12). Im wyższa częstotliwość danej sinusoidy, tym mniejsza wartość W_A, co powoduje zmniejszenie parametru e_{trj}. Proces ten odpowiada sposobowi odbierania wysokich dźwięków przez ludzkie ucho. Wraz ze wzrostem częstotliwości, maleje czułość ludzkiego ucha na bodźce dźwiękowe [161].



Rysunek 3.12: Wykres krzywej ważenia A w zakresie 1 kHz – 22 kHz, dla częstotliwości próbkowania 44100 Hz.

Krzywe ważenia definiują w jaki sposób odbieramy dźwięki z określonego zakresu częstotliwości oraz danego zakresu poziomu głośności dźwięku. W proponowanym algorytmie zastosowano krzywą A, która odpowiada poziomowi głośności dźwięku z zakresu 0 – 40 [fonów]. Jest to poziom głośności odpowiadający normalnym warunkom odsłuchowym.

Zastosowana w pracy krzywa ważenia A, została zaprojektowana jako transmitancja filtru, zgodnie z zaleceniem standardu IEC [61]. Parametry filtru są następujące:

$$H(z) = \frac{0.2557 - 0.5115z^{-1} - 0.2557z^{-2} + 1.0230z^{-3} - 0.2557z^{-4} - 0.5115z^{-5} + 0.2557z^{-6}}{1 - 4.0196z^{-1} + 6.1894z^{-2} - 4.4532z^{-3} + 1.4208z^{-4} - 0.1418z^{-5} + 0.0044z^{-6}} \tag{3.18}$$

Algorytm selekcji trajektorii sinusoidalnych

Na podstawie wzoru (3.17) konstruowany jest zbiór E, taki że:

$$E = \{e_{trj}\}_{trj=1}^{N_{trj}}, \tag{3.19}$$

opisujący miarę perceptualnej istotności trajektorii. Każdemu elementowi zbioru E, odpowiada element ze zbioru P:

$$P = \{p_{trj}\}_{trj=1}^{N_{trj}}. \tag{3.20}$$

Element p_{trj} przedstawia wektor parametrów opisujących trajektorię sinusoidalną trj, tzn. wektor amplitud A_{trj} i częstotliwości f_{trj}.

Zbiór E wykorzystywany jest w algorytmie selekcji trajektorii sinusoidalnych - rysunki 3.10 oraz 3.11. Algorytm jest iteracyjną procedurą, podczas której w każdym kroku odrzucana jest pojedyncza trajektoria ze zbioru P. Trajektoria o indeksie trj wybierana jest zgodnie z warunkiem:

$$trj = \arg\min_{trj} \{e_{trj}\}, \qquad trj \in \{1, \ldots N_{trj}\}. \tag{3.21}$$

Następnie, na podstawie indeksu trj odrzucana jest trajektoria ze zbioru P:

$$P = P \setminus \{p_{trj}\}. \tag{3.22}$$

Jednocześnie, aktualizacji podlega liczba trajektorii N_{trj}:

$$N_{trj} = N_{trj} - 1. \tag{3.23}$$

Miara e_{trj} preferuje trajektorie sinusoidalne długim czasie trwania. Ponieważ trajektorie sinusoidalne kodowane są predykcyjnie, tego typu strategia pozwala znacząco ograniczyć prędkość bitową strumienia generowanego przez model sinusoidalny (B_{Sin}). Należy

również podkreślić, że część trajektorii o długości mniejszej niż $L_{trj,thr}$, została wcześniej usunięta w bloku selekcji składowych tonalnych (rozdział 3.4.2). Pozwala to na pominięcie w procesie kodowania krótkotrwałych trajektorii o dużej energii, które w większości przypadków należą do transjentu lub szum [13].

Wstępna selekcja trajektorii sinusoidalnych

Obydwa warianty sterowania pracą kodera wykorzystują blok wstępnej selekcji trajektorii sinusoidalnych. Na tym etapie wybierane jest N_{trj} trajektorii sinusoidalnych, dla których średnia liczba sinusoid na ramkę - $\overline{N}_{Sin}$, nie przekracza $\overline{N}_{Sin,max}$. Bieżąca wartość $\overline{N}_{Sin}$ obliczana jest na podstawie wzoru:

$$\overline{N}_{Sin} = \frac{\sum_{trj=1}^{N_{trj}} L_{trj}}{R}, \tag{3.24}$$

gdzie:

trj – indeks trajektorii,

N_{trj} – liczba trajektorii,

L_{trj} – długość trajektorii trj,

R – liczba ramek.

3.7.4. Kodowanie Huffmana

Parametry modelu sinusoidalnego, czyli chwilowa amplituda i częstotliwość zostały zakodowane wykorzystując algorytm Huffmana.

W pracy wykorzystano implementację [145] kodowania Huffmana stworzoną przez Karla Skrettinga [144].

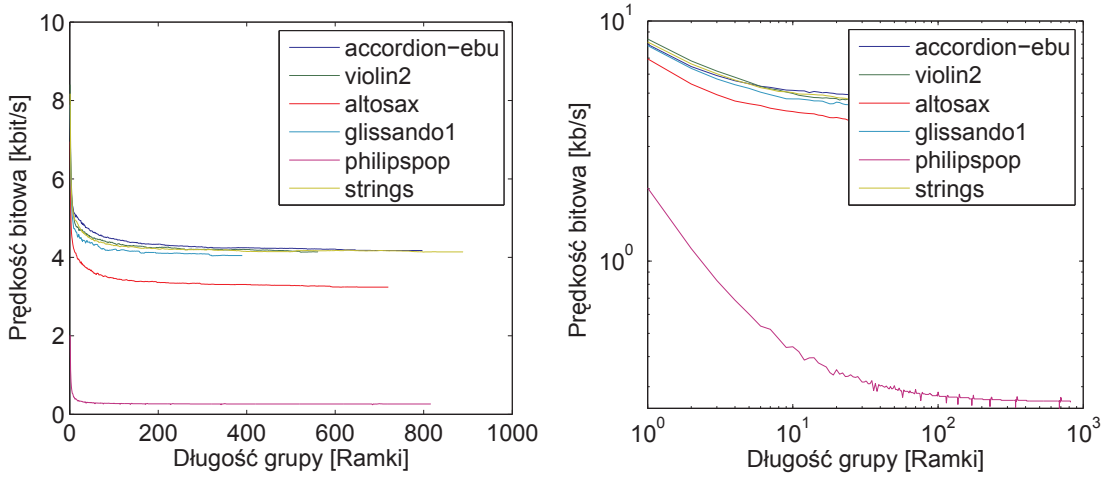
Kodowaniu podlega błąd predykcji dla kolejnych wartości trajektorii sinusoidalnej p_{trj}. W proponowanej technice zastosowano odnawianie cyklu układu predykcji co 250 ramek. Oznacza to, że pamięć predyktora jest zerowana w przybliżeniu co 2,9 s. Jest to czas swobodnego dostępu do strumienia binarnego t_R (ang. *random access*), czyli maksymalny czas jaki potrzebuje dekoder, aby poprawnie zdekodować dany fragment sygnału. Należy podkreślić, że początki trajektorii kodowane są bezwzględnie, tzn. ich skwantowane wartości umieszczane są bezpośrednio w strumieniu. Oznacza to, że czas t_R może ulec skróceniu, w zależności od struktury trajektorii sinusoidalnych.

Czas swobodnego dostępu do strumienia binarnego został przyjęty na podstawie standardu MPEG-4 AAC HE. W standardzie zdefiniowano czas zerowania pamięci predyktorów w technice LTP (ang. *Long Term Prediction*) na 2,9 s [67].

W zastosowanym algorytmie, przyjęto model kodowania entropijnego ze stalą tablicą Huffmana, przesyłaną raz na początku strumienia. Tablica wyznaczana jest dla każdego sygnału niezależnie, bazując na rozkładzie symboli w danym nagraniu.

W pracy zbadano wpływ zastosowanej stałej tablicy kodów Huffmana, na wielkość strumienia, generowanego przez proponowaną technikę BWE. W tym celu, przeprowadzono eksperyment, w którym:

- tablica kodów wyznaczana była co pewien interwał czasu, wyrażony w liczbie ramek, nazwana dalej - grupą ramek,
- w danej grupie ramek obowiązuje pojedyncza tablica kodów,
- tablica kodów Huffmana przesyłana jest raz na grupę ramek,
- długość grupy ramek zmieniała się w przedziale od 1 do 800 ramek ze skokiem równym 1,
- w eksperymentach wykorzystano sygnały foniczne z tabeli 1.1.



(a) Skala liniowa

(b) Skala logarytmiczna loglog

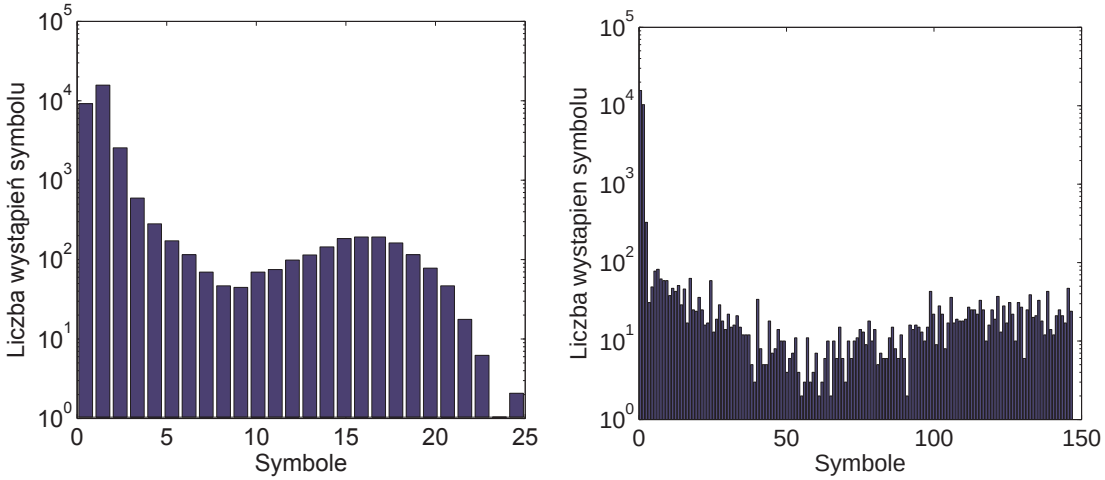
Rysunek 3.13: Zależność prędkości bitowej od długości grupy ramek, objętych kodowaniem tą samą tablicą Huffmana. Jest to porównanie dla wszystkich sygnałów fonicznych biorących udział w badaniu, zgodnie z tabelą 1.1. Tablica kodów entropijnych przesyłana jest raz na każdą grupę ramek, np. dla długości grupy równej 100, tablica jest umieszczana w strumieniu bitowym co 100 ramek. Szczegółowe wyniki znajdują się w dodatku A.

Na podstawie wyników badań, przedstawionych na rysunku 3.13, możemy zaobserwować, że jeśli długość grupy jest większa od 50 ramek, wówczas wyraźnie spada prędkość bitowa generowanego strumienia. Następnie, prędkość bitowa stabilizuje się na określonym poziomie, zależnym od badanego sygnału. Interwał 50 ramek odpowiada czasowi około 0,5 s dla częstotliwości próbkowania $f_s = 44100$ Hz. Biorąc pod uwagę powyższą obserwację, można założyć, że przedział czasu równy 0,5 s jest wystarczający, aby zebrać odpowiednią statystykę danych określających parametry modelu sinusoidalnego oraz zbudować na tej podstawie tablicę kodów Huffmana. Dodatkowo, możliwe jest również zastosowanie grupy ramek o dowolnej długości większej od 50, dla którego prędkość bitowa zakodowanych parametrów pozostanie taka sama. Na tej podstawie w dalszej części pracy, w badaniach symulacyjnych, wykorzystano pojedynczą tablicę kodów Huffmana, wyznaczaną tylko raz dla całego sygnału.

W aktualnie stosowanych standardach transmisji telewizji cyfrowej, takich jak MPEG-4 AVC/H.264 [75], jednostką synchronizacji strumienia danych wizyjnych i fonicznych jest grupa obrazów. Czas trwania grupy obrazów dla aktualnie stosowanych implementacji wynosi około 0,5 s. Na tej podstawie można założyć, że proponowana technika BWE spełnia wymagania odnośnie swobodnego dostępu do strumienia danych. Wyznaczony interwał czasu zapewnia również poprawną synchronizację danych strumienia fonicznego oraz wizyjnego.

W pracy przeprowadzono również serię eksperymentów sprawdzających, czy wystarczy konstrukcja jednej uniwersalnej tablicy Huffmana dla szerokiej klasy sygnałów. Podczas badań zebrano statystykę skwantowanych wartości amplitudy oraz częstotliwości na podstawie wszystkich sygnałów testowych. Wyniki przedstawia rysunek 3.14.

Możemy zaobserwować, że rozkład wystąpień symboli dla skwantowanych wartości amplitudy i częstotliwości ma wyraźne maksimum w okolicach zera. Z tego powodu, dane przedstawione na wykresie 3.14 możemy podzielić na wartości małe w zakresie od 0 do 3 oraz pozostałe wartości duże.



(a) Amplituda (b) Częstotliwość

Rysunek 3.14: Liczba wystąpień symboli dla skwantowanych wartości dla amplitudy i częstotliwości. Wykresy są w skali logarytmicznej.

Rozkład wartości dużych jest zbliżony do równomiernego, co oznacza, że nie jest możliwe skonstruowanie pojedynczej uniwersalnej tablicy kodów Huffmana. Taka tablica musiałaby reprezentować bardzo dużą liczbę danych. Z drugiej strony, wartości duże reprezentowane byłyby przez słowa kodowe o podobnej długości. Oznacza to, że prawdopodobieństwa wystąpienia symboli są równe, a zatem w tym wypadku kodowanie Huffmana jest nieefektywne.

Biorąc jednak pod uwagę fakt, że liczba wystąpień wartości dużych dla pojedynczego nagrania jest od 100 do 1000 krotnie mniejsza od wartości małych, w pracy zastosowano pojedynczą tablicę kodów Huffmana, generowaną dla każdego sygnału osobno. Takie podejście pozwala skonstruować tablicę kodów dopasowaną do rozkładu wartości w danym sygnale.

3.8. Proces dekodowania sygnału

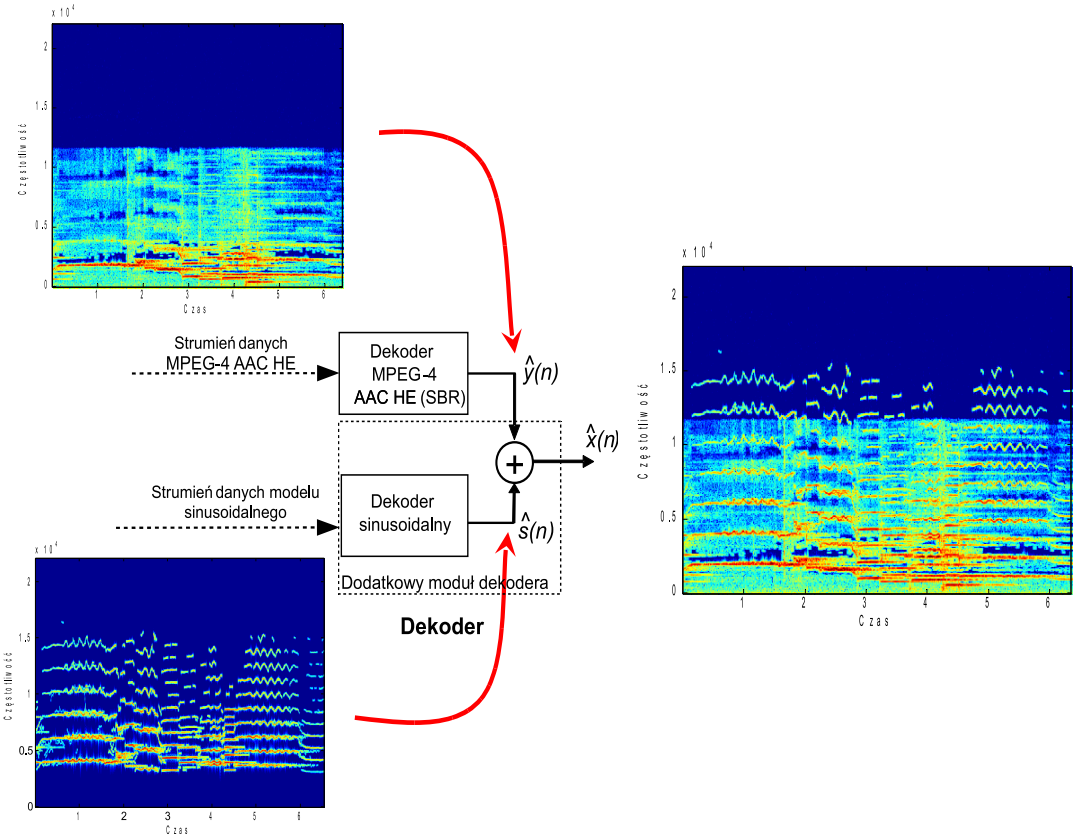
Zrekonstruowany sygnał jest sumą sygnału zsyntezowanego z parametrów modelu sinusoidalnego oraz sygnału zdekodowanego przez MPEG-4 AAC HE. Zadaniem dekodera SBR jest rekonstrukcja składowej szumowej HF oraz transjentów. Należy zauważyć, że jeśli dekoder nie jest przystosowany do odbioru parametrów modelu sinusoidalnego, wówczas w dalszym ciągu możliwa jest rekonstrukcja sygnału HF. W tym wypadku odtworzona zostanie składowa szumowa HF. Zrekonstruowany w ten sposób sygnał będzie niższej jakości

niż zdekodowany za pomocą MPEG-4 AAC HE, jednak w subiektywnej ocenie słuchaczy będzie wyżej oceniony niż sygnał ograniczony pasmowo.

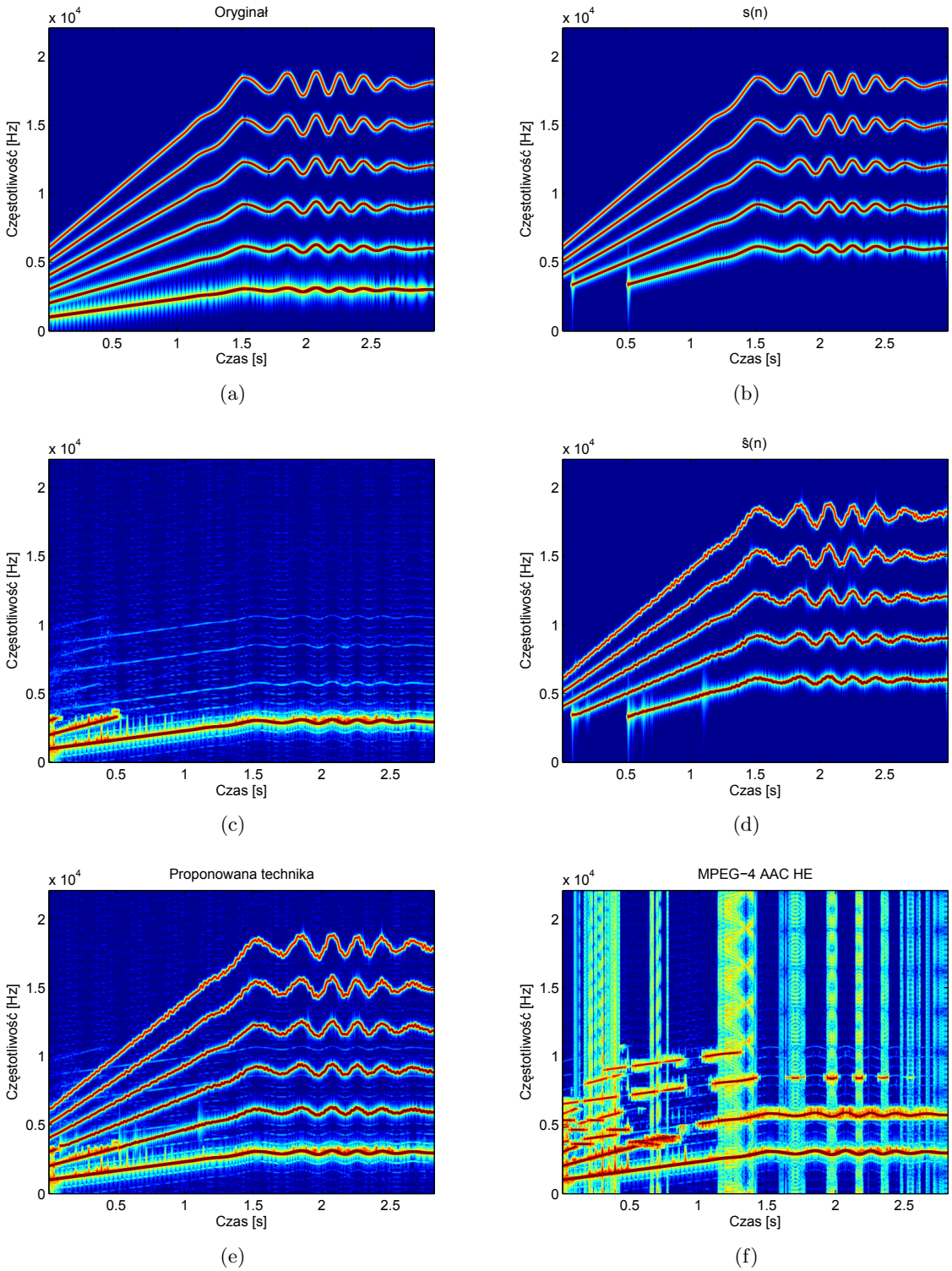
Na rysunkach 3.16 i 3.17 pokazany został proces analizy oraz syntezy (rekonstrukcji) sygnału, odpowiednio syntetycznego oraz naturalnego. Na rysunkach (b) przedstawione są spektrogramy wykrytych oraz zsyntezowanych składowych tonalnych. Natomiast rysunki (c) pokazują zdekodowany w dekoderze MPEG-4 AAC HE sygnał y ze stłumionymi składowymi tonalnymi. W przypadku sygnału syntetycznego składowe tonalne zostały całkowicie usunięte z sygnału wejściowego x, a w paśmie składowych HF nie ma żadnych składowych. W przypadku sygnału naturalnego przedstawionego na rysunku 3.17 w zakresie dużych częstotliwości pozostała tylko składowa szumowa.

Rysunki (e) i (f) przedstawiają porównanie odpowiednio proponowanej techniki BWE w porównaniu MPEG-4 AAC HE. W przypadku proponowanej techniki BWE, sygnał został wiernie zrekonstruowany w całym zakresie częstotliwości. Natomiast, MPEG-4 AAC HE zrekonstruował sygnał jedynie do częstotliwości $f_{SBR_{max}} = 10$ kHz. Dodatkowo nie została poprawnie odtworzona struktura harmoniczna. Mechanizm replikacji widma nie był w stanie odtworzyć szybkozmiennych składowych tonalnych - widocznych w zakresie od 0 do 1 sekundy. Dla przedziału czasu od 1,5 s do 3 s, dekoder SBR odtworzył drugą harmoniczną kopiując w dziedzinie częstotliwości pierwszą harmoniczną sygnału. W ten sposób nie zostały zachowane chwilowe modulacje częstotliwości (efekt vibratto), które powinny proporcjonalnie rosnąć wraz ze wzrostem częstotliwości - stanowiąc wielokrotność częstotliwości podstawowej.

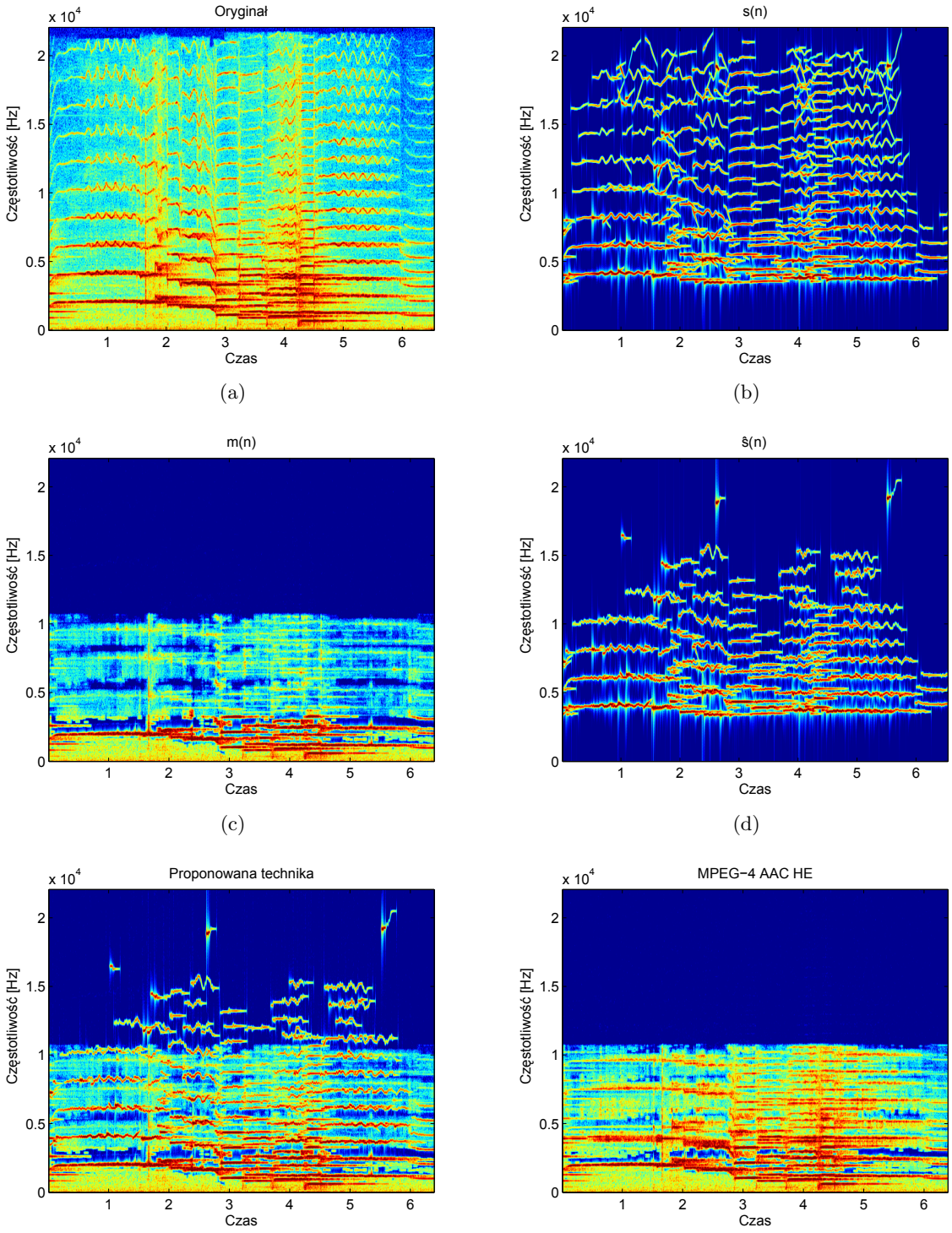
Podobne zniekształcenia można zauważyć na rysunku 3.17f, podczas gdy proponowana technika BWE poprawnie rekonstruuje strukturę harmoniczną składowych tonalnych. Jednocześnie proponowana technika BWE generuje dodatkowe składowe powyżej częstotliwości $f_{SBR_{max}}$, zwiększając tym samym szerokość pasma sygnału $\hat{x}$ w stosunku do MPEG-4 AAC HE.



Rysunek 3.15: Rekonstrukcja sygnału – sygnał wyjściowy jest sumą sygnału $\hat{m}$ odtworzonego w dekoderze MPEG-4 AAC HE oraz sygnału $\hat{s}$ odtworzonego w dekoderze sinusoidalnym. Zrekonstruowany sygnał $\hat{x}$ posiada poprawną strukturę harmoniczną składowych tonalnych oraz szersze pasmo w stosunku do sygnału dekodowanego bezpośrednio w dekoderze MPEG-4 AAC HE.



Rysunek 3.16: Spektrogramy przedstawiające poszczególne etapy kodowania sygnału z zastosowaniem proponowanej techniki: a) sygnał wejściowy, b) zsyntezowane składowe tonalne s, c) sygnał y po procesie tłumienia, d) zsyntezowane składowe tonalne po stronie dekodera - sygnał $\hat{s}$, e) zrekonstruowany sygnał - proponowana technika, f) zrekonstruowany sygnał za pomocą MPEG-4 AAC HE. Kodowaniu podlegał sygnał syntetyczny o harmonicznej strukturze. Sygnały zakodowano dla prędkości bitowej równej 16 kb/s.



Rysunek 3.17: Spektrogramy przedstawiające poszczególne etapy kodowania sygnału z zastosowaniem proponowanej techniki: a) sygnał wejściowy, b) zsyntezowane składowe tonalne s, c) sygnał y po procesie tłumienia, d) zsyntezowane składowe tonalne po stronie dekodera - sygnał $\hat{s}$, e) zrekonstruowany sygnał - proponowana technika, f) zrekonstruowany sygnał za pomocą MPEG-4 AAC HE. Kodowaniu poddany został fragment nagrania skrzypiec - sygnał *violin2*. Sygnały zakodowano dla prędkości bitowej równej 16 kb/s.

3.9. Właściwości konstrukcyjne kodeka

3.9.1. Zgodność proponowanej techniki BWE z MPEG-4 AAC HE

Proponowana technika BWE została zaprojektowana jako opcjonalne rozszerzenie standardu MPEG-4 AAC HE. Jednoczesne wykorzystanie wspomnianych technik, wymaga ustalenia grupy parametrów, których kontrola ma wpływ na otrzymaną funkcjonalność oraz cechy użytkowe. Poniżej opisane zostaną parametry, które należy uwzględnić podczas projektowania schematu kompresji z proponowaną techniką BWE.

- Długość okna analizy N oraz przesunięcie (skok) H - w proponowanej technice BWE przyjęto $N = 2048$ i $H = 512$. Dla porównania, w koderze podstawowym AAC, będący częścią MPEG-4 AAC HE, parametry te wynoszą odpowiednio $N = 1024$ i $H = 512$, a w dekoderze AAC $N = 2048$ i $H = 1024$. Dla kodera SBR, odpowiednio $N = 320$ i $H = 32$ oraz dekoder SBR $N = 640$ i $H = 64$.

  Za pomocą parametrów N i H można regulować dokładność siatki czasowo-częstotliwościowej. Zwiększenie rozdzielczości siatki prowadzi do wzrostu jakości zdekodowanego sygnału, kosztem liczby dodatkowych danych do zakodowania.

  Rozdzielczość czasowa siatki ma również wpływ na czas swobodnego dostępu do strumienia, który rośnie wraz ze wzrostem skoku H. Rozdzielczość siatki czasowo-częstotliwościowej proponowanej techniki jest większa od kodera AAC oraz mniejsza od kodera SBR. Parametry N i H zostały dobrane doświadczalnie i stanowią kompromis pomiędzy dokładnością estymacji parametrów w analizie czasowo-częstotliwościowej, czasem obliczeń oraz liczą danych wymaganych do zakodowania.

- Cykliczne zerowanie predyktorów - w proponowanej technice BWE predykcja wykorzystywana jest do kompresji trajektorii sinusoidalnych. W AAC zerowanie predyktorów odbywa się co 250 ramek i jest stosowana niezależnie dla pasm skalowania (w technice AAC jest 80 pasm skalowania). Taki sam cyklu zerowania predykcji, czyli około 2,9 s, zastosowano w proponowanej technice BWE. W proponowanej technice predyktory przełączają się, podobnie jak w AAC, niesynchronicznie co kilka ramek (punkt 3.7.4).

- Tablica kodów Huffmana - w AAC tablica kodów jest opisywana w standardzie, co umożliwia dowolny dostęp do strumienia. W proponowanej technice BWE, dla

celów eksperymentalnych, tablica przesyłana jest raz na początku strumienia. Dokładna dyskusja tego problemu została opisana w rozdziale 3.7.4. Należy zaznaczyć, że opracowanie uniwersalnych tablic kodów dla proponowanej techniki BWE byłoby procesem czasochłonnym i znacząco wydłużyłoby czas badań.

- Strumień binarny - proponowane rozwiązanie nie ingeruje w format standardowego strumienia MPEG-4 AAC HE. Generowany jest dodatkowy strumień danych, który może być zignorowany w dekoderze, lub wykorzystany do poprawy rekonstrukcji tonalnych składowych sygnału HF.

Dodatkowo, możemy wyróżnić następujące cechy użytkowe proponowanej techniki BWE:

- uniwersalność - możliwe jest zastosowanie opisywanego algorytmu BWE, jako rozszerzenie innych technik kodowania sygnałów fonicznych. Dotyczy to zwłaszcza technik wykorzystujących replikację widma do rekonstrukcji składowych HF, np. Dolby Digital Plus (EAC-3) [40].

- złożoność obliczeniowa - problem złożoności obliczeniowej nie był badany w pracy, której głównym celem było stworzenie modelu techniki BWE. Głównymi elementami, które wpływają na złożoność obliczeniową jest analiza sinusoidalna oraz pętla sterowania przepływnością bitową. W zależności od potrzeb można jednak wpływać na złożoność obliczeniową, np. poprzez zmianę parametrów N i H lub ograniczając liczbę trajektorii w bloku „wstępna selekcja trajektorii sinusoidalnych".

- elastyczność - proponowaną technikę można dostosować do potrzeb odbiorcy, a w szczególności można sterować:

  – złożonością obliczeniową, poprzez zmianę parametrów kodeka,

  – dostępem do strumienia binarnego - swobodny dostęp do strumienia zależy od parametrów N i H oraz częstotliwości przesyłanej tablicy kodów Huffmana (rozdział 3.7.4),

  – rozmiarem strumienia binarnego - ograniczenie strumienia binarnego można uzyskać poprzez odrzucenie części trajektorii sinusoidalnych, zgodnie z regułą przedstawioną w rozdziale 3.7.3.

3.9.2. Określenie parametrów wstępnych modelu sinusoidalnego

W drodze procesu analizy sinusoidalnej otrzymujemy bardzo dużo parametrów opisujących trajektorie sinusoidalne. Przykładowo jeśli średnio na ramkę zostanie wykryte około 100 sinusoid to w efekcie możemy otrzymać nawet kilka tysięcy trajektorii dla kilkusekundowego sygnału. Zakładając, że parametry modelu sinusoidalnego będą kwantowane w sposób nie wprowadzający zauważalnych zniekształceń, wówczas prędkość bitowa wygenerowanego strumienia binarnego może wynosić około kilkadziesiąt kb/s.

Prędkość bitowa strumienia danych reprezentującego parametry modelu sinusoidalnego powinna zawierać się w zakresie 3–6 kb/s - prędkością bitową porównywalną z kodekiem SBR. Należy więc poszukiwać rozwiązań ograniczających liczbę trajektorii sinusoidalnych do percepcyjnie istotnych. Z drugiej strony bardzo trudno jest wyeliminować część danych w procesie pętli sterowania przepływnością, z powodu braku odpowiedniej miary zniekształceń, która określi czy dany stopień kwantyzacji jest odpowiedni.

Innym sposobem zmniejszenia liczby danych modelu sinusoidalnego, mogłoby być zastosowanie modelu psychoakustycznego, który określa czy dane składowe sygnału są słyszalne, tzn. ich energia jest większa od progu maskowania [28, 161]. Jednak i w tym przypadku zastosowanie znanych modeli (np. 1 i 2 model psychoakustyczny ze standardu MPEG ISO/IEC 11172-3 [63]) jest nieskuteczne, ponieważ modele te działają w obrębie krótkiego odcinka czasu, tzn. w ramach pojedynczego okna analizy sygnału. Natomiast trajektorie sinusoidalne mogą mieć długość nawet kilku sekund, obejmując tym samym wiele okien analizowanego sygnału. Dodatkowo model psychoakustyczny ma tendencję do maskowania składników o małej energii lub bliskich sobie. Problem w tym, że z punktu widzenie efektywności kompresji oraz poprawnej syntezy sygnału wiele składowych sygnału o niskiej energii powinno być zachowane. Sinusoidy te wchodzą w skład trajektorii sinusoidalnych. Usuwanie pojedynczych sinusoid powoduje przerywanie trajektorii, co zwiększa prędkość transmisji, z powodu konieczności zakodowania początków nowych trajektorii.

Dodatkowo zaczyna być odczuwalny efekt gwałtownego pojawienia się lub wygaśnięcia trajektorii, co powoduje „trzaski” lub efekt szumu muzycznego [38].

W związku z powyższymi problemami przeprowadzono serię badań wstępnych, których celem było ustalenie parametrów wejściowych modelowania sinusoidalnego, tj.:

- kwantyzacja amplitud i częstotliwości,

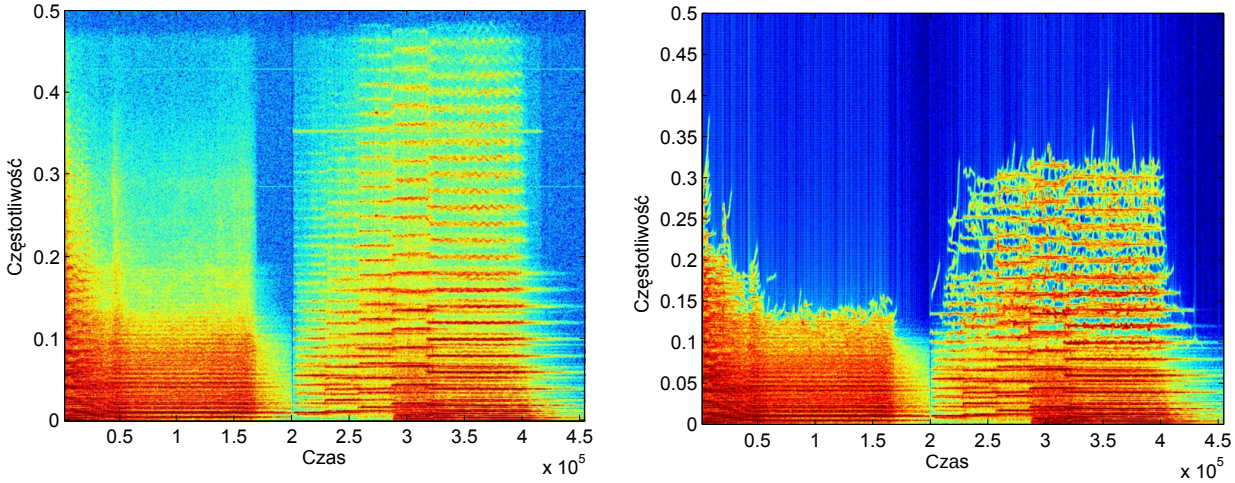
- liczba trajektorii sinusoidalnych w bloku „wstępna selekcja trajektorii sinusoidalnych", pętli sterowania przepływnością.

Zbadano również wpływ powyższych parametrów na jakość rekonstrukcji sygnału HF. Badania były przeprowadzone zgodnie z metodologią testów subiektywnych MUSHRA [78].

Przykładowy spektrogram sygnału poddanego testom, został przedstawiony na rysunku 3.18. Sygnały testowe są sumą:

- oryginału ograniczonego pasmowo do 4 kHz,
- sygnału syntetycznego, wygenerowanego za pomocą modelu sinusoidalnego, reprezentującego składowe tonalne HF powyżej 4 kHz.

Częstotliwość podziału 4 kHz, odpowiada zdefiniowanej w rozdziale 1 granicy podziału pomiędzy składowymi małych i dużych częstotliwości f_{BWE}.



(a) Oryginał (b) Sygnał zakodowany

Rysunek 3.18: Przykładowe spektrogramy sygnałów testowych wykorzystanych w badaniu.

3.9.3. Kwantyzacja amplitud i częstotliwości

W badaniu kwantyzacji poddano amplitudy i częstotliwości, opisujące trajektorie modelu sinusoidalnego. Parametry zostały poddane kwantyzacji równomiernej w skali logarytmicznej, zgodnie z opisem w rozdziale B.4.2. Na podstawie wstępnych badań eksperymentalnych ustalono poniższe przedziały kwantyzacji:

- częstotliwość - 20, 40, 80 centów,
- amplituda - 3, 6 dB,

W badaniu wzięło udział 22 słuchaczy, którzy oceniali 11 sygnałów testowych pochodzących z różnych źródeł [42, 73, 151]. Wyniki badań przedstawia rysunek 3.19. Możemy zaobserwować, że oceny dla różnych progów kwantyzacji są zbliżone do siebie, co uniemożliwia podjęcie decyzji o wyborze właściwego progu kwantyzacji. W tej sytuacji, próg kwantyzacji dla amplitudy i częstotliwości wybrano na podstawie najgorszego przypadku, pokazanego na rysunku 3.19 (b). Możemy zaobserwować, że słuchacze wyraźnie niżej ocenili sygnały zakodowane z krokiem kwantyzacji dla amplitudy równym 3 dB oraz dla częstotliwości równym 40 centów (f20-a3). Ocena MUSHRA w tym przypadku wynosi 4,55, natomiast dla kolejnych kroków kwantyzacji spada poniżej 4, czyli słuchacze dostrzegają zniekształcenia wprowadzane przez kwantyzację.

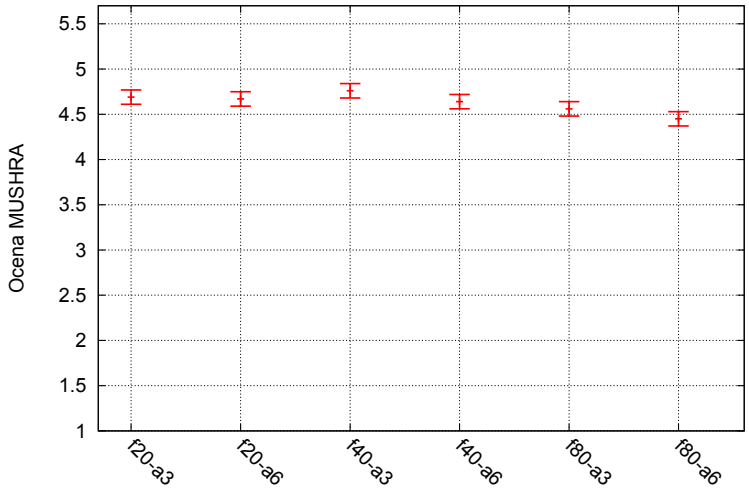
Na podstawie powyższych obserwacji zdecydowano się na wybór następujących przedziałów kwantyzacji:

- przedział kwantyzacji amplitudy: 3 dB,
- przedział kwantyzacji częstotliwości: 40 centów.

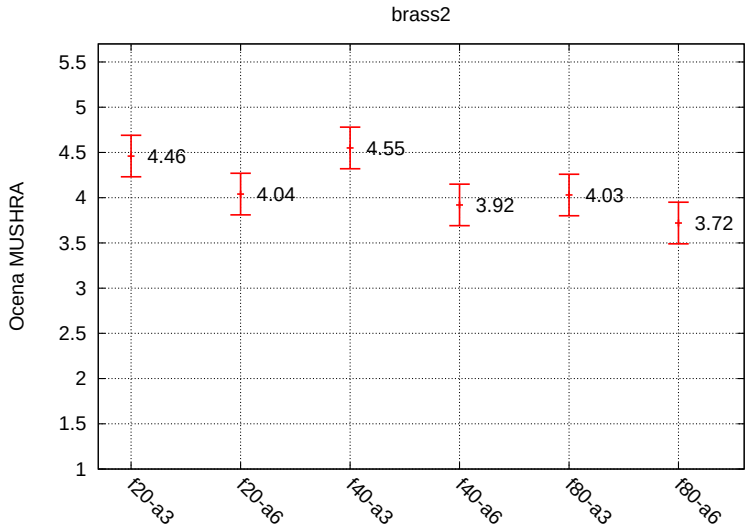
3.9.4. Proponowana strategia ograniczenia liczby trajektorii sinusoidalnych

W bloku „wstępna selekcja trajektorii sinusoidalnych", pokazanym na rysunkach 3.10 i 3.11, liczba trajektorii N_{trj} ograniczana jest do średniej liczby sinusoid na ramkę $\overline{N}_{Sin,max}$. Aby wyznaczyć $\overline{N}_{Sin,max}$ przeprowadzono badanie, którego celem było określenie minimalnej liczby sinusoid niezbędnych do poprawnej rekonstrukcji składowych tonalnych w ramce.

Wyniki badań zostały przedstawione na rysunku 3.20. W pierwszym teście $\overline{N}_{Sin,max}$ ustalono na 6, 8, 10 i 15 sinusoid na ramkę. Rysunek przedstawia wpływ $\overline{N}_{Sin,max}$ na jakość zdekodowanego sygnału. Otrzymane wyniki są bardzo zbliżone do siebie, co uniemożliwiło podjęcie decyzji. Dlatego przeprowadzono dodatkowy test dla $\overline{N}_{Sin,max} = 5$, 10, 20, 30, 40 i 50 - rysunek 3.20 (b).



(a) Testy wstępne: 22 osoby - średnia ocena dla wszystkich nagrań biorących udział w badaniu



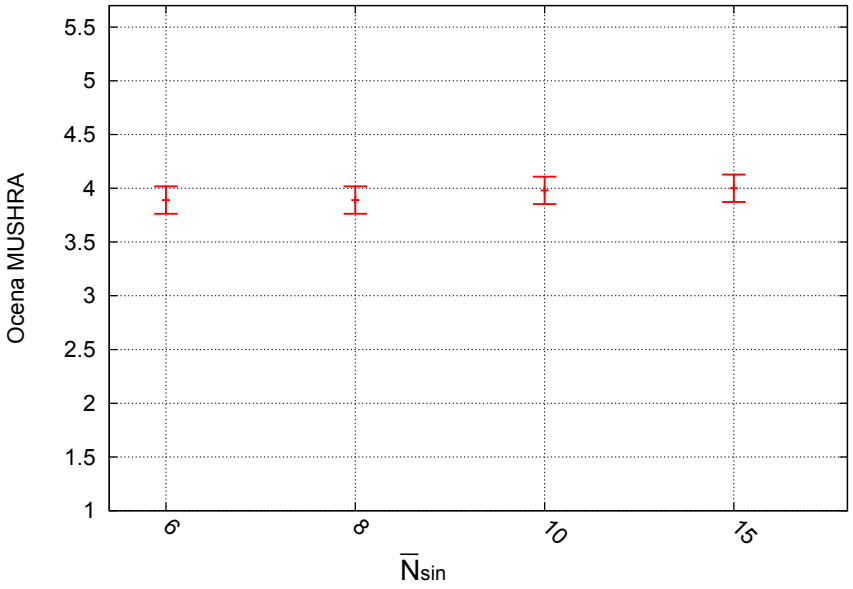
(b) Najgorszy przypadek

Rysunek 3.19: Wyniki eksperymentów dla kwantyzacji amplitud i częstotliwości. Objaśnienie: $f_{xx} - a_y$, gdzie xx - przedział kwantyzacji częstotliwości w centach, y - przedział kwantyzacji amplitudy w dB. Na przykład $f_{20} - a_3$ oznacza składowe kodowane były z przedziałem kwantyzacji częstotliwości równym 20 centów i amplitudy równym 3 dB.

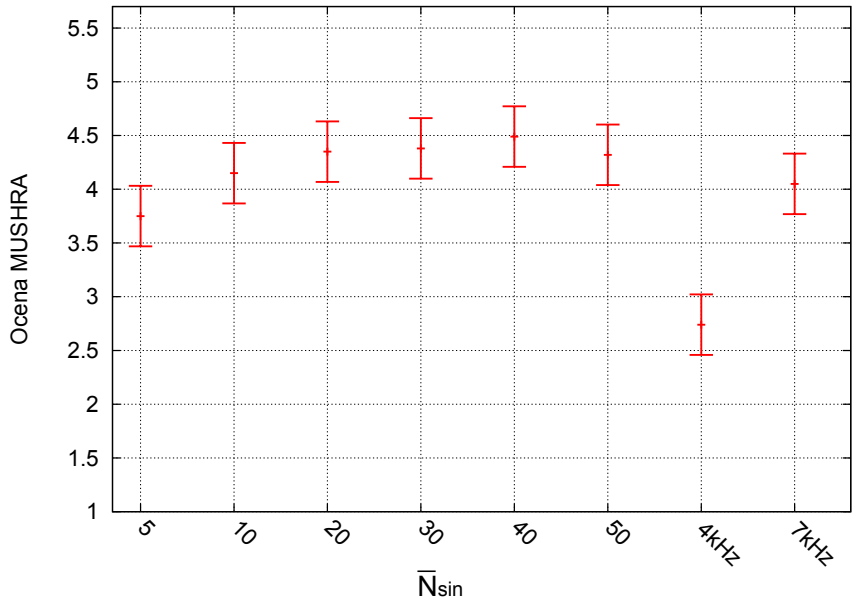
Na podstawie otrzymanych wyników można wyciągnąć wniosek, że należy ustalić wstępną, średnią liczbę sinusoid na ramkę $\overline{N}_{Sin,max}$ na około 20. W ten sposób, zrekonstruowany sygnał uzyskuje wynik percepcyjnie „dobry” na skali MUSHRA (4 – dobry).

Można również zauważyć, że zwiększanie średniej liczby sinusoid powyżej 40 spowodowało pogorszenie jakości ocenianego sygnału. Wynika to z błędnego wyboru składników tonalnych przez analizę sinusoidalną. W tym wypadku, szum reprezentowany był przez

składniki tonalne, co powoduje powstanie negatywnego efektu szumu muzycznego i jednocześnie pogorszenie jakości zdekodowanego sygnału. Szerzej ten problem opisany został w rozdziale 3.4.2 i 3.5.



(a) Testy wstępne: 22 osoby



(b) Testy dodatkowe: 3 osoby

Rysunek 3.20: Wyniki badań odsłuchowych pozwalające oszacować niezbędną liczbę sinusoid w górnej części pasma. Objaśnienie: sin_{ss} - gdzie ss to średnia liczba sinusoid na ramkę $\overline{N}_{Sin}$. 4kHz i 7kHz to sygnały ograniczone pasmowo odpowiednio do 4 i 7 kHz.

Na podstawie badań ustalono $\overline{N}_{Sin,max} = 20$, jako minimalną liczbę niezbędną do poprawnej rekonstrukcji sygnału.

3.10. Podsumowanie

W rozdziale przedstawiono nową, autorską metodę rozszerzania widma wykorzystującą modelowanie sinusoidalno-szumowe, która umożliwia poprawną rekonstrukcję składowych tonalnych dużych częstotliwości. Nowa technika może być wykorzystana w połączeniu z istniejącymi już technikami BWE, pozwalając na zmniejszenie zniekształceń powodowanych przez replikację widma stosowaną w obecnych standardach kompresji sygnałów fonicznych (np. Dolby EAC-3 [40], MPEG-4 AAC HE [68]). W pracy zastosowano połączenie kodeka standardu MPEG-4 AAC HE oraz modelowania sinusoidalnego.

W rozdziale zaproponowano podstawowe elementy funkcjonalne techniki BWE oraz przykład implementacji bloków kodeka. Dodatkowo zaproponowano schemat wyboru składowych tonalnych, który zapewnia skuteczną separację składowych tonalnych i szumowych. Przedstawiono również sposób ograniczenia złożoności obliczeniowej algorytmu poprzez zastosowanie detektora tonalności i możliwe pominięcie przez koder sinusoidalny wskazanych fragmentów sygnału.

W rozdziale zaproponowano również dwa warianty algorytmu kompresji składowych tonalnych z włączoną lub wyłączoną pętlą sterowania prędkością bitową. W tym celu skonstruowano miarę perceptualnej istotności trajektorii sinusoidalnych, umożliwiającą podjęcie decyzji o odrzuceniu danej trajektorii podczas kodowania.

Przeprowadzono liczne badania symulacyjne oraz testy subiektywne, mające na celu ustalenie odpowiednich poziomów kwantyzacji parametrów modelu sinusoidalnego. W efekcie ustalono schemat kompresji z rozszerzaniem widma, przy pomocy którego możliwa jest kompresja sygnału dla prędkości bitowych rzędu 16 – 24 kb/s. Prędkość bitowa strumienia danych opisującego składowe tonalne jest rzędu 3 – 6 kb/s.

Do najważniejszych zalet proponowanej techniki rozszerzania widma należy:

- poprawna rekonstrukcja struktury harmonicznej oraz zmian na płaszczyźnie czasowo-częstotliwościowej składowych tonalnych dużych częstotliwości,
- uzyskanie wyższej jakości zdekodowanego sygnału w stosunku do kodeka MPEG-4 AAC HE.

Rozdział 4

Metoda rozszerzania widma wykorzystująca skalowanie częstotliwości

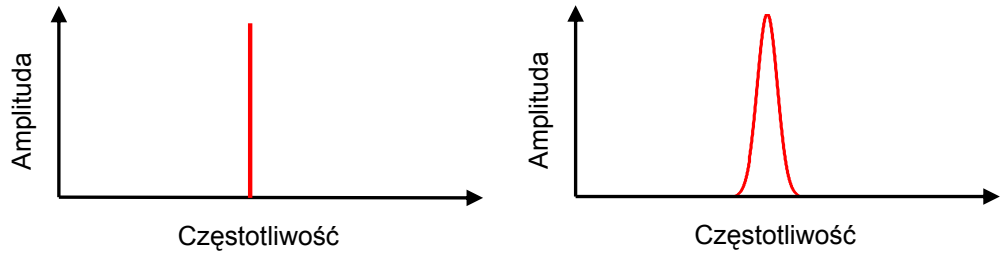
4.1. Wprowadzenie

W rozdziale przedstawiona zostanie nowa, autorska metoda kompresji sygnałów fonicznych z wykorzystaniem rozszerzania widma. Proponuje się metodę, która jest alternatywą dla replikacji widma (SBR) stosowanej w MPEG-4 AAC HE. Główna idea pracy dotyczy rekonstrukcji struktury tonalnej składowych dużych częstotliwości z wykorzystaniem informacji o częstotliwościach podstawowych wyznaczonych w koderze. Sygnał jest dekomponowany na grupy sygnałów wąskopasmowych, uzyskanych w procesie demodulacji przy wykorzystaniu chwilowych częstotliwości składowych tonalnych. Składowe dużych częstotliwości (HF) rekonstruowane są poprzez modulację sygnału wąskopasmowego z uwzględnieniem odpowiednio przeskalowanej częstotliwości chwilowej tego sygnału. Podejście to oferuje poprawną syntezę szybkozmiennych składowych tonalnych sygnału, jak również poprawną rekonstrukcję szeregów harmonicznych. Dla zachowania właściwej obwiedni widma decydującej o brzmieniu muzyki zastosowano również korekcję energii przeskalowanych składowych tonalnych. Technika ta stanowi oryginalny wkład autora rozprawy w rozwój metod rozszerzania widma. Część rezultatów wstępnych badań zaprezentowanych w rozdziale została zawarta we wcześniejszych publikacjach autora rozprawy [163].

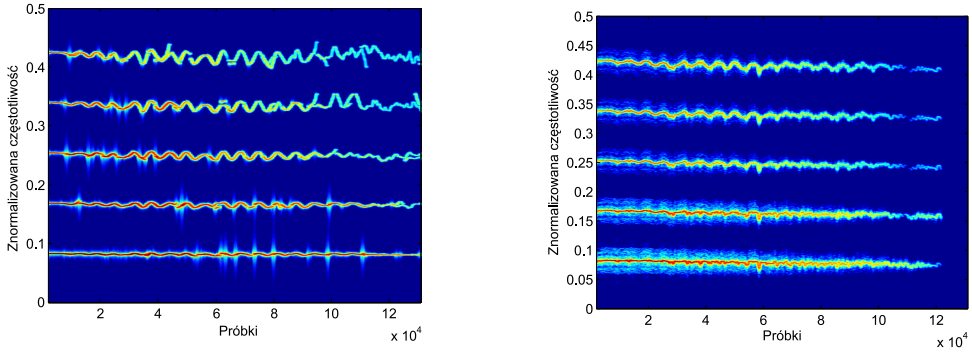
4.2. Koncepcja metody

Rzeczywiste sygnały (w tym muzyka) zawierają składowe pseudosinusoidalne, czyli sinusoidy przekształcone mikro-modulacjami AM i FM [85, 111, 115], które nadają sygnałowi charakterystyczną barwę. Mikro-modulacje mają częściowo charakter losowy, co powoduje, że widmo amplitudowe takiego sygnału jest szersze niż widmo pojedynczej sinusoidy i może być nawet interpretowane jako wąskopasmowy szum. Model parametryczny opisujący tego typu zjawiska powinien być w stanie reprezentować centralną częstotliwość oraz szerokość widma składowej tonalnej. Problem ten został już poruszony w pracach Fitza i Hakena [48, 49], w których zaprezentowano model sinusoidalny, w którym składowe sinusoidalne modulowane są wąskopasmowym szumem. W ten sposób możliwe jest kontrolowanie stopnia tonalności każdej składowej sinusoidalnej, co odbywa się przez zastosowanie zmiennej w czasie modulacji AM lub FM po stronie dekodera.

W dalszej części rozdziału przez składową tonalną rozumieć będziemy wąskopasmowy sygnał będący sinusoidą, której chwilowe parametry amplitudy i częstotliwości zawierają mikro modulacje AM i FM. Różnicę pomiędzy składową tonalną reprezentowaną przez pojedynczą sinusoidę oraz sygnał wąskopasmowy przedstawia rysunek 4.1.



(a) sinusoida (b) składowa wąskopasmowa



(c) vibrato - sygnał syntetyczny (d) vibrato - sygnał naturalny

Rysunek 4.1: Widmo sygnału sinusoidalnego a) i widmo składowej tonalnej b). c) i d) Spektrogramy sygnałów składających się odpowiednio z sygnałów z rysunków a) i b).

Dla sygnału wąskopasmowego reprezentującego składową tonalną można zdefiniować chwilową częstotliwość $f(t)$, fazę $\varphi(t)$ i amplitudę $A(t)$ [160]. Jeśli dla sygnału wąskopasmowego:

$$x(t) = A(t)\cos(\varphi(t)), \tag{4.1}$$

zapiszemy odpowiadający mu sygnał analityczny:

$$z(t) = x(t) + j\mathcal{H}\{x(t)\}, \tag{4.2}$$

gdzie $\mathcal{H}\{x(t)\}$ jest transformatą Hilberta sygnału $x(t)$. Wówczas:

$$z(t) = A(t)e^{j\varphi(t)}. \tag{4.3}$$

Amplitudę chwilową sygnału $x(t)$ możemy zapisać jako:

$$A(t) = |z(t)|. \tag{4.4}$$

Natomiast częstotliwość chwilowa sygnału $x(t)$ jest pochodną argumentu sygnału analitycznego:

$$f(t) = \frac{1}{2\pi}\frac{d}{dt}\arg\{z(t)\}. \tag{4.5}$$

Źródła dźwięku wielu naturalnych instrumentów rozpatrywać możemy jako rezonatory, które generują wąskopasmową energię, gdy są pobudzone przez losowy proces, np. podmuch powietrza, tarcie struny. Wygenerowany w ten sposób sygnał można interpretować jako sumę szeregów harmonicznych o wartościach chwilowych częstotliwości i amplitud, które zmieniają się losowo w niewielkim zakresie. W tego typu sygnałach, pojedyncze składowe identyfikowane są w dziedzinie krótkookresowego widma jako wąskopasmowe skupienie energii.

W pracy szeregiem harmonicznym $s_h(t)$ nazywa się sumę składowych pseudo-sinusoidalnych (czyli składowych sinusoidalnych o zmiennej w czasie amplitudzie i częstotliwości):

$$s_h(t) = \sum_{k=1}^{K} A_k(t)\cos\left(2\pi k\int_0^t f_0(\tau)d\tau + \varphi_k(0)\right), \tag{4.6}$$

gdzie k jest indeksem harmonicznej, K jest liczbą harmonicznych, $A_k(t)$ oznacza amplitudę chwilową, $f_0(\tau)$ jest częstotliwością chwilową, $\varphi_k(0)$ fazą początkową składowej pseudo-sinusoidalnej.

Przez *składową harmoniczną* określać będziemy k-tą składową szeregu harmonicznego i częstotliwości chwilowej będącej całkowitą wielokrotnością częstotliwości podstawowej $f_0(t)$.

Główna idea techniki prezentowanej w bieżącym rozdziale polega na odtworzeniu wąskopasmowych sygnałów reprezentujących składowe tonalne dużych częstotliwości, w taki sposób aby odtworzyć ich mikro-modulacje AM i FM. Można to zrobić przekształcając składowe małych częstotliwości poprzez skalowanie ich parametrów chwilowych ($A(t)$ i $f(t)$). Jeśli dla danego szeregu harmonicznego (4.6) częstotliwość chwilowa k-tej składowej harmonicznej wynosi:

$$f_k(t) = k \cdot f_0(t), \qquad k = 1, ..., K, \tag{4.7}$$

gdzie $f_0(t)$ jest chwilową częstotliwości podstawową szeregu harmonicznego, wówczas sygnał wąskopasmowy reprezentujący k-tą harmoniczną opisany jest wzorem:

$$x_k(t) = Re\{A_k(t)\, e^{j2\pi k f_0(t)}\}. \tag{4.8}$$

Dodatkowo wąskopasmowy sygnał reprezentujący składową zmieniającą się wraz z częstotliwością podstawową opisujemy:

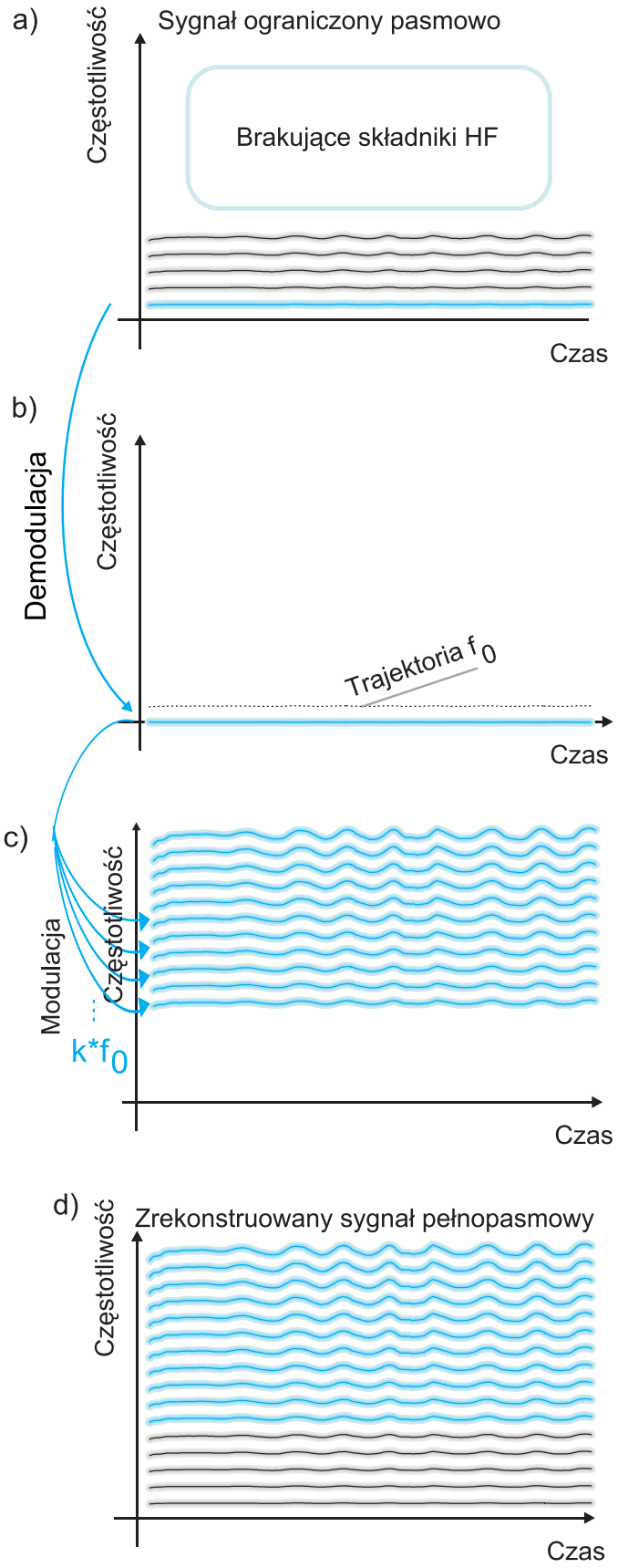
$$x_1(t) = Re\{A_1(t)\, e^{j2\pi f_1(t)}\}. \tag{4.9}$$

4.3. Analiza sygnału w dekoderze

W proponowanej technice koder estymuje wartości częstotliwości chwilowych $f(t)$ oraz amplitudy $A(t)$ danego szeregu harmonicznego. Zakodowane parametry przesyłane są do dekodera, w którym odbywa się proces analizy sygnału na podstawie informacji z kodera. Przykład procesu analizy sygnału skalowania częstotliwości (wzór (4.9)) pokazany jest na rysunku 4.2. Analiza sygnału polega na wyodrębnieniu sygnału $x_0(t)$ i na jego podstawie rekonstrukcji składowych dużych częstotliwości. Pierwszym krokiem algorytmu jest demodulacja sygnału wejściowego $x(t)$ zgodnie z przebiegiem częstotliwości podstawowej $f_0(t)$ (rysunek 4.2b). Następnie, sygnał poddany jest filtracji dolnoprzepustowej, w celu otrzymania reprezentacji sygnału $x_0(t)$ w paśmie podstawowym. Otrzymany sygnał $x_0(t)$ poddany jest modulacji zgodnie z częstotliwością $f_0(t)$ przez co uzyskujemy wąskopasmową reprezentację składowej tonalnej. Następnie częstotliwość sygnału $x_0(t)$ jest przeskalowywana kolejnymi, całkowitymi wielokrotnościami częstotliwości f_0 zgodnie ze wzorem (4.9). W wyniku modulacji otrzymujemy sygnał wąskopasmowy $x_{HT}(t)$ będący sumą składowych tonalnych dużych częstotliwości (rysunek 4.2c). Powyższa procedura

powtarzana jest dla każdej trajektorii f_0. Sygnał przedstawiony na rysunku 4.2d stanowi sumę wąskopasmowych składowych tonalnych $x_k(t)$. Wszystkie składowe $x_k(t)$ mają tą samą energię, aby więc poprawnie odtworzyć sygnał $x_{HT}(t)$ należy przeskalować amplitudową obwiednię widma otrzymanego sygnału. Proces skalowanie obwiedni zostanie przedstawiony w dalszej części pracy.

Powyżej zaprezentowano ogólną koncepcję kodeka, który odtwarza składowe sinusoidalne dużych częstotliwości wykorzystując algorytm skalowania częstotliwości. Wyniki badań przedstawione zostaną w rozdziale 5. W dalszych punktach rozdziału 4 zostanie opisana konstrukcja kodeka implementująca powyżej opisaną ideę.



Rysunek 4.2: Proces analizy i syntezy sygnału w proponowanej technice BWE: a) spektrogram sygnału ograniczonego pasmowo rekonstruowanego w dekoderze, b) wąskopasmowa składowa tonalna, c) zrekonstruowane składowe tonalne HF, d) zrekonstruowany sygnał stanowiący sumę składowych HF oraz sygnału zdekodowanego w dekoderze podstawowym.

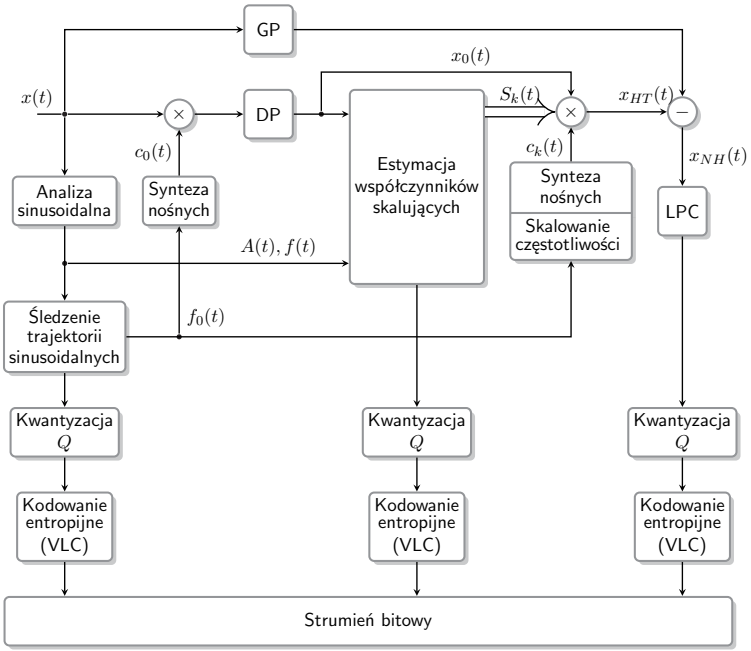
4.4. Opis proponowanej techniki BWE

4.4.1. Wprowadzenie

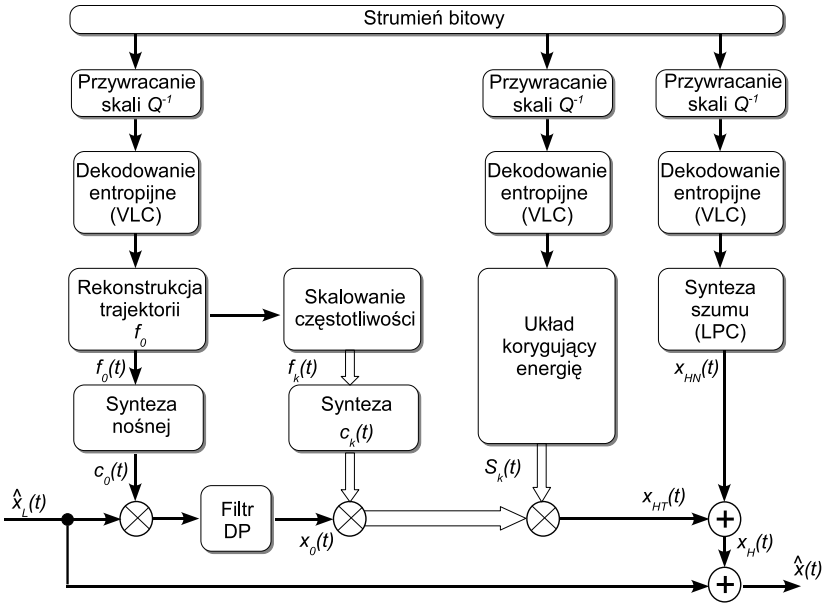
W proponowanej technice wartość granicy podziału pomiędzy małymi i dużymi częstotliwościami $f_{BWE} = f_{SBR}$ i zależy od docelowej prędkości bitowej zgodnie z tabelą 2.1. Podczas badań, do kodowania małych częstotliwości zastosowano koder podstawowy MPEG-4 AAC HE [2, 67].

Po stronie kodera (rysunek 4.3) sygnał wejściowy $x(t)$ dekomponowany jest do postaci sumy sygnałów wąskopasmowych $x_{HT}(t)$. Sygnały te reprezentują poszczególne składowe tonalne, wchodzące w skład danego szeregu harmonicznego i zmieniające się zgodnie z częstotliwością podstawową f_0. Pierwszym etapem analizy w koderze jest estymacja częstotliwości podstawowych szeregów harmonicznych $f_0(t)$. Jednocześnie algorytm analizy sinusoidalnej wyznacza parametry chwilowe składowych tonalnych $A(t)$ i $f(t)$. Na podstawie chwilowych częstotliwości podstawowych $f_0(t)$ wyznaczana jest demodulująca fala nośna $c_0(t)$. Sygnał będący wynikiem demodulacji $x(t)$ i $c_0(t)$ poddawany jest filtracji dolnoprzepustowej w filtrze oznaczonym DP na rysunku 4.3. Na podstawie częstotliwości $f_0(t)$ syntezowane są nośne modulujące $c_k(t)$, których częstotliwości są przeskalowane zgodnie ze wzorem (4.7). Dodatkowo na podstawie parametrów $A(t)$ i $f(t)$ wyznaczana jest funkcji opisującej widmo amplitudowe $\tilde{A}_m$ dla m-tej ramki analizowanego sygnału $x(t)$. Spróbkowana w punktach kf_0 obwiednia amplitudowa $\tilde{A}_m$ tworzy zbiór współczynników skalujących $S_k(t)$ dla k-tej harmonicznej. W wyniku modulacji sygnałów $x_0(t)$ i $c_k(t)$ poddanych skalowaniu $S_k(t)$ otrzymujemy sygnał składowych tonalnych dużych częstotliwości $x_{HT}(t)$ powyżej częstotliwości f_{BWE}. Parametry opisujące współczynniki skalujące $S_k(t)$ oraz trajektorie $f_0(t)$ podlegają kodowaniu i przesyłane są do dekodera.

Dodatkowym elementem w proponowanej technice jest wyznaczenie składowej szumowej dużych częstotliwości $x_{HN}(t)$. Sygnał $x_{HN}(t)$ uzyskiwany jest poprzez usunięcie składowych tonalnych $x_{HT}t$ z sygnału wejściowego stosując algorytm tłumienia widma (opisany w punkcie 3.4.3). Otrzymany sygnał $x_{HN}(t)$ traktowane jest jak szum i poddany kodowaniu LPC.



Rysunek 4.3: Schemat kodera



Rysunek 4.4: Schemat dekodera

Po stronie dekodera (rysunek 4.4) składowe dużych częstotliwości rekonstruowane są poprzez skalowanie częstotliwości chwilowych sygnału $x_0(t)$ uzyskanego poprzez demodulację z sygnału małych częstotliwości $\hat{x}_L(t)$. Proces skalowania częstotliwości realizowany jest w podobny sposób jak w koderze.

Strumień bitowy generowany przez proponowany koder uwzględnia dane trajektorii częstotliwości podstawowych $f_0(t)$, współczynniki skalujące obwiednię widma $S_k(t)$ oraz współczynniki predyktora LP opisujące składową szumową dużych częstotliwości.

4.4.2. Estymacja parametrów

W proponowanej technice BWE składowe składowe tonalne podlegają procesowi detekcji opisanemu w rozdziale 3.4.2. Dodatkowo śledzone i łączone w trajektorie są częstotliwości f_0, które następnie są kodowane różnicowo i przesyłane do dekodera.

Jednym z podstawowych etapów w proponowanej technice BWE jest estymacja częstotliwości podstawowych f_0 sygnału. Częstotliwości f_0 wykorzystywane są w procesie skalowania częstotliwości (modulacji) oraz podczas wyznaczania szeregów harmonicznych w sygnale wejściowym. Mając daną częstotliwość podstawową można wyznaczyć częstotliwości kolejnych składowych harmonicznych. Przy czym zakładamy, że częstotliwości harmonicznych są całkowitymi wielokrotnościami f_0.

Pierwszym krokiem analizy sygnału składowych HF jest zastosowanie modelu harmonicznego [113, 140], który umożliwia detekcję szeregów harmonicznych w badanym sygnale. Szeregi harmoniczne wykorzystywane są następnie do poprawnej rekonstrukcji obwiedni składowych HF. Sygnał wejściowy $x(t)$ podlega analizie w następujących po sobie oknach czasowych, z uwzględnieniem odpowiedniej rozdzielczości czasowo-częstotliwościowej. W tym celu wykorzystane jest przekształcenie STFT o długości okna $N = 2048$ i skoku H=512. Wartości amplitud $A(t)$ i częstotliwość $f(t)$ estymowane są, w każdym oknie analizy, na podstawie algorytmu QIFFT [4] - opisanego w rozdziale 3.4.2. Następnie, parametry sinusoid łączone są w trajektorie sinusoidalne.

Dane uzyskane w wyniku analizy harmonicznej - amplitudy i częstotliwości sinusoid oraz częstotliwości f_0 - wykorzystywane są do aproksymacji obwiedni widma $\tilde{A}_m$ sygnału w danym oknie analizy m. Następnie, na ich podstawie wyznaczane są współczynniki skalujące obwiednię energetyczną składowych HF.

W proponowanej technice BWE analiza sinusoidalna wykorzystywana jest do wyznaczenia wartości chwilowych amplitud $A(t)$ i częstotliwości $f(t)$ składowych tonalnych. Dane te następnie wykorzystywane są do:

- estymacji obwiedni widma, wyznaczanej wspólnie dla wszystkich szeregów harmonicznych,
- określenia częstotliwości podstawowej f_0 dla powyższych szeregów harmonicznych,
- wyznaczenia trajektorii sinusoidalnej, która zostanie wykorzystana w procesie skalowania i modulacji.

4.4.3. Skalowanie częstotliwości

Głównym zadaniem procesu skalowania jest wyodrębnienie wąskopasmowego sygnału $x_0(t)$ z sygnału wejściowego $x(t)$ oraz rekonstrukcja składowych tonalnych HF. Proces skalowania odbywa się poprzez modulację sygnału wąskopasmowego $x_0(t)$ częstotliwością chwilową wybranej trajektorii sinusoidalnej f_{mod}. Skutkuje to przesunięciem widma sygnału $x_0(t)$ o wartość częstotliwości f_{mod}. Częstotliwość ta jest całkowitą wielokrotnością f_0.

$$f_{mod} = kf_0, \tag{4.10}$$

gdzie k jest indeksem harmonicznej. Należy podkreślić, że współczynnik k dla naturalnych sygnałów nie musi być liczbą całkowitą. W pracy, dla uproszczenia rozważań, nie uwzględniono współczynnika nieharmoniczności (punkt 2.3.1) w procesie rekonstrukcji sygnału.

Poniżej opisana zostanie procedura skalowania częstotliwości, gdzie częstotliwość $f_{mod} = f_0$.

- Częstotliwość trajektorii f_0 wyznaczona przez analizę sinusoidalną jest spróbkowana w węzłach wyznaczonych przez zastosowaną długość okna analizy oraz jego skok. Aby uzyskać przesunięcie w dziedzinie częstotliwości, które podąża za zmianami częstotliwości podstawowej każdego szeregu harmonicznego, należy najpierw poddać interpolacji punkty częstotliwości f_0. Na podstawie uzyskanej trajektorii możliwa jest synteza sygnału nośnej demodulującej $c_0(t)$.

- Funkcja demodulująca $c_0(t)$ jest zespolonym przebiegiem harmonicznym o zmiennej częstotliwości:
$$c_0(t) = \exp\left(-j2\pi \int_0^t f_0(t)dt\right). \tag{4.11}$$

- Wynik demodulacji $x_0(t)$ poddany zostaje filtracji dolnopasmowej filtrem $h_{DP}(t)$. Otrzymany w ten sposób sygnał wąskopasmowy $x_0(t)$ charakteryzuje się wolno zmienną częstotliwość chwilową zbliżoną do 0. Sygnał ten reprezentuje pojedynczą składową tonalną rozpatrywanego szeregu harmonicznego w paśmie podstawowym.
$$x_0(t) = (c_0(t) \cdot x(t)) * h_{DP}(t), \tag{4.12}$$
gdzie częstotliwość graniczna filtru dolnoprzepustowego DP
$$f_{DP} = \frac{\min f_0(t)}{2}. \tag{4.13}$$

Sygnał $x_0(t)$ podlega filtracji w każdym oknie analizy m. Częstotliwość graniczna filtru powinna zmieniać się wraz z szerokością pasma składowej tonalnej. Dla uproszczenia w pracy zastosowano filtr, który w pewnym stopniu ma zabezpieczać przed nakładaniem się widm sąsiednich składowych tonalnych.

- Tonalne składowe dużych częstotliwości rekonstruowane są poprzez modulację sygnału $x_0(t)$ zestawem funkcji modulujących, których częstotliwości są całkowitymi wielokrotnościami $f_0(t)$.

$$c_k(t) = \exp\left(j2\pi \int_0^t f_k(t)dt\right) \qquad k = k_{start}, \ldots, k_{end}. \tag{4.14}$$

Sygnały $c_k(t)$ wyznaczane są dla pozostałych harmonicznych szeregu. Proces skalowania musi być powtórzony dla wszystkich szeregów i odpowiadających im trajektorii f_0 (f_{mod}).

$$x_{HT}(t) = \sum_{k=k_{start}}^{k_{end}} c_k(t) \cdot x_0(t), \tag{4.15}$$

gdzie:

$$k_{start} = \left\lceil \frac{f_{BWE}}{\max\left(f_0(t)\right)} \right\rceil, \tag{4.16}$$

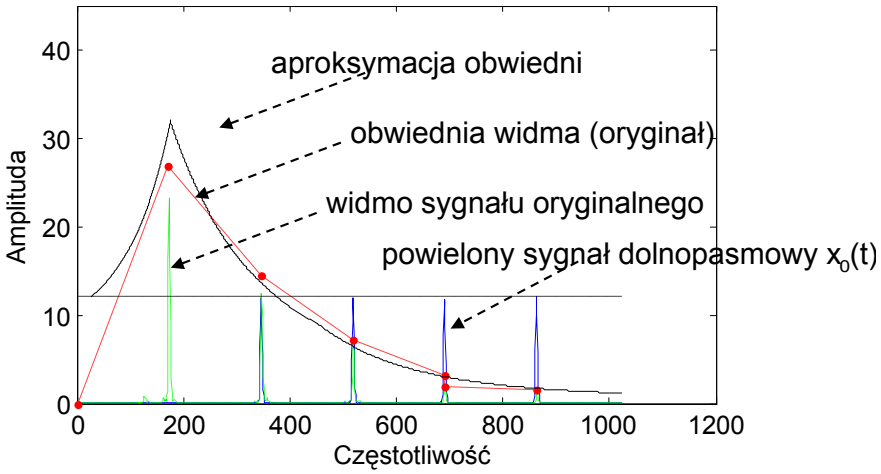
$$k_{end} = \left\lfloor \frac{0{,}5 \cdot f_s}{\max\left(f_0(t)\right)} \right\rfloor. \tag{4.17}$$

gdzie f_s jest częstotliwością próbkowania sygnału.

4.4.4. Skalowanie amplitudowej obwiedni widma

Otrzymany w drodze skalowania sygnał $x_{HT}(t)$ posiada strukturę harmoniczną zbliżoną do oryginału, co zostało zaprezentowane na rysunku 4.2. Poszczególne zrekonstruowane składowe sygnału $x_{HT}(t)$ mają taką samą energię równą energii sygnału $x_0(t)$. Obwiednia widma sygnału oryginalnego G_m oraz sygnału przeskalowanego są różne. Istotne jest więc znalezienie odpowiedniej reprezentacji amplitudowej obwiedni widma. W proponowanej technice wykorzystano do tego celu współczynniki skalujące S_k, które obliczane są po stronie kodera i przesyłane do dekodera.

Współczynniki skalujące wyznaczane są w kolejnych, następujących po sobie oknach analizy, zgodnie z zastosowanym algorytmem STFT. Rysunek 4.5 ilustruje proces estymacji współczynników skalujących. Pierwszym krokiem jest wyznaczenie obwiedni widma



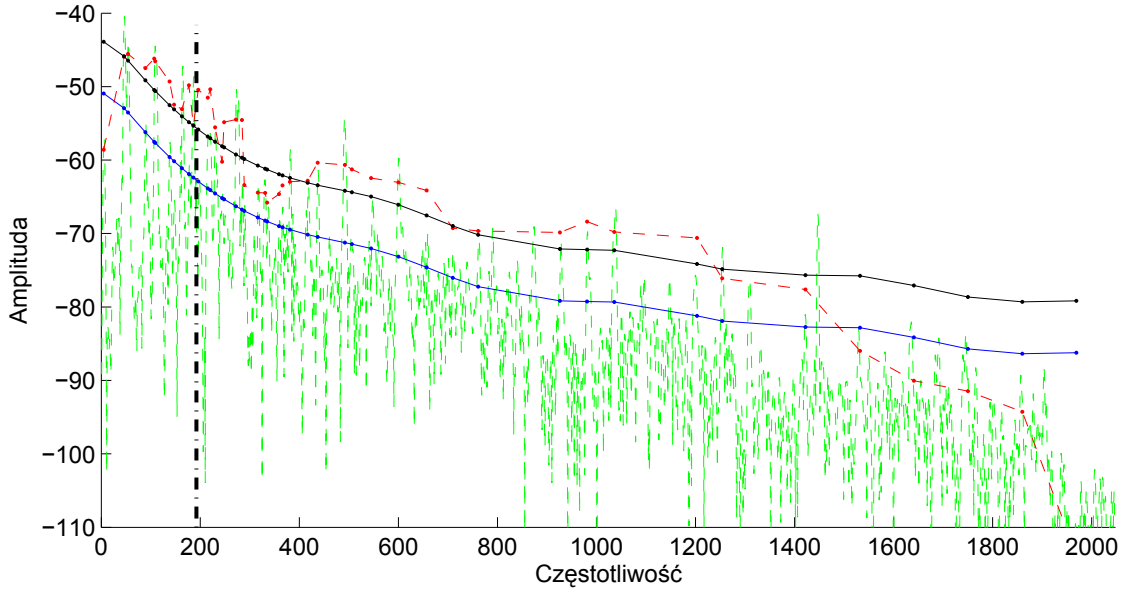
Rysunek 4.5: Wyznaczanie współczynników skalujących - przykład obwiedni energetycznej dla pojedynczego szeregu harmonicznego w obrębie jednej ramki analizy.

G_m sygnału oryginalnego w danym oknie analizy m (rysunek 4.5 - czerwona linia). W tym celu zastosowano metodę predykcji liniowej LPC (ang. *Linear Predictive Coding*), w której dla każdego krótkookresowego widma (STFT) wyznaczane są współczynniki predyktora LPC. Następnie charakterystyka amplitudowa predyktora $|G_{LPC_m}(f)|$ jest próbkowana w miejscach $p = kf_0$, gdzie p jest częstotliwością składowej tonalnej, $k = 1, \ldots, k_{end}$.

Otrzymujemy współczynniki skalujące poszczególne składowe tonalne p w danym oknie m i oznaczamy jako funkcję skalującą widmo $|G_{LPC_m}(p)|$. W pracy zastosowano filtr LPC 10 rzędu. Współczynniki filtru oraz wzmocnienie podlegają kompresji i są przesyłane do dekodera.

Uzyskana, w drodze przekształcenia LPC, transmitancja $G_{LPC}(f)$ oddaje jedynie przybliżony charakter obwiedni widma sygnału oryginalnego. Związane jest to z zastosowanym niskim rzędem filtru predykcji, co z kolei podyktowane jest ograniczoną pulą bitów niezbędną do zakodowania współczynników filtru. Problem ten przedstawiony jest na rysunku 4.6. Niebieska linia przedstawia obwiednię LPC, natomiast czerwona linia - połączone wartości amplitud sinusoid $A_m(p)$.

Charakterystyka amplitudowa filtru $|G_{LPC}(f)|$ nie może być interpretowana jako obwiednia widma, ponieważ jaj charakterystyka nie przechodzi przez maksima widma amplitudowego. $|G_{LPC}(f)|$ przybliża widmo amplitudowe w sensie średniokwadratowym. Natomiast kształt charakterystyki amplitudowej $|G_{LPC}(f)|$ jest zbliżony do obwiedni widma. Bazując na tym spostrzeżeniu w pracy zastosowano dodatkowy współczynnik wzmocnienia



Rysunek 4.6: Przykład skalowania energii dla pojedynczego okna analizy. Linia o kolorze niebieskim reprezentuje $G_{LPC_m}(f)$, linia o kolorze czarnym $G_D G_{LPC_m}(f)$, a przerywana linia o kolorze czerwonym oznacza połączone wartości amplitud chwilowych sinusoid wyznaczonych przez analizę sinusoidalną.

G_D określający przybliżone zależności pomiędzy charakterystyką amplitudową $|G_{LPC}(f)|$, a rzeczywistą obwiednią sygnału G_m.

Zadaniem techniki BWE jest poprawna rekonstrukcja składowych powyżej częstotliwości f_{BWE}. Z tego powodu obwiednia widma G_m powinna odpowiadać poziomowi amplitud powyżej f_{BWE}. Spełnienie tego warunku jest trudne ze względu na ograniczoną pulę bitów, a przez to, niski rząd predyktora $G_{LPC_m}(f)$. Z tego powodu, zastosowano ograniczony zakres częstotliwości $f_{bw} \in < f_{BWE}, f_s/4 >$, w którym wartości obwiedni widma $|G_{LPC}(z)|$ dopasowywane są do obwiedni widma sygnału oryginalnego G_m. Zakres f_{bw} został dobrany doświadczalnie, biorąc pod uwagę ograniczoną pulę bitów oraz fakt, że percepcyjnie najistotniejsze składowe sygnału znajdują się w początkowym zakresie dużych częstotliwości [161].

Na dokładność aproksymacji sygnału za pomocą metody LPC ma wpływ rząd predyktora, który w proponowanym rozwiązaniu wynosił 10. Parametr ten dobrano na podstawie nieformalnych testów. Celem doboru rzędu predyktora było uzyskanie możliwie „gładkiej" funkcji skalującej, przybliżającej wierzchołki obwiedni widma sygnału. Z drugiej strony dobór rzędu predykcji podyktowany jest również ograniczoną pulą bitów dostępną dla techniki rozszerzania widma.

Aby wyznaczyć współczynnik wzmocnienia G_D należy najpierw wyznaczyć za pomocą modelowania sinusoidalnego wartości amplitud $A_m(p)$ w danym oknie analizy. Przez A_{env} oznaczymy zbiór amplitud uzyskanych podczas analizy sinusoidalnej, wówczas możemy zapisać:

$$A_{env} = \bigcup_{i \in I} A_{env,i}, \tag{4.18}$$

gdzie $A_{env,i}$ jest zbiorem amplitud pojedynczego szeregu harmonicznego, i jest indeksem i-tego szeregu harmonicznego, I zbiorem indeksów występujących w sygnale szeregów harmonicznych. Wartości amplitud A wyznaczone zostały w bloku analizy sinusoidalnej. Dla wielu nagrań muzycznych dane pochodzące z analizy sinusoidalnej są jednak niekompletne. Z tego powodu w pracy zastosowano dodatkowy zbiór amplitud A_{orig}, który zawiera brakujące parametry składowych tonalnych odczytanych bezpośrednio z widma amplitudowego.

Zbiór amplitud dla pojedynczego szeregu harmonicznego opisuje wzór:

$$A_{env,i} = A_{sin,i}(m) \cup A_{orig,i}(m), \tag{4.19}$$

gdzie:

m – indeks okna analizy,

$A_{sin,i}(m) = A_s, s \in S$ – zbiór amplitud składowych tonalnych otrzymanych z modelu sinusoidalnego dla i-tego szeregu harmonicznego,

S – zbiór indeksów składowych harmonicznych uzyskanych podczas analizy sinusoidalnej,

$A_{orig,i}(m) = A_n, n \in N$ – uzupełniający zbiór brakujących amplitud uzyskanych z widma amplitudowego w procesie próbkowania widma dla $F_{sin,i}$,

N – zbiór indeksów brakujących składowych tonalnych, stanowiących wielokrotności częstotliwości podstawowej f_0.

$$N = K_i \setminus S, \tag{4.20}$$

gdzie:

$K_i = \{1, 2, \ldots, k_{end}\}$ - zbiór indeksów wszystkich sinusoid i-tego szeregu harmonicznego.

Biorą pod uwagę zbiór współczynników $|G_{LPC_m}(p)|$ oraz zbiór amplitud A_{env}, można wyznaczyć współczynnik wzmocnienia G_D amplitud $|G_{LPC_m}(p)|$, dla ograniczonego

zakresu częstotliwości f_{bw}:

$$G_{D(dB)} = \frac{1}{L} \sum_{pp \in P(f_{bw})} \left(20 \log_{10} \left(A_{env_m}(pp)\right) - 20 \log_{10} \left(G_{LPC_m}(pp)\right)\right), \tag{4.21}$$

gdzie:

P – zbiór indeksów składowych tonalnych,

$P(f_{bw})$ – zbiór indeksów składowych tonalnych ograniczonego do zakresu częstotliwości f_{bw},

L – liczba elementów zbioru $P(f_{bw})$.

Wartość współczynnika skalującego obwiednię widma ulega dodatkowemu tłumieniu w zależności od wartości współczynnika tonalności dla danego okna analizy. Operacja tłumienia istotna jest zwłaszcza w okolicach transjentów, które przeważnie mają charakter szerokopasmowego impulsów. Demodulacja i skalowanie częstotliwości sygnału wąskopasmowego $x_0(t)$ mogłyby doprowadzić do niepoprawnej rekonstrukcji fragmentów widma zawierających transjenty. W efekcie mogłyby zostać wygenerowane dodatkowe, niepożądane składniki tonalne. Aby zapobiec tego typu zniekształceniom, fragmenty zrekonstruowanego sygnału składowych tonalnych zostają stłumione proporcjonalnie do stopnia tonalności danego fragmentu sygnału. Współczynnik tłumienia G_T bazuje na współczynniku tonalności i wyznaczony jest na podstawie wzoru (3.14).

Przez $G_m(p)$ oznaczymy wartości współczynników skalujących uzyskanych poprzez próbkowanie funkcji G_m w miejscach występowania składowych tonalnych o indeksach p. Wartość funkcji skalującej obwiednię widma dla danego okna analizy wyznaczana jest według wzoru:

$$G_m(p) = G_T(m) \cdot G_D(m) \cdot |G_{LPC_m}(p)|, \tag{4.22}$$

gdzie $G_m(p)$ jest macierzą o rozmiarach $m \times p$. Wartości w wierszach macierzy $G_m(p)$ podlegają interpolacji, na podstawie której otrzymujemy p wolno-zmiennych w czasie sygnałów $S_{k,i}$, reprezentujących funkcje skalujące obwiednię widma kolejnych składowych tonalnych.

Na podstawie wzoru (4.15) oraz funkcji skalującej obwiednię widma $S_{k,i}$ można określić wzór opisujący składowe tonalne HF:

$$x_{HT}(t) = \sum_{i \in I} \sum_{k=k_{start}}^{k_{end}} S_{k,i} \cdot c_{k,i}(t) \cdot x_{0,i}(t). \tag{4.23}$$

4.5. Rekonstrukcja sygnału

W dekoderze (rysunek 4.4) składowe dużych częstotliwości rekonstruowane są z ograniczonego pasmowo sygnału wąskopasmowego $x_0(t)$ oraz na podstawie dodatkowych parametrów ze strumienia (tabela 4.2).

W pierwszej kolejności dekodowane są częstotliwości podstawowe $f_0(t)$ każdego szeregu harmonicznego. Następnie trajektorie f_0 wykorzystywane są do demodulacji składowych tonalnych na podstawie sygnału otrzymanego z dekodera podstawowego (rysunek 4.2). Proces skalowania jest podobny do zastosowanego w koderze (wzór (4.15)), z tą różnicą, że:

$$\hat{x}_{HT}(t) = 2Re\left\{\sum_{k=k_{start}}^{k_{end}} c_k^*(t)\cdot\hat{x}_0(t)\right\}, \tag{4.24}$$

gdzie $\hat{x}_0(t)$ to sygnał wąskopasmowy wyznaczony na podstawie równania (4.12), $\hat{c}_k(t)$ to nośna modulująca, która różni się od nośnej z równania (4.14) przeciwnym znakiem argumentu funkcji wykładniczej. Z całego sygnału brana jest tylko część rzeczywista. W celu przywrócenia energii z oryginalnego sygnału wynik mnożony jest przez 2.

Należy podkreślić, że sygnał $x_0(t)$ uzyskany po stronie dekodera oraz wartości f_0 zostały poddane kwantyzacji. Mamy więc do czynienia z błędem kwantyzacji. Błąd ten można zminimalizować poprzez odpowiednie określenie częstotliwości granicznej dolnoprzepustowego filtru adaptacyjnego $h_{DP}(t)$ wykorzystywanego do uzyskania sygnału $x_0(t)$. W pracy zastosowano uproszczoną regułę wyznaczania częstotliwości granicznej filtru DP opisaną wzorem (4.13).

4.6. Kodowanie parametrów

W proponowanej technice BWE kodowane są następujące parametry:

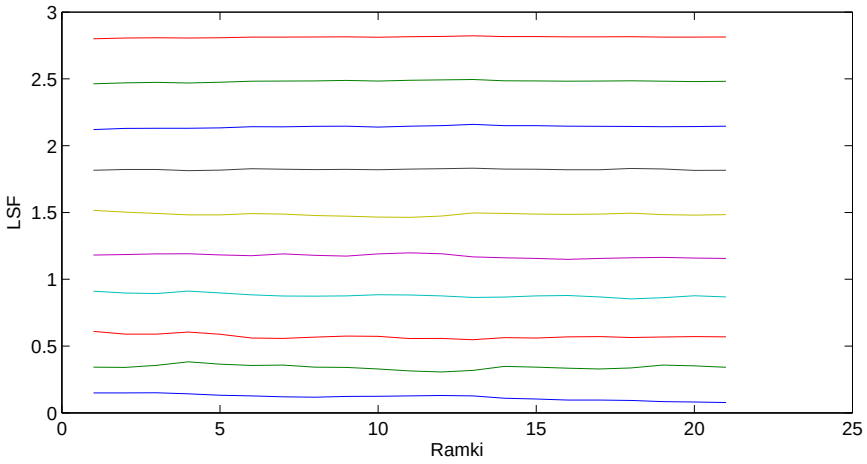
- dla składowych tonalnych:
  - trajektorie częstotliwości chwilowej f_0,
  - współczynniki skalujące obwiedni widma $G_{LPC_m}(p)$,
- opcjonalne dla składowej szumowej:
  - współczynniki LPC reprezentowane przez częstotliwości linii widmowych - LSF (ang. *Line Spectral Frequencies*).

4.6.1. Kodowanie trajektorii częstotliwości podstawowych

Do kodowania trajektorii f_0 wykorzystano kodowanie różnicowe z predyktorem adaptacyjnym ADPCM (opisane w dodatku B), kwantyzację równomierną oraz kodowanie Huffmana. Zastosowano model predykcji liniowej w wariancie Burga [98]. Rząd predykcji oraz ilość poprzednich próbek wykorzystanych do estymacji współczynników predyktora zależy od rozdzielczości czasowej kodowanych trajektorii sinusoidalnych. Podczas badań parametry predykcji ustalono zgodnie z zaleceniami autorów metody śledzenia trajektorii sinusoidalnych na podstawie predykcji liniowej [98]. Rozdzielczość czasowa analizy wynosi 11,6 ms, gdzie zastosowano predyktor 6-tego rzędu operującego na 20 poprzednich próbkach trajektorii sinusoidalnej. Prędkość bitowa uzyskana po kodowaniu trajektorii f_0 wynosi około 0,1 kb/s (tabela 4.2).

4.6.2. Kodowanie współczynników skalujących obwiednię widmową

Jak zostało opisane w punkcie 4.4.4, obwiednia widmowa aproksymowana jest przy pomocy przekształcenia LPC. W pracy współczynniki LPC podlegają przekształceniu na częstotliwości linii widmowych LSF (ang. *Line Spectral Frequencies*) [71, 90]. Sposób przekształcenia współczynników LPC na LSF opisany został w pracy [114].



Rysunek 4.7: Przykład zmiany współczynników LSF w kolejnych oknach analizowanego sygnału.

Współczynniki LSF są mniej podatne na szum kwantyzacji w stosunku do LPC. Dodatkowo współczynniki LSF zmieniają się w niewielkim zakresie w kolejnych oknach analizy

(rysunek 4.7). Umożliwia to efektywne kodowanie LSF za pomocą techniki ADPCM - kodowany jest błąd predykcji w kolejnych oknach analizy.

4.6.3. Reprezentacja składowej szumowej

Składowa szumowa stanowi różnicę pomiędzy sygnałem oryginalnym $x(t)$, a zrekonstruowanymi składowymi tonalnymi HF $x_{HT}(t)$, co możemy zapisać jako:

$$x_{HN}(t) = x(t) - x_{HT}(t). \tag{4.25}$$

Operacja uzyskania składowej szumowej $x_{HN}(t)$ realizowana jest poprzez mechanizm tłumienia widmowego, opisanego w punkcie 3.4.3. Zakłada się, że sygnał residuum $x_{HN}(t)$ w głównej mierze zawiera szum. Otrzymany sygnał podlega modelowaniu i kodowaniu szumowemu za pomocą metody LPC dla kolejnych okien analizy wyznaczonych za pomocą algorytmu STFT. Uzyskane parametry predyktora służą do generowania szumu syntetycznego po stronie dekodera.

Filtr LPC pobudzany jest szumem białym $w(n)$ o zerowej wartości średniej i wariancji σ_W^2 generuje sygnał $\hat{x}_{HN}(n)$ opisany wzorem:

$$\hat{x}_{HN}(n) = \sum_{i=1}^{M} a_i \hat{x}_{HN}(n-i) + G_{HN} w(n), \tag{4.26}$$

gdzie: G_{HN} jest wzmocnieniem szumu białego.

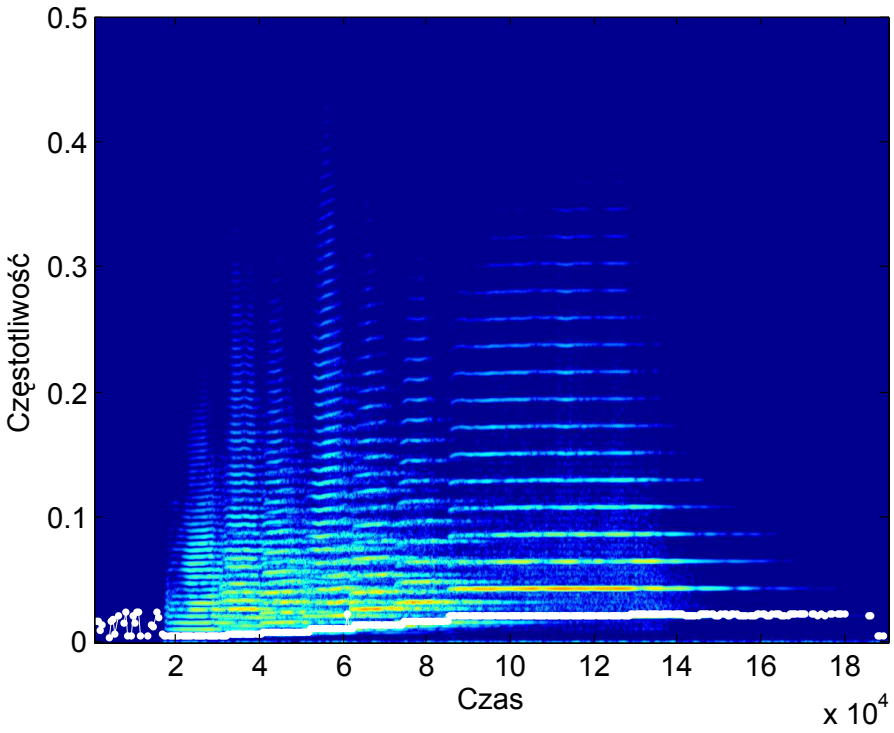
Obwiednia widmowa szumu również reprezentowana jest przez model LPC, możliwie niskiego rzędu, w tym wypadku 5-tego rzędu. Współczynniki filtru syntezy LPC zakodowane są w strumieniu w często stosowanej postaci PARCOR, której zaletą jest mała podatność na kwantyzację i łatwo uzyskiwalna gwarancja stabilności filtru poprzez zebezpieczenie, że ich wartość bezwzględna nie osiągnie 1 (w ten sposób nawet przy bardzo silnej kwantyzacji nie istnieje niebezpieczeństwo utraty stabilności filtru syntezy). Kwantyzacja współczynników PARCOR odbywa się w skali logarytmicznej, z rozdzielczością 4 bitów. Prędkość bitowa generowanego dla składowej szumowej strumienia wynosi około 0,4 kb/s.

4.7. Pozostałe bloki funkcjonalne kodera

4.7.1. Estymacja częstotliwości podstawowych f_0

Częstotliwości podstawowe szeregów harmonicznych f_0 wykorzystywane są w procesie skalowania częstotliwości (punkt 4.4.3) w celu uzyskania sygnału wąskopasmowego $x_0(t)$ reprezentującego składową tonalną.

W implementacji proponowanej techniki BWE wykorzystano algorytm detekcji f_0 zaproponowany przez Cheveigne i Kawaharę [29]. Wyjście detektora f_0 podlega operacji śledzenia, w którym wykorzystano algorytm śledzenia składowych sinusoidalnych. Zastosowanie tego typu algorytmu pozwala na uzupełnienie brakujących punktów f_0 (powstałych np. w wyniku błędów oktawowych estymacji f_0), wartościami „zombie". Trajektoria $f_0(m)$, wyznaczona dla chwil czasu m, jest następnie nadpróbkowana wykorzystując funkcje sklejane, dzięki czemu wartości trajektorii f_0 są dostępne dla każdej chwili czasu. Przykład działania algorytmu detekcji f_0 oraz śledzenia został przedstawiony na rysunku 4.8.



Rysunek 4.8: Wykrywanie częstotliwości podstawowej

4.7.2. Wybór trajektorii dla procesu modulacji

Jak zostało opisane w poprzednim punkcie, sygnał $x_o(t)$ można wyznaczyć poprzez demodulacje sygnału według częstotliwości podstawowej f_0. Tego typu podejście nie zawsze się jednak sprawdza ze względu na błędną estymację f_0. Błędy mogą powstawać na skutek interferencji składowej tonalnej o częstotliwości f_0 z innymi składowymi sygnału. Składowa o częstotliwości f_0 może również być sygnałem o bardzo małej energii, co dodatkowo utrudnia estymację częstotliwości chwilowej.

Jeśli estymacja częstotliwości będzie obarczona błędem oktawowym, wówczas dekoder zrekonstruuje składowe o częstotliwościach będących wielokrotnością wybranej składowej. Oznacza to, że odtworzony szereg harmoniczny będzie niekompletny. W efekcie zmieni się barwa oraz głośność postrzeganego dźwięku. Tego typu błędy nie są jednak bardzo dokuczliwe dla słuchaczy. Odtworzony w ten sposób sygnał będzie postrzegany podobnie jak sygnały ze stłumionymi wyższymi składowymi lub sygnały ograniczone pasmowo.

Poważniejszym problemem są błędy estymacji wynikające z niepoprawnego śledzenia trajektorii f_0. Wyższe harmoniczne zrekonstruowanego w ten sposób sygnału nie będą miały zachowanych poprawnych relacji częstotliwościowych w stosunku do składowych LF. Spowoduje to powstanie zniekształceń zbliżonych do generowanych przez techniki rozszerzania widma wykorzystujące replikację (punkt 2.6.3).

Z tego powodu w pracy zaproponowano dodatkowy, alternatywny algorytm wyznaczania częstotliwości wykorzystywanej w procesie demodulacji f_{mod} - wzór (4.10) w stosunku do bezpośredniej estymacji f_0. W tej metodzie trajektoria sinusoidalna o zmiennej w czasie wartości częstotliwości f_{mod} wyznaczana jest na podstawie częstotliwości podstawowej f_0 oraz trajektorii sinusoidalnych uzyskanych w procesie analizy sinusoidalnej i śledzenia. Algorytm opisujący proces detekcji f_{mod} składa się z następujących kroków:

- Trajektorie sinusoidalne grupowane są według wielokrotności częstotliwości f_0. Otrzymujemy w ten sposób *obiekt harmoniczny* zawierający składową sinusoidalną o częstotliwości podążającej za f_0 oraz składowe o częstotliwości będącej wielokrotnością f_0. Podczas tego procesu otrzymujemy zestaw obiektów harmonicznych dla danego fragmentu sygnału. Liczba obiektów harmonicznych odpowiada licznie trajektorii sinusoidalnych o częstotliwości f_0.

- Dla każdego obiektu harmonicznego wybierana jest trajektoria, która zostanie wykorzystana w procesie demodulacji estymacji sygnału $x_0(t)$. Zmieniająca się częstotliwości chwilowa tej trajektorii oznaczona jest jako f_{mod} i wyznaczana zgodnie poniższą regułą:

$$k_{mod} = \arg\max_{k} \ (E_k \cdot L_k) \qquad k \in \{1, \dots k_{end}\}, \tag{4.27}$$

gdzie:

k_{mod} - indeks wybranej k-tej harmonicznej,

L_k - długość k-tej harmonicznej (liczba węzłów),

E_k - energia k-tej harmonicznej:

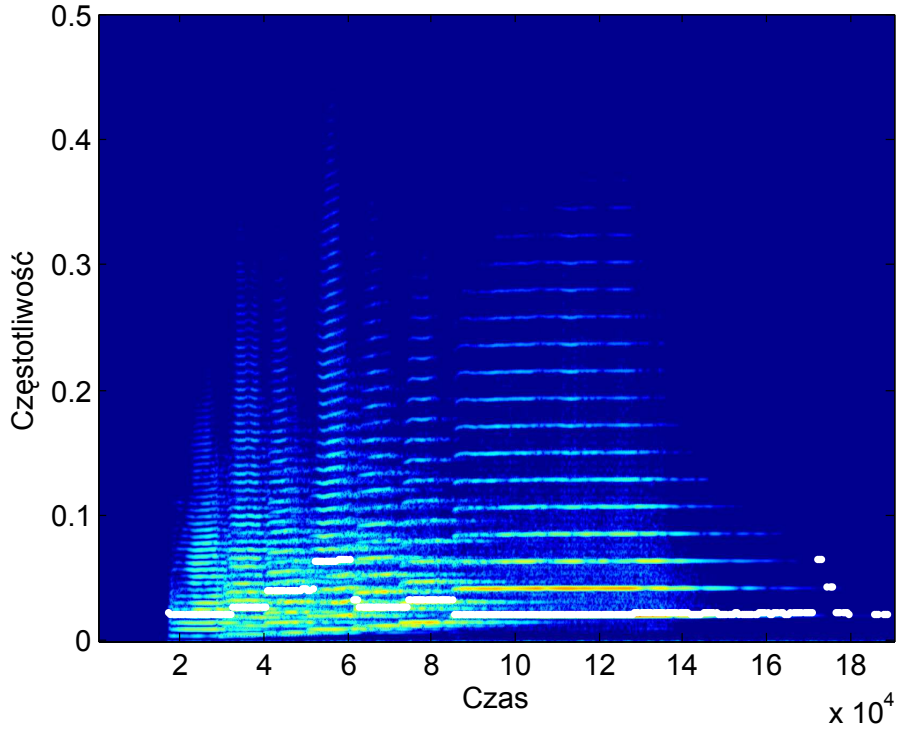
$$E_k = \sum_{l=1}^{L_k} |A_l|^2, \tag{4.28}$$

Zgodnie ze wzorem (4.27) wybrana zostaje trajektoria o dominującej energii dla danego obiektu harmonicznego. Dodatkowo, istotne znaczenie ma długość trajektorii sinusoidalnej (L_k). Preferowane są trajektorie o długim czasie trwania, co pozwala zmniejszyć strumień bitowy generowany przez parametry opisujące wybrane trajektorie.

Przykładowy wybór trajektorii sinusoidalnej f_{mod} na podstawie reguły (4.27) został zaprezentowany na rysunku 4.9.

4.7.3. Detektor tonalności

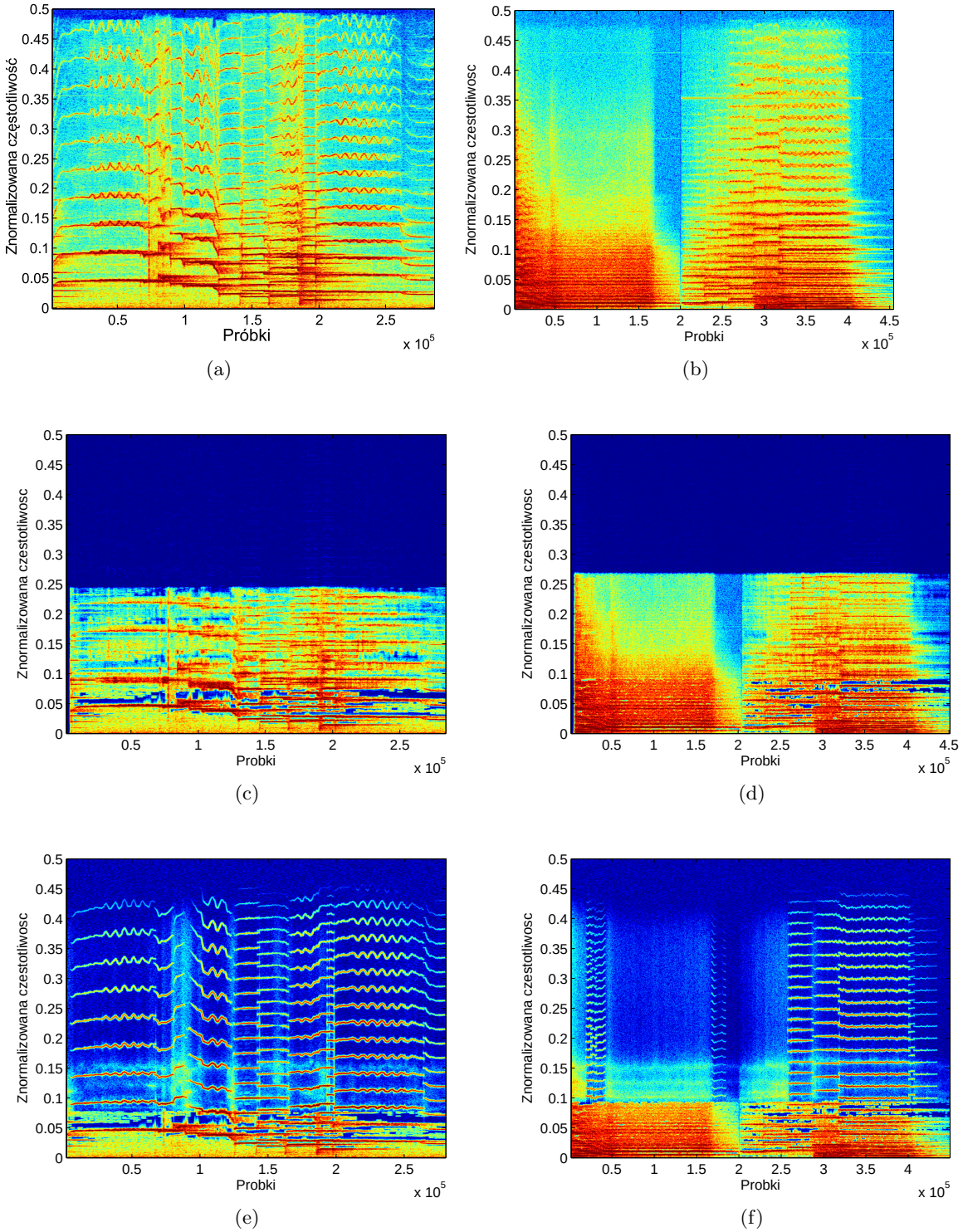
W proponowanej technice BWE zastosowano detektor tonalności, którego zasada działania została przedstawiona w punkcie 3.6. Zadaniem klasyfikatora jest podjęcie decyzji czy dana fragment sygnału będzie kodowany z wykorzystaniem techniki skalowania częstotliwości, lub zostanie zrekonstruowana przez model szumowy. Analiza tonalności przeprowadzana jest w zakresie częstotliwości: $f_{BWE} - -f_s/2$.



Rysunek 4.9: Trajektoria wykorzystania w algorytmie skalowania częstotliwości, spełniająca kryterium (4.27).

4.8. Przykład działania techniki i wyniki eksperymentalne

Rysunek 4.10 przedstawia przykład działania proponowanej techniki poszerzania widma. Możemy zauważyć, że technika MPEG-4 AAC HE dekoduje sygnał, który mimo zastosowania techniki replikacji widma SBR w dalszym ciągu ma ograniczone pasmo. To zjawisko nie występuje dla proponowanej techniki BWE. W tym wypadku algorytm skalowania częstotliwości umożliwia rekonstrukcję składowych dla pełnego zakresu częstotliwości HF. Dodatkowo poprawnie reprezentowane są składowe tonalne, co widać zwłaszcza na przykładzie sygnału *violin* (rysunek 4.10e). Możliwe to było dzięki poprawnej detekcji częstotliwości podstawowej f_0. Poza składowymi tonalnymi w dekoderze rekonstruowany jest również szum. W tym wypadku (rysunek 4.10e) obwiednia składowej szumowej nie została wiernie zrekonstruowana. Związane to jest z ograniczoną pulą bitów dostępną dla kodowania szumu. Jednak wyniki przedstawione w rozdziale 5 oraz dodatku D pokazują, że nieprecyzyjna reprezentacja składowej szumowej nie jest negatywnie postrzegana przez słuchaczy. Znacznie gorzej oceniane są sygnały, w których źle zostały zdekodowane składowe tonalne, przykładem jest technika MPEG-4 AAC HE (rysunki 4.10c i 4.10d).



Rysunek 4.10: Przykład rekonstrukcji sygnału violin2 dla prędkości bitowej 16 kb/s - rysunki a), c), e) oraz rekonstrukcji sygnału strings dla prędkości bitowej 20 kb/s - rysunki b), d), f). Spektrogramy przedstawiają: a), b) sygnał oryginalny, c), d) sygnał zakodowany przy pomocy techniki MPEG-4 AAC HE d), e) sygnał zdekodowany przy użyciu proponowanej techniki BWE. Należy zauważyć, że maksymalny zakres rekonstruowanych częstotliwości przez technikę MPEG-4 AAC HE jest ograniczony przez mechanizm sterowania prędkości bitowej. Tego typu efekt nie występuje w proponowanej technice BWE - rekonstruowany jest pełen zakres częstotliwości.

Tabela 4.1: Parametry kwantyzacji ustalone dla proponowanej techniki BWE. Skwantowane wartości podlegają kodowaniu ADPCM.

Trajektorie częstotliwości podstawowych (f_0)	**40 [cent]**
Współczynniki LSF skalujące obwiednię widma	**0.5 dB**
Współczynniki wzmocnienia LPC dla obwiednii widma	**0.1 dB**
Współczynniki LPC (residuum szumowe)	**4 bity**
Współczynniki wzmocnienia LPC (residuum szumowe)	**3 dB**

Tabela 4.2: Prędkości bitowe strumieni opisujących poszczególne parametry proponowanego kodeka.

Trajektorie częstotliwości podstawowych (f_0)	$\sim$ **0.1 kb/s**
Współczynniki LSF skalujące obwiednie widma	$\sim$ **2 kb/s**
Współczynniki wzmocnienia LPC dla obwiednii widma	$\sim$ **0.5 kb/s**
Residuum szumowe (LPC)	$\sim$ **0.4 kb/s**
Całkowita prędkość bitowa strumienia	**1 - 3 kb/s**

Interesującym przypadkiem jest spektrogram pokazany na rysunku 4.10f. Algorytm detektora tonalności (punkt 4.7.3) zdecydował, że część sygnału dla zakresu próbek od 0 do $2{\cdot}10^5$ ma charakter szumu. W tym zakresie koder nie zastosował algorytmu skalowania częstotliwości i rekonstrukcji składowych tonalnych, co spowodowało wygenerowaniem szumu. Brak składowych tonalnych umożliwił przeznaczenie większej liczby bitów na kodowanie szumu. Pozwoliło to spowodowało uzyskać lepszą reprezentację szumu w dekodowanym sygnale.

Parametry kompresji oraz prędkości bitowe poszczególnych elementów strumienia danych zostały zestawione w tabelach 4.1 i 4.2. Najmniejszy strumień bitowy opisuje parametryczną reprezentacja trajektorii f_0. Spowodowane jest to efektywną kompresją algorytmu kodowania trajektorii sinusoidalnych (dodatek B) oraz niewielką liczbą wykrywanych szeregów harmonicznych w sygnale. Liczba wykrytych szeregów zależy natomiast od dokładności estymacji f_0. Największą część strumienia, podobnie jak w technice SBR, zawierają współczynniki skalujące obwiednię widma. Dlatego należy poszukiwać efektywniejszych metod reprezentacji oraz kodowania obwiedni widmowej.

4.9. Podsumowanie

W rozdziale zostało zaproponowane nowe, oryginalne podejście do realizacji techniki BWE oferującej wysoką jakość dekodowanego sygnału przy jednoczesnej małej prędkości bitowej, rzędu 16 – 24 kb/s, natomiast 1 – 3 kb/s dla składowych HF. Główną zaletą proponowanej techniki jest poprawna rekonstrukcja energii tonalnych składowych dużych częstotliwości oraz prawidłowa rekonstrukcja zmiennych w czasie i częstotliwości składowych sygnału. Proponowana technika BWE pozwala na osiągnięcie wyższej jakości zdekodowanego sygnału w porównaniu do obecnie istniejących metod, czyli MPEG-4 AAC HE, zwłaszcza dla sygnałów posiadających silne składowe tonalne w zakresie dużych częstotliwości. Wyniki badań zostały zaprezentowane w rozdziale 5.

Rozdział 5

Ocena eksperymentalna

5.1. Wprowadzenie

W rozdziale przedstawione zostaną wyniki eksperymentów dla zaproponowanych technik kodowania sygnałów fonicznych wykorzystujących rozszerzanie widma. Celem eksperymentów było porównanie efektywności kompresji zaproponowanych technik z efektywnością MPEG-4 AAC HE dla prędkości bitowych w zakresie 16 – 24 kb/s. Podczas badań wykorzystano metodologię subiektywnych testów odsłuchowych dla małych i średnich zniekształceń według standardu ITU-R BS.1534-1 (MUSHRA) [78]. Opis metody przeprowadzenia testów odsłuchowych znajduje się w rozdziale 1.4.

5.2. Przygotowanie testów odsłuchowych

Praca dotyczy w szczególności kompresji sygnałów reprezentujących muzykę. Z tego powodu, w badaniach nie uwzględniono sygnałów mowy. Wykorzystano sygnały foniczne, których spis znajduje się w tabeli 1.1. Sygnały testowe zostały podzielone na dwie kategorie:

- fragmenty nagrań utworów zagranych na *pojedynczym instrumencie*,
- fragmenty nagrań utworów zagranych na *wielu instrumentach* – muzyka.

Sygnały foniczne użyte w badaniach są sygnałami monofonicznymi o częstotliwości próbkowania 44,1 kHz. W trakcie badań ocenie podlegała jakość dźwięku sygnałów zdekodowanych przez testowane kodeki, dla prędkości bitowych 16, 20 i 24 kb/s.

W dalszej części rozprawy wykorzystano następujące oznaczenia technik BWE:

- hfr-sin – łączna metoda wykorzystująca MPEG-4 AAC HE oraz modelowanie sinusoidalne (opisana w rozdziale 3). Technika została przebadana pod kątem zastosowania pętli sterowania przepływnością opisanej w punkcie 3.7.2. Zaproponowano dwa warianty:
  - hfr-sin-v1 – wariant bez pętli sterowania przepływnością,
  - hfr-sin-v2 – wariant z pętlą sterowania przepływnością,
- hfr-dem – technika rozszerzania widma poprzez skalowanie częstotliwości (opisana w rozdziale 4),
- he-aac – MPEG-4 AAC HE zawierający technikę SBR.

Dodatkowo podczas badań wykorzystano nieskompresowane sygnały odniesienia (kotwice) o częstotliwości próbkowania 44100 Hz, rozdzielczości 16 bitów na próbkę. Wykorzystano następujące oznaczenia:

- oryginał – sygnał oryginalny o jakości nagrania na płycie CD,
- filtr-7-kHz – sygnał oryginalny ograniczony pasmowo do 7 kHz,
- filtr-3,5-kHz – sygnał oryginalny ograniczony pasmowo do 3,5 kHz,

Oznaczenia kodeków, stosowane w dalszej części pracy zawierają również informację o docelowej prędkości bitowej strumienia generowanego przez badany kodek. Przykładowo, „hfr-dem-20kb/s" oznacza sygnał zdekodowany przez technikę rozszerzania widma wykorzystującą skalowanie częstotliwości (rozdział 4), dla strumienia o prędkości bitowej 20 kb/s.

W badaniach wzięło udział 17 osób w wieku od 20 do 38 lat. Uczestnicy testów mieli już doświadczenie związane z testami odsłuchowymi. Osoby te zostały również zaznajomione z metodologią badań, sygnałami testowymi oraz rodzajem zniekształceń wprowadzanych przez techniki kompresji sygnałów fonicznych.

Każda osoba wzięła udział w pojedynczej *sesji*, w której odsłuchano 21 *zestawów testowych*. Każdy zestaw testowy zawierał sygnał oryginalny, jego odpowiedniki ograniczone pasmowo (kotwice) oraz sygnały zakodowane technikami BWE, co daje łącznie 7 sygnałów

fonicznych w zestawie. Średni czas trwania pojedynczej sesji odsłuchowej wynosił około 50 minut.

Testy zostały przeprowadzone w specjalnie do tego celu przygotowanym pomieszczeniu odsłuchowym w Katedrze Telekomunikacji Multimedialnej i Mikroelektroniki, Politechniki Poznańskiej. Pomieszczenie to spełnia wymagania akustyczne stawiane przez normę ITU-R BS.1534-1 (MUSHRA) [78].

5.3. Analiza statystyczna

Wykresy zaprezentowane w tej części rozprawy, przedstawiają wartości średnie oceny subiektywnej, uzyskane z analizy statystycznej danych eksperymentalnych. Dodatkowo wyznaczone zostały 95% przedziały ufności dla każdego uzyskanego wyniku. Analiza statystyczna została wykonana zgodnie z zaleceniami standardu testów odsłuchowych BS.1534 [78]. Wartość średnia subiektywnej oceny MUSHRA, została wyznaczona na podstawie wzoru:

$$\mu_{jk} = \frac{1}{n_{jk}} \sum_{i=1}^{n_s} x_{ijk}, \tag{5.1}$$

gdzie n_{jk} jest liczbą osób oceniających k-ty sygnał i j-ty kodek ($j = 1, \ldots, J$), J jest liczbą badanych grup - w tym wypadku $J = 7$ i jest to suma liczby badanych kodeków oraz rodzajów zastosowanych kotowic (sygnałów odniesienia). Natomiast, ocena oddana przez i-tego słuchacza dla j-tego kodeka (kotwicy) i k-tego sygnału, oznaczona jest jako x_{ijk}. Wartości średnie μ_j i μ_k wyznaczone są w podobny sposób.

W pracy zastosowano jednoczynnikową analizę wariancji (ang. *analysis of variance*) ANOVA [14, 155], która ma na celu odpowiedzieć na pytanie, czy analizowane średnie μ_{jk} różnią się między sobą. Przy czym, liczba kodeków jest czynnikiem wpływającym na badaną jakość zdekodowanego sygnału. W celu wyznaczenia przedziałów ufności w pracy zastosowano test „prawdziwie istotnej różnicy" HSD Tukeya (ang. *honestly significant difference*) [155] , który wykorzystuje studentyzowany rozstęp w celu porównania wszystkich par średnich między badanymi kodekami. Statystyka testowa ma postać:

$$HSD_\alpha = q_{\alpha;J;N-J} \sqrt{\frac{MS_E}{N}}, \tag{5.2}$$

gdzie:

$q_{\alpha;J;N-J}$ – wartość krytyczna studentyzowanego rozstępu studenta dla poziomu istotności α, liczby badanych grup J. Wartość studentyzowanego rozstępu odczytuje się z tablic matematycznych [162],

$N = \sum_{j=1}^{J} n_{jk}$,
MS_E – średni kwadrat błędu:

$$MS_E = \frac{\sum_{j=1}^{J} \sum_{i=1}^{n_{jk}} (x_{ij} - \hat{\mu}_j)^2}{N - J}. \tag{5.3}$$

Ostatecznie, 95% przedziały ufności wyznaczane są na podstawie wzoru:

$$\left\langle \mu_{jk} - \frac{HSD_\alpha}{2}, \mu_{jk} + \frac{HSD_\alpha}{2} \right\rangle. \tag{5.4}$$

Przedziały ufności zaznaczone są na każdym wykresie zamieszczonym w bieżącym rozdziale. Jeśli przedziały nie nachodzą na siebie wówczas z 95% pewnością można stwierdzić, że jakość sygnału zdekodowanego przez jeden kodek jest lepsza od innego. Analiza statystyczna została wykonana z pomocą dostępnego publicznie oprogramowania o dostępnym kodzie źródłowym „FRIEDMAN" [45].

5.4. Wyniki - ocena subiektywna

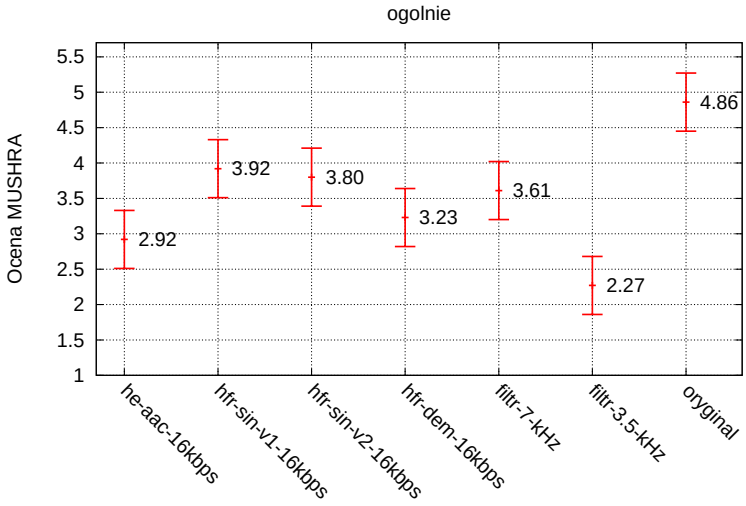
5.4.1. Porównanie kodeków wykorzystujących techniki rozszerzania widma dla sygnałów muzycznych

Na podstawie rysunku 5.1 możemy zaobserwować że, jakość sygnału reprezentowanego przez kotwicę 3,5 kHz została oceniona jako „słaba" (tabela 1.2), natomiast przez kotwicę 7 kHz jako „zadowalająca". Możemy również zaobserwować, że sygnał oryginalny został poprawnie rozpoznany przez słuchaczy i oceniony jako „doskonały". Podobne wnioski i spostrzeżenia możemy sformułować dla pozostałych wyników badań przedstawionych w tym punkcie. Szczegółowe wynik dla poszczególnych sygnałów można znaleźć w dodatku D.1.

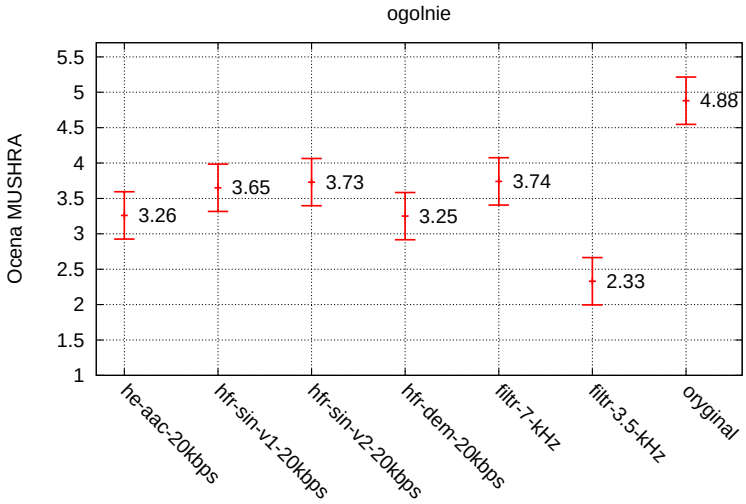
Kodek MPEG-4 AAC HE (he-aac) otrzymał ocenę „zadowalający" i został oceniony zdecydowanie niżej niż proponowane w rozprawie techniki rozszerzania widma. Jakość sygnału zdekodowanego za pomocą proponowanych technik BWE mieści się również w przedziale „dobra". Wyjątkiem jest technika „hfr-dem", która została oceniona jako „zadowalająca". Technika „hfr-dem' otrzymała jednak ocenę o 0.31 punktu (około 8%) wyższą w porównaniu do kodeka MPEG-4 AAC HE.

Rysunki 5.2 i 5.3 przedstawiają odpowiednio wyniki badań dla prędkości bitowych 20 i 24 kb/s. Można zaobserwować, że wraz ze wzrostem prędkości bitowej zwiększa się

również subiektywna ocena kodeka MPEG-4 AAC HE, od 2,92 dla 16 kb/s do 3,93 dla 24 kb/s.

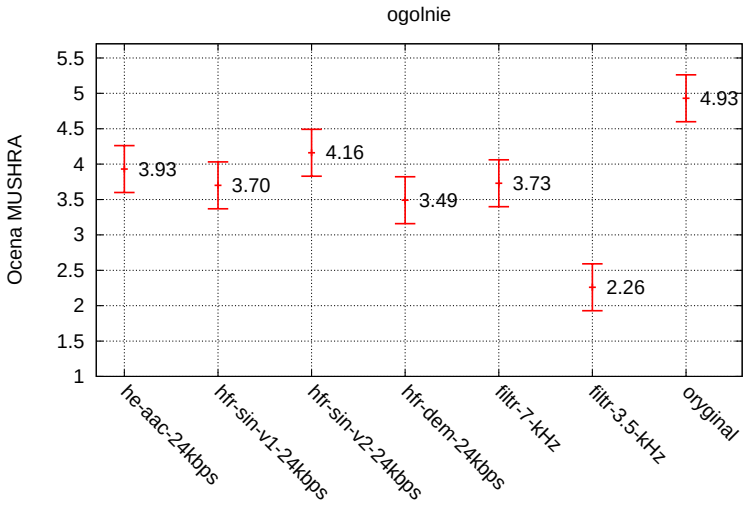


Rysunek 5.1: Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 16 kb/s z 95% przedziałem ufności.



Rysunek 5.2: Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 20 kb/s z 95% przedziałem ufności.

Należy podkreślić, że proponowane techniki rozszerzania widma otrzymały bardzo zbliżone oceny dla wszystkich testowanych prędkości bitowych. Badanie to pokazuje, że proponowane techniki (w zastosowaniu dla szerokiego zakresu materiału muzycznego i prędkości bitowych z zakresu 16 – 24 kb/s), uzyskują nie gorsze oceny niż aktualnie uważana za najlepszą technika MPEG-4 AAC HE. Z drugiej strony, stosując proponowane techniki



Rysunek 5.3: Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 24 kb/s z 95% przedziałem ufności.

BWE możemy ograniczyć prędkość bitową nawet do 16 kb/s uzyskując tę samą jakość zdekodowanego sygnału jak dla 24 kb/s.

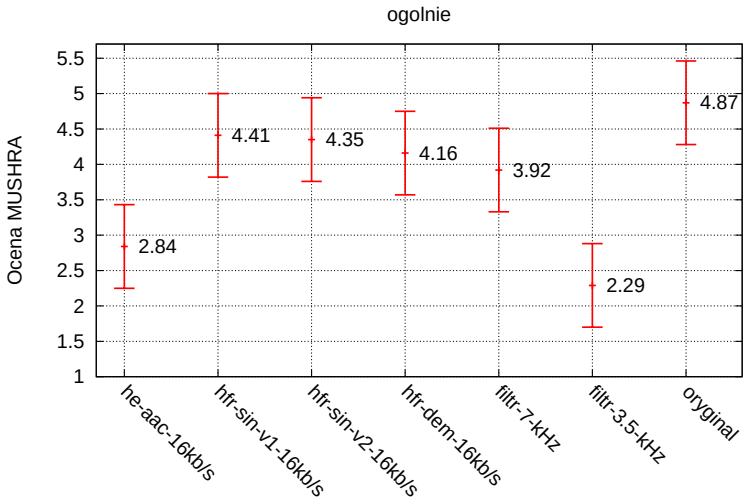
Jeśli porównamy badane kodeki z sygnałami kotwic, wówczas zauważymy, że proponowane techniki BWE uzyskują wyższą ocenę MUSHRA niż sygnał ograniczony pasmowo do 3,5 kHz. Na podstawie tabeli 2.1 wiemy, że częstotliwość podziału f_{BWE} testowanych kodeków dla 16 kb/s jest zbliżona i wynosi 3,44 kHz. Testowane techniki BWE nie wprowadzają więc dodatkowych zniekształceń do sygnału i gwarantują wyższą jakość dźwięku niż sygnał ograniczony pasmowo.

Możemy zauważyć, że oceny testowanych technik BWE są zbliżone lub niższe od oceny uzyskanej przez sygnał ograniczonego pasmowo do 7 kHz. Na tej podstawie można określić górną granicę jakości zdekodowanego sygnału, jaką mogą zaoferować badane kodeki dla sygnałów muzycznych.

5.4.2. Porównanie kodeków wykorzystujących techniki rozszerzania widma dla sygnałów zawierających silne składowe tonalne

W tym punkcie przedstawione zostaną wyniki badań eksperymentalnych zebrane dla sygnałów posiadających silne składowe tonalne o zmiennej częstotliwości w pasmie dużych częstotliwości. Sumaryczne wyniki przedstawiają rysunki 5.4, 5.5 i 5.6 odpowiednio dla 16, 20 i 24 kb/s. Szczegółowe wynik dla poszczególnych sygnałów można znaleźć w dodatku D.2.

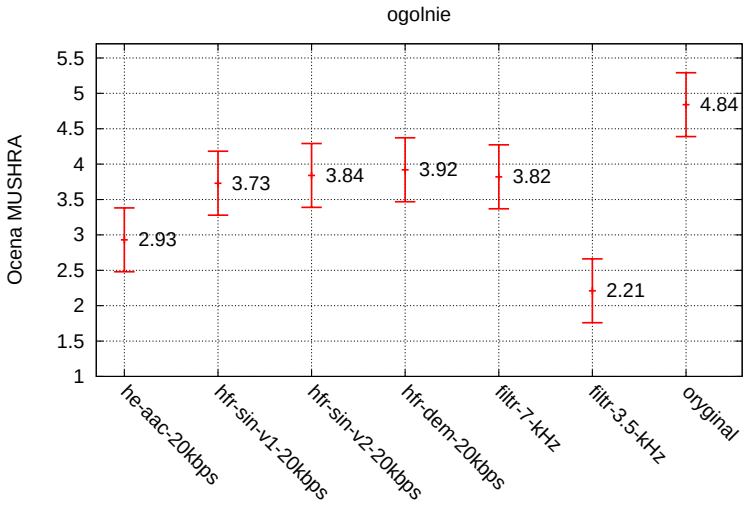
Na podstawie zaprezentowanych rysunków można zauważyć, że efektywność proponowanych rozwiązań zdecydowanie wzrasta dla badanej klasy sygnałów zawierających silne składowe tonalne. Proponowane techniki otrzymały ocenę subiektywną powyżej „dobrej" dla 16 i 24 kb/s i „dobrą" dla 20 kb/s. Dla 16 kb/s najwyższą ocenę otrzymała technika „hfr-sin-v1" (4,41). Wynik ten jest o 1.57 punktu wyższy od oceny uzyskanej przez MPEG-4 AAC HE (2,84 - jakość „zadowalająca"). Wynik ten jest również wyższy o 0.92 punktu w stosunku do tej samej proponowanej technik BWE dla 16 kb/s i sygnałów reprezentujących nagrania muzyczne (porównanie rysunków 5.1 i 5.4). Podobnie jak dla szerokiej klasy sygnałów muzycznych, tak i w tym przypadku, proponowane techniki oferują zbliżoną jakość zdekodowanego sygnału dla prędkości bitowych w zakresie 16 – 24 kb/s.



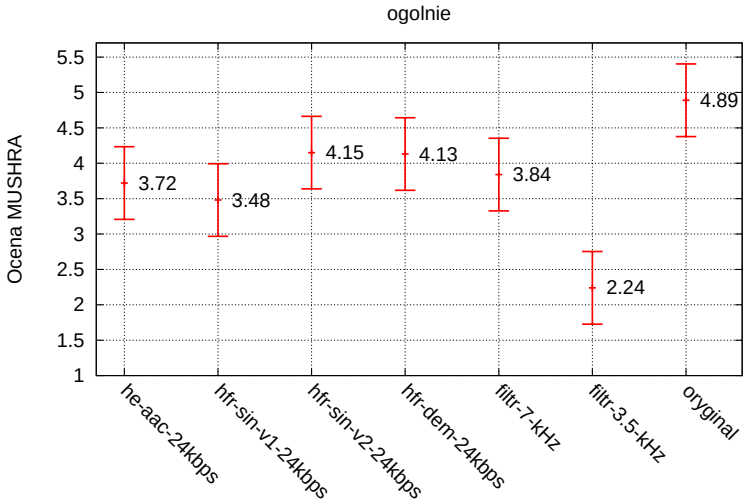
Rysunek 5.4: Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 16 kb/s z 95% przedziałem ufności.

Wyniki ocen subiektywnych proponowanych technik dla 16 kb/s są wyższe niż dla 20 i 24 kb/s. Na tego typu wyniki miały wpływ zmiany częstotliwości granicznej f_{BWE} zgodnie z tabelą 2.1. Wraz ze wzrostem prędkości bitowej zwiększa się również wartość f_{BWE}. W tym wypadku, składowe sygnału na granicy f_{BWE} są wierniej reprezentowane przez kodek sinusoidalny niż przez kodek podstawowy (AAC).

Ponadto, sygnały zdekodowane za pomocą proponowanych technik otrzymały wyższą ocenę subiektywną niż sygnały ograniczone pasmowo do 7 kHz.



Rysunek 5.5: Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 20 kb/s z 95% przedziałem ufności.



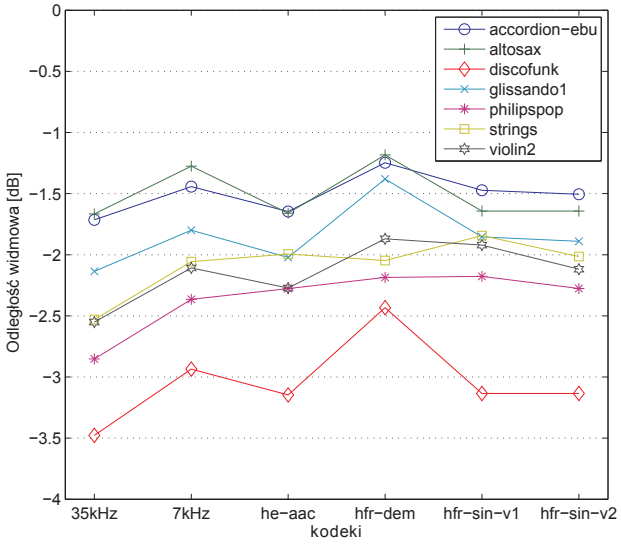
Rysunek 5.6: Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 24 kb/s z 95% przedziałem ufności.

5.5. Wyniki - ocena obiektywna

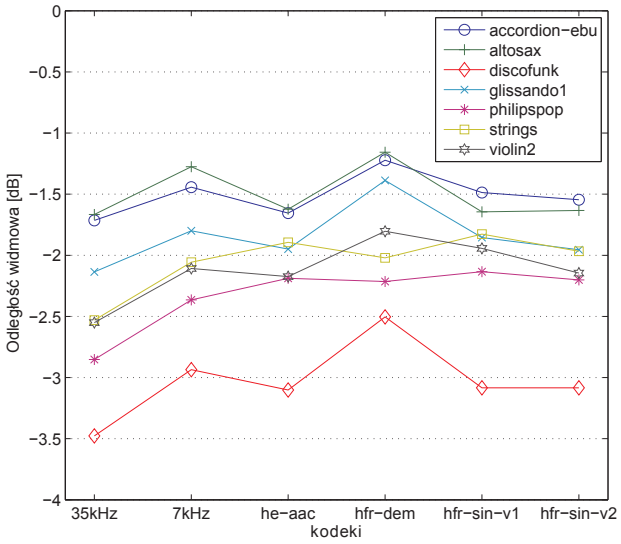
W celu weryfikacji subiektywnych badań eksperymentalnych, przeprowadzono dodatkowe badania porównujące techniki rozszerzania widma, posługując się obiektywną miarą zniekształceń. W tym przypadku zastosowano miarę logarytmicznej odległości widmowej [71] (ang. *log-spectral distance*) (LSD) wyrażonej wzorem:

$$LSD_i = \sqrt{\frac{1}{n_{high} - n_{low}} \sum_{n=n_{low}}^{n_{high}} \left[10 \log_{10} \frac{P_i(n)}{\hat{P}_i(n)} \right]^2}, \tag{5.5}$$

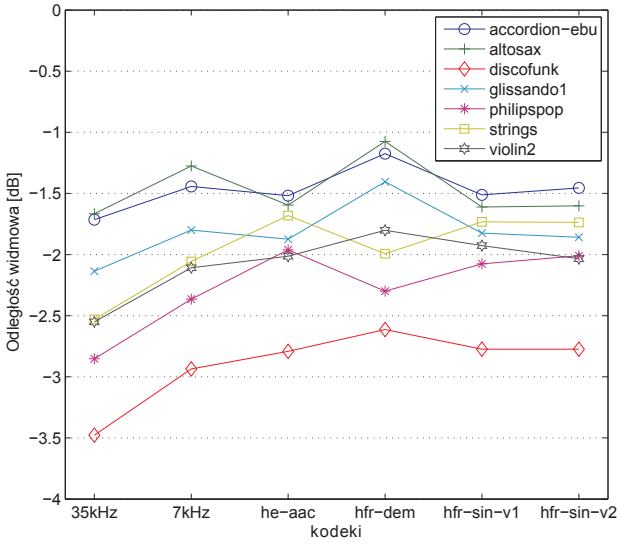
gdzie LSD_i jest odległością widmową dla okna analizy i, P_i jest widmem mocy sygnału oryginalnego, $\hat{P}_i$ jest widmem mocy sygnału zdekodowanego za pomocą proponowanych technik. LSD zostało wyznaczone dla zakresu $< n_{low}, n_{high} >$ indeksów FFT, co odpowiada zakresowi częstotliwości $< f_{bwe}, f_s/2 >$. LSD jest dobrze opisaną w literaturze miarą, która bazuje na informacji, że sposób postrzegania głośności przez człowieka ma charakter logarytmiczny. Jednak ignoruje ważne właściwości narządu słuchu, na przykład maskowanie w dziedzinie częstotliwości i czasu. Z tego powodu, wyniki przedstawione w tym punkcie mają charakter poglądowy i służą weryfikacji oceny subiektywnej.



Rysunek 5.7: Odległość widmowa dla prędkości bitowej 16 kb/s.



Rysunek 5.8: Odległość widmowa dla prędkości bitowej 20 kb/s.



Rysunek 5.9: Odległość widmowa dla prędkości bitowej 24 kb/s.

Miara LSD zawiera się w zakresie od -4 do 0, gdzie wartość 0 oznacza ten sam charakter widma amplitudowego porównywanych sygnałów. Wartości mniejsze wskazują na różnice pomiędzy sygnałami. Wykresy 5.7, 5.8 i 5.9 wskazują, że proponowane techniki otrzymał wyższe oceny niż porównywana technika MPEG-4 AAC HE. Najwyższą wartość miary LSD otrzymała technika „hfr-dem". Tak dobry wynik może być spowodowany tym, że technika „hfr-dem" rekonstruuje składowe w pełnym zakres częstotliwości.

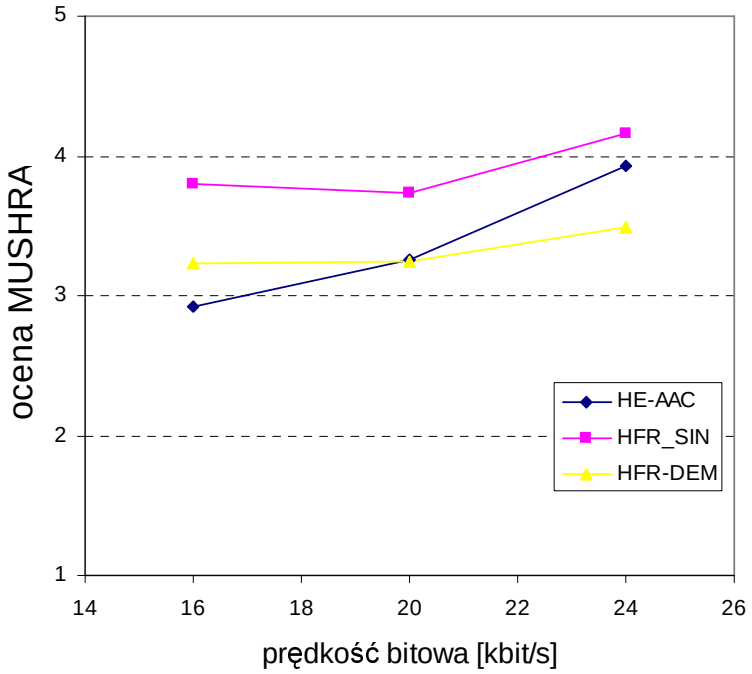
5.6. Dyskusja wyników

Zaprezentowana w rozdziale 3 technika rozszerzania widma wykorzystująca modelowanie sinusoidalne, zbadana została w dwóch wariantach z i bez pętli sterowania przepływnością. Na podstawie analizy wyników można stwierdzić, że wariant z pętlą sterowania przepływnością („hfr-sin-v2") pozwala na wyższą efektywność kompresji w stosunku do wariantu bez pętli sterowania. Dlatego w zagregowanych wynikach (rysunki 5.10, 5.11 i 5.12) symbolem „hfr-sin" oznaczany będzie kodek z pętlą sterowania przepływnością.

Zestawione wyniki dla całego zestawu sygnałów muzycznych wykorzystanych w eksperymentach pokazują, że technika „hfr-sin" otrzymała znacznie wyższe oceny niż porównywana technika MPEG-4 AAC HE (rysunek 5.10). Dla techniki „hfr-dem" otrzymane wyniki są zbliżone do MPEG-4 AAC HE.

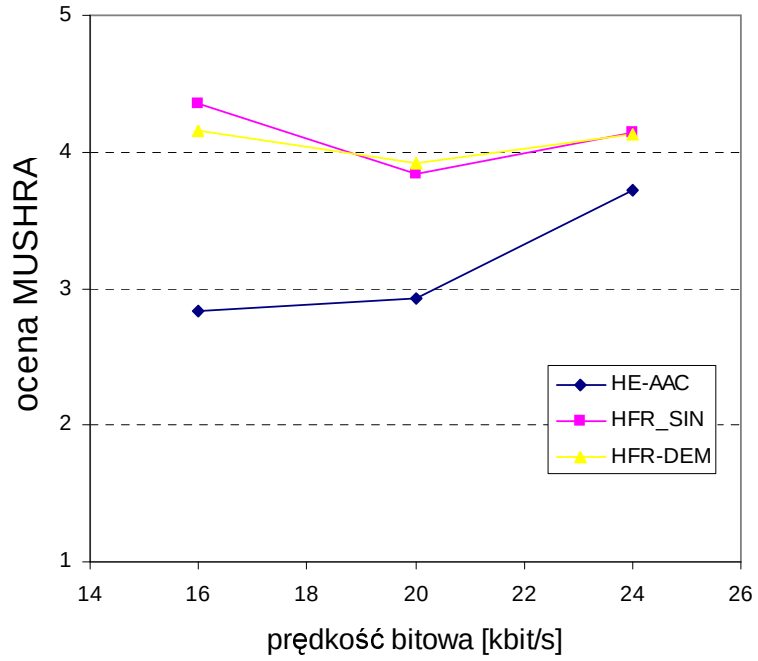
Dla zestawienia zawierającego sygnały o silnych składowych tonalnych (rysunek 5.11) obie proponowanej techniki są wyraźnie lepsze od MPEG-4 AAC HE zwłaszcza dla małych prędkości bitowych rzędu 16 kb/s.

Powyższe spostrzeżenia potwierdzają również wyniki badań obiektywnych otrzymane dla miary odległości widmowej (rysunek 5.12).

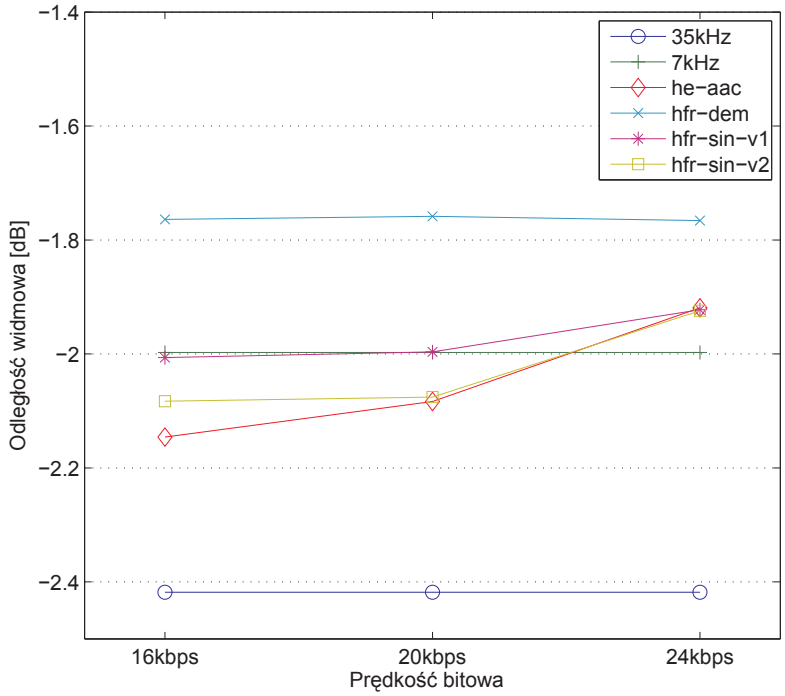


Rysunek 5.10: Zależność subiektywnej oceny od prędkości bitowej dla wszystkich badanych sygnałów. „hfr-sin" oznacza technikę opisaną w rozdziale 3 z pętlą sterowania przepływnością.

Technika rozszerzania widma wykorzystująca skalowanie częstotliwości - „hfr-dem" (rozdział 4) w zestawieniu dla sygnałów muzycznych, otrzymała oceny porównywalne do MPEG-4 AAC HE. Niższa ocena „hfr-dem" w stosunku do techniki „hfr-sin" spowodowana jest tym, że w zrekonstruowanym sygnale („hfr-dem") słyszalne są „artefakty" o metalicznym brzmieniu. Efekt ten związany jest z niedokładną separacją wąskopasmowego sygnału (rozdział 4.4.3). W uzyskanym sygnale wąskopasmowym poza głównym składnikiem sinusoidalnym pozostają inne składowe, które niewłaściwie skalują się wraz z częstotliwością. Efekt ten jest szczególnie dokuczliwy dla sygnałów zawierających wiele składowych tonalnych, np. sygnał *strings* lub *discofunk* (dodatek D). Równocześnie proponowana technika „hfr-dem" otrzymała bardzo wysokie oceny dla sygnałów *glissando1*



Rysunek 5.11: Zależność subiektywnej oceny od prędkości bitowej dla nagrań zawierających silne składowe tonalne. „hfr-sin" oznacza technikę opisaną w rozdziale 3 z pętlą sterowania przepływnością.



Rysunek 5.12: Zależność obiektywnej oceny (odległości widmowej) od prędkości bitowej dla wszystkich badanych sygnałów.

(4,43), *violin2* (4,67) i prędkości bitowej 24 kb/s. Oznacza to, że może ona stanowić dobrą alternatywę dla kodowania składowych tonalnych HF w stosunku do metody łączącej modelowanie sinusoidalne i SBR.

Wyniki badań opisane w rozdziale pokazują, że proponowane techniki BWE dają szczególnie dobre rezultaty dla sygnałów fonicznych zawierających nagrania instrumentów o silnych składowych tonalnych oraz składowych o zmiennej częstotliwości. Dla tego typu sygnałów subiektywna ocena proponowanych rozwiązań mieści się w przedziale pomiędzy „dobrą" a „doskonałą" jakością na skali MUSHRA (około 4,5 / 5 punktów). Porównywana technika MPEG-4 AAC HE otrzymuje dla tych sygnałów, ocenę poniżej jakości „zadawalającej" na skali MUSHRA (poniżej 3 punktów). Należy podkreślić, że dla szerokiego spektrum sygnałów uwzględniających muzykę, proponowane techniki zostały ocenione w podobny sposób jak technika MPEG-4 AAC HE. Oznacza to, że zaproponowane w rozprawie rozszerzenie techniki MPEG-4 AAC HE (rozdział 3) zwiększa efektywność kompresji standardowej techniki MPEG-4 AAC HE dla sygnałów zawierających silne składowe tonalne. Jednocześnie proponowana technika „hfr-sin" zachowuje zbliżoną efektywność kompresji w stosunku do MPEG-4 AAC HE dla pozostałych klas sygnałów.

Istotną zaletą proponowanych technik jest uzyskanie „dobrej" jakości dźwięku bez względu na wybraną prędkość bitową (w badanym zakresie 16-24 kb/s), zarówno dla sygnałów muzycznych (rysunek 5.10) jak i dla sygnałów zawierających dźwięki o silnej strukturze harmonicznej (rysunek 5.11).

Rozdział 6

Podsumowanie i dyskusja

6.1. Wnioski

Uzyskane wyniki badań *potwierdzają tezę pracy*. W rozprawie zostały zaproponowane techniki kompresji sygnału fonicznego, które łącznie wykorzystują rozszerzanie widma oraz modelowanie sinusoidalno-szumowe. Nowe techniki kompresji umożliwiają rekonstrukcję sygnału fonicznego o jakości lepszej niż przy zastosowaniu MPEG-4 AAC HE dla prędkości bitowych rzędu 16 – 24 kb/s. Równocześnie, zaproponowane rozwiązania poprawnie rekonstruuje składowe tonalne sygnału.

Poniżej przedstawione zostały najważniejsze wnioski rozprawy.

- Zastosowanie modelowania sinusoidalnego pozwala na poprawną reprezentację składowych tonalnych sygnału dużych częstotliwości oraz ich efektywną kompresję.
- Zastosowanie modelowania sinusoidalnego w połączeniu z techniką - MPEG-4 AAC HE, pozwala na poprawę jakości zdekodowanego sygnału dla tej samej prędkości bitowej.
- Efektywność kompresji dla techniki wykorzystującej łączny mechanizm rozszerzania widma i modelowania (rozdział 3) w dużym stopniu zależy od poprawnej klasyfikacji składowych tonalnych. Na proces klasyfikacji istotny wpływ ma obecność szumu w sygnale oraz charakter poszczególnych składowych tonalnych. Szybkozmienne składowe w krótkookresowym widmie mogą zostać zakwalifikowane jako szum. Parametry trajektorii sinusoidalnych kodowane są różnicowo, jeśli więc zostaną zakłócone przez szum wówczas efektywność kompresji zmniejszy się.

- W punkcie 3.9.4 pokazano, że istnieje minimalna wartość $\overline{N}_{Sin,max}$ średniej liczby sinusoid na okno analizowanego sygnału, umożliwiająca poprawną rekonstrukcję sygnału, czyli uzyskania wyniku „dobry" na skali MUSHRA. Dla zaproponowanej techniki granica ta wynosi 20 sinusoid na okno analizy.

- W punkcie 3.9.4 przedstawiono wyniki badań, które wskazują, że istnieje górna granica liczby kodowanych trajektorii, której przekroczenie nie poprawia jakości zrekonstruowanego sygnału fonicznego. W tej sytuacji dochodzi do redukcji jakości dekodowanego sygnału. Spowodowane jest to błędną klasyfikacją składowych tonalno-szumowych. Podczas badań ustalono, że dla wykorzystanych sygnałów testowych redukcja jakości następuje po przekroczeniu średniej liczby 40 – 50 sinusoid na okno analizy.

- Na podstawie wyników badań przedstawionych w punkcie 3.9.3 ustalono, że parametry modelu sinusoidalnego powinny być kwantowane równomiernie w skali logarytmicznej z krokiem 40 centów dla częstotliwości oraz 3 dB dla amplitud. W zaproponowanych rozwiązaniach zrezygnowano z kodowania parametru fazy dla dużych częstotliwości. Zastosowana strategia kodowania pozwoliła na efektywną kompresję oraz poprawną rekonstrukcję sygnałów.

- Istotnym parametrem wpływającym na jakość zrekonstruowanego sygnału jest ciągłość parametrów modelu sinusoidalnego, zarówno amplitud jak i częstotliwości. Niekompletne lub krótkotrwałe trajektorie postrzegane są jako składowe szumowe lub powodują nieprzyjemne artefakty związane z gwałtowną zmiana głośności na początku i końcu trajektorii (punkty 3.4.2 i 3.5).

- Sterowanie pracą kodera sinusoidalnego musi odbywać się poprzez zastosowanie odpowiedniego warunku psychoakustycznego. W innym wypadku, koder może odrzucić trajektorie percepcyjnie istotne (punkt 3.7.3).

- W zaproponowanych rozwiązaniach zastosowano kodowanie Huffmana z dynamicznie tworzoną tablicą przesyłaną co pewien interwał czasu. Na podstawie wyników

badań przedstawionych w punkcie 3.7.4 stwierdzono, że dla tablicy Huffmana przesyłanej co około 0,5 s możliwa jest rekonstrukcja sygnału o wysokiej jakości. Jednocześnie jest to odstęp czasu umożliwiający właściwą synchronizację ze strumieniem obrazów ruchomych (punkt 3.7.4).

- Na podstawie badań wstępnych ustalono, że składowe tonalne warto rozpatrywać jako sygnały wąskopasmowe o szerokości pasma zależnej od lokalnych modulacji AM i FM. Ponieważ reprezentacja oparta na takim podejściu (punkt 4.2) oferuje wierniejszą, naturalniej brzmiącą rekonstrukcję sygnału, w porównaniu do reprezentacji typowej dla modelu sinusoidalnego, która brzmi syntetycznie z powodu zbyt uproszczonego przedstawienia zmienności amplitud i częstotliwości jako funkcji wolnozmiennej.

- Badania wstępne (punkt 2.6) przeprowadzone z wykorzystaniem referencyjnej implementacji kodera MPEG-4 AAC HE wykazały istotne znaczenie modulacji amplitudy i częstotliwości poszczególnych składowych tonalnych dla subiektywnej oceny jakości sygnału. Ucho ludzkie wrażliwe jest na zniekształcenia modulacji częstotliwości wprowadzane przez koder standardu MPEG-4 AAC HE w zakresie dużych częstotliwości. Wnioski te zostały również potwierdzone podczas badań przedstawionych w rozdziale 5.

6.2. Oryginalne osiągnięcia

Głównym osiągnięciem autora jest opracowanie dwóch efektywnych technik kompresji wykorzystujących rozszerzanie widma i jednocześnie mogących być uzupełnieniem aktualnie stosowanych standardów w tej dziedzinie (MPEG-4 AAC HE).

Technika opisana w rozdziale 3 umożliwia dokładniejszą rekonstrukcję składowych tonalnych w stosunku do standardowej techniki SBR. Zrealizowano to, poprzez łączne zastosowanie w tym samym paśmie częstotliwości modelowania sinusoidalnego oraz techniki SBR. Proponowana technika stanowi uzupełnienie techniki MPEG-4 AAC HE.

Druga technika przedstawiona w rozdziale 4 wykorzystuje skalowanie częstotliwościowe składowych tonalnych z dolnej części pasma w celu odtworzenia brakujących wyższych składowych harmonicznych. Wąskopasmowy sygnał reprezentuje składową tonalną jako

składową sinusoidalną przekształconą mikro-modulacjami AM i FM. Technika ta jest alternatywnym rozwiązaniem dla techniki SBR.

Obie techniki rozszerzania widma są autorskimi rozwiązaniami i stanowią oryginalny wkład w dziedzinie kompresji sygnałów wykorzystujących rozszerzanie widma.

Nowe techniki zostały przebadane eksperymentalnie zgodnie z obowiązującymi standardami testów subiektywnych. Parametry proponowanych kodeków ustalono na podstawie wielu badań eksperymentalnych. Uzyskane wyniki pozwoliły na implementację kodeków, a następnie poddanie ich badaniom eksperymentalnym mającym na celu porównanie proponowanych rozwiązań z techniką MPEG-4 AAC HE.

Wspomniane wyżej prace eksperymentalne wykonanie około 10 serii testów odsłuchowych, w których łączny czas trwania prezentowanych nagrań wynosił ponad 150 godzin. Na ten czas składa się końcowe porównanie efektywności proponowanych rozwiązań, jak również czas poświęcony na dobór wstępnych parametrów kodeków.

Należy podkreślić fakt, iż testy odbywały się w specjalnie do tego celu przygotowanym pomieszczeniu odsłuchowym. W tym miejscu w danym czasie mogła przebywać jedna osoba, co znacząco komplikowało i wydłużało czas badań. Realny czas trwania eksperymentów szacowany jest na 3 - 4 miesiące codziennych badań.

Przeprowadzenie i przygotowanie tak dużej liczby eksperymentów było zadaniem trudnym ze względów logistycznych i dużej liczby osób uczestniczących w eksperymentach. Na trud powstania pracy składa się również czas poświęcony na przetworzenie i analizę wyników.

Wyniki badań przedstawione w rozdziale 5 pokazują, że proponowane techniki oferują wyższą efektywność kompresji w stosunku do standardowej techniki MPEG-4 AAC HE. Szczególnie dobre rezultaty uzyskały dla sygnałów fonicznych zawierających dźwięki o silnych składowych tonalnych i/lub zmiennej częstotliwości. Dla tego typu sygnałów subiektywna ocena proponowanych rozwiązań mieści się w przedziale pomiędzy „dobrą" a „doskonałą" jakością na skali MUSHRA (około 4,5 / 5 punktów). Porównywana technika MPEG-4 AAC HE otrzymuje, dla tych sygnałów, ocenę poniżej jakości „zadawalającej" na skali MUSHRA (poniżej 3 punktów). Należy podkreślić, że dla prędkości bitowych poniżej 24 kb/s i dla pozostałych klas sygnałów uwzględniających muzykę, proponowane techniki uzyskały wyższą efektywność kompresji w stosunku do MPEG-4 AAC HE. Dla

prędkości bitowych równych 24kb/s porównywane techniki oferują podobną efektywność kompresji.

Poniżej przedstawiono pozostałe osiągnięcia autora pracy.

- Zaproponowanie metody klasyfikacji składowych tonalnych bazującej na łącznej analizie parametrów widma chwilowego oraz trajektorii sinusoidalnych (punkt 3.4.2).
- Zaproponowanie klasyfikatora tonalności, którego celem jest przełączanie pracy kodera pomiędzy aktualnym trybem, a proponowanymi technikami (punkty 3.6 i 4.7.3).
- Przedstawienie strategii ograniczenia liczby kodowanych trajektorii do percepcyjnie najistotniejszych (punkt 3.9.4).
- Przedstawienie algorytmu sterownia pracą kodera sinusoidalnego (punkt 3.7.2), bazując na percepcyjnym kryterium ograniczania liczby kodowanych trajektorii sinusoidalnych (punkt 3.9.4).
- Pokazanie aktualnego stanu wiedzy w dziedzinie kompresji z rozszerzaniem widma (rozdział 2).
- Implementacja aplikacji służącej do prowadzenia testów odsłuchowych subiektywnej oceny jakości sygnałów fonicznych według metodologii MUSHRA (punkt 1.4).

Istotnym osiągnięciem przedstawionym w pracy jest analiza zniekształceń wprowadzanych przez technikę MPEG-4 AAC HE (punkt 2.5 i 2.6). Na podstawie badań wskazano mechanizmy powodujące powstawanie zniekształceń oraz zaproponowano możliwe rozwiązania.

Autor przedstawił w pracy również analizę właściwości modelowania sinusoidalno-szumowego, które uwzględniały problemy klasyfikacji składowych (punkty 3.4.2 i 3.6), estymację parametrów oraz wskazał źródła zniekształceń kompresji dla modelownania sinusoidalnego (punkt 3.5).

W dodatku C pokazano także wyniki testów autorskiej implementacji kodeka sinusoidalno-szumowego w porównaniu do MPEG-4 SSC [69].

6.3. Dalsze badania

Dalsze badania powinny koncentrować się na redukcji prędkości bitowej strumienia reprezentującego składowe dużych częstotliwości. Osiągnąć to można poprzez zastosowanie bardziej zaawansowanych metod kodowania trajektorii sinusoidalnych, na przykład wykorzystując relacje pomiędzy składowymi harmonicznymi. Możliwe jest również zastąpienie analizy szumu metodą Linear Predictive Coding (LPC) przez inny algorytm, który efektywniej aproksymuje obwiednię widma sygnału, na przykład przesyłając energię każdego pasma krytycznego.

Z punktu widzenia analizy tonalnej sygnału zasadne wydaje się poszukiwanie nowych metod klasyfikacji składowych tonalnych i szumowych. Pozwoliłoby to na efektywniejsze działania algorytmu śledzenia trajektorii sinusoidalnych, a jednocześnie dokładniejszą rekonstrukcję sygnału i efektywniejszą kompresję.

Ponadto, można:

- poszukiwać rozwiązań mogących zmniejszyć złożoność obliczeniową modelowania sinusoidalnego;
- uzupełnić proponowaną technikę rozszerzania widma z wykorzystaniem skalowania częstotliwości o modele transjentów;
- poszukiwać rozwiązań, które dla techniki przedstawionej w rozdziale 4 umożliwią uzyskanie składowych tonalnych reprezentowanych jako wąskopasmowe sygnały, pozbawionych zakłóceń pochodzących od sąsiednich składowych sygnału;
- przenieść kod programu napisany w środowisku MATLAB do implementacji w języku C. Język C jest powszechnie wykorzystywany w referencyjnych implementacjach grupy MPEG. Pozwoli to na przedstawienie własnych rozszerzeń do standardów grupy MPEG, jak również umożliwi wykorzystanie w komercyjnych aplikacjach.

Spis literatury

[1] "EBU subjective listening tests on low-bitrate audio codecs," European Broadcasting Union (EBU), Doc. 3296, Jun. 2003.

[2] 3GPP TS 26.410, "Enhanced aacPlus source code, version V8.0.0, release 8," 3rd Generation Partnership Project, Tech. Rep., 2009.

[3] 3rd Generation Partnership Project (3GPP). [Online]. Available: http://www.3gpp.org/

[4] M. Abe and J. Smith, "AM/FM rate estimation for time-varying sinusoidal modeling," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 2005, pp. 201–204.

[5] M. Arora, J. Lee, and S. Park, "High Quality Blind Bandwidth Extension of Audio for Portable Player Applications," in *120th Convention of the Audio Engineering Society*, Paris, May 2006.

[6] M. Bartkowiak and T. Żernicki, "The impact of finite word length on IMDCT computation in MPEG-2 AAC decoding," in *International Workshop on Systems, Signals and Image Processing, IWSSIP'04*, Poznań, 2004, pp. 343–346.

[7] M. Bartkowiak and T. Żernicki, "Software audio codec for interactive television," in *International Workshop on Systems, Signals and Image Processing, IWSSIP'04*, Poznań, 2004, pp. 469–472.

[8] M. Bartkowiak i T. Żernicki, "Hybrydowy psychoakustyczny kodek dźwięku dla transmisji strumieniowej", *XI Krajowe Sympozjum Nauk Radiowych URSI*, Poznań, Polska, 7–8 Kwiecień 2005, s. 395–400.

[9] M. Bartkowiak and T. Żernicki, "A simple adaptive matrixing scheme for efficient coding of stereo sound," in *Proceedings of IX European Signal Processing Conference, EUSIPCO'05*, Antalya, Turkey, Sep. 2005.

[10] M. Bartkowiak i T. Żernicki, "Improved partial tracking technique for sinusoidal modeling of speech and audio", *Poznańskie Warsztaty Telekomunikacyjne - PWT '07*, Poznań, Polska, 2007. [Online]. Available: http://www.multimedia.edu.pl/publications/

[11] M. Bartkowiak and T. Żernicki, "Harmonic sinusoidal + noise modeling of audio based on multiple f0 estimation," in *125th Convention of the Audio Engineering Society*, ser. AES Preprint 7510, San Francisco, USA, 2-5 Oct. 2008.

[12] M. Bartkowiak i T. Żernicki, "Hybrydowy parametryczno-transformatowy kodek dźwięku", *Krajowa Konferencja Radiokomunikacji, Radiofonii i Telewizji, KKRRiT*, Wrocław, Polska, 9–11 Kwiecień 2008, s. 507–510.

[13] J. Beauchamp, *Analysis, synthesis, and perception of musical sounds: the sound of music*. Springer Verlag, 2007.

[14] S. Bech, N. Zacharov, and F. Wickelmaier, *Perceptual Audio Evaluation—Theory, Method and Application*. Wiley, 2006.

[15] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'79)*, vol. 4, 1979.

[16] B. Bessette, R. Salami, R. Lefebvre, M. Jelinek, J. Rotola-Pukkila, J. Vainio, H. Mikkola, and K. Jarvinen, "The adaptive multirate wideband speech codec (amr-wb)," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 8, pp. 620–636, 2002.

[17] A. Biswas, "Perceptually biased linear prediction," *Journal of the Audio Engineering Society*, vol. 54, Dec. 2006.

[18] A. Biswas, rozmowa prywatna (private communication), 2008.

[19] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[20] M. Bosi and R. Goldberg, *Introduction to Digital Audio Coding and Standards*. Kluwer Academic Publishers Norwell, MA, USA, 2002.

[21] G. Bouvigne, "48kbps AAC public test results, MP3'Tech," http://www.mp3-tech.org/tests/aac_48/results.html, Mar. 2006.

[22] K. Brandenburg and Karlheinz, "Introduction to perceptual coding," in *AES Conference:Collected Papers on Digital Audio Bit-Rate Reduction*, May 1996.

[23] K. Brandenburg and D. Seitzer, "OCF: coding high quality audio with data rates of 64 kbit/sec," in *85th Convention of the Audio Engineering Society*, Nov. 1988, pp. 1993–1996.

[24] A. W. Bronkhorst, "Localization of real and virtual sound sources," *The Journal of the Acoustical Society of America*, vol. 98, no. 5, pp. 2542–2553, 1995.

[25] C. Budsabathon and A. Nishihara, "Bandwidth Extension with Hybrid Signal Extrapolation for Audio Coding," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 90, pp. 1564–1569, 2007.

[26] O. Cappe and J. Laroche, "Evaluation of short-time spectral attenuation techniques for therestoration of musical recordings," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 84–93, 1995.

[27] Coding Technologies. [Online]. Available: http://www.codingtechnologies.com/

[28] A. Czyżewski, *Dźwięk cyfrowy. Wybrane zagadnienia teoretyczne, technologia, zastosowanie*, II ed., ser. Problemy współczesnej nauki. Teoria i zastosowania. Akademicka Oficyna Wydawnicza EXIT, 2001.

[29] A. de Cheveigné b and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[30] M. Dempsey *et al.*, "Method for introducing harmonics into an audio stream for improving three dimensional audio positioning," 10 Apr. 2001, uS Patent 6,215,879.

[31] A. den Brinker, J. Breebaart, P. Ekstrand, J. Engdegard, F. Henn, K. Kjorling, W. Oomen, and H. Purnhagen, "An Overview of the Coding Standard MPEG-4 Audio Amendments 1 and 2: HE-AAC, SSC, and HE-AAC v2," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2009, pp. 1–21, 2009.

[32] A. den Brinker and A. Oomen, "Fast ARMA modelling of power spectral density functions," in *Proceedings of XII European Signal Processing Conference, EUSIPCO2000*, Tampere, Finland, Sep. 2000, pp. 1229 – 1232.

[33] P. Depalle, G. Garcia, X. Rodet, and P. IRCAM, "Tracking of partials for additive sound synthesis using hiddenMarkov models," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-93*, vol. 1, 1993.

[34] M. Drobner, *Instrumentoznawstwo i akustyka*, 7th ed. Polskie Wydawnictwo Muzyczne, 2004.

[35] B. Edler and H. Purnhagen, "Concepts for Hybrid Audio Coding Schemes Based on Parametric Techniques," *Proc. 105th*, p. 4808, 1998.

[36] B. Edler and H. Purnhagen, "Parametric audio coding," in *5th International Conference on Signal Processing Proceedings, (WCCC-ICSP 2000)*, vol. 1, 2000.

[37] B. Edler, H. Purnhagen, and C. Ferekidis, "ASAC–analysis/synthesis audio codec for very low bit rates," in *100th Convention of the Audio Engineering Society*, Copenhagen, Denmark, May 1996.

[38] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, 1984.

[39] Y. Ephraim and D. Malah, "Speech enhancement using a minimun mean-square error log-spectral amplitude estimator," *IEEE transactions on acoustics, speech, and signal processing*, vol. 33, no. 2, pp. 443–445, 1985.

[40] ETSI TS 102 366, *Digital Audio Compression (Enhanced AC-3) Standard*, European Telecommunications Standards Institute / European Broadcasting Union (ETSI/EBU) Std., 2005.

[41] ETSI TS 201 980, *Digital Radio Mondiale (DRM): System Specification*, European Telecommunications Standards Institute / European Broadcasting Union (ETSI/EBU) Std. 3.1.1, 2009.

[42] European Broadcasting Union, "Sound Quality Assessment Material," compact disk, Apr. 1988, a part of an EBU publication Tech 3253.

[43] C. Faller, "Parametric coding of spatial audio," Ph.D. dissertation, École Polytechnique Fédérale de Lausanne, 2004.

[44] A. Ferreira and D. Sinha, "Accurate spectral replacement," in *118th Convention of the Audio Engineering Society*, Barcelona, Spain, May 2005.

[45] ff123 and Matthew Belmonte and Gary Perlman, "Statistical analysis of listening test data." [Online]. Available: http://ff123.net/friedman/stats.html

[46] L. D. Fielder, M. Bosi, G. Davidson, M. Davis, C. Todd, and S. Vernon, "AC-2 and AC-3: Low-Complexity Transform-Based Audio Coding," in *AES Conference: Collected Papers on Digital Audio Bit-Rate Reduction*, May 1996.

[47] L. Fielder et al., "Introduction to Dolby digital plus, an enhancement to the Dolby digital coding system," in *117th Convention of the Audio Engineering Society*. San Francisco, CA, USA: Dolby Laboratories, 28 - 31 Oct. 2004.

[48] K. Fitz, L. Haken, and P. Christensen, "A new algorithm for bandwidth association in bandwidth-enhanced additive sound modeling," in *International Computer Music Conference Proceedings*, 2000, pp. 384–387.

[49] K. Fitz and L. Haken, "Lemur: A bandwidth-enhanced sinusoidal modeling system," *The Journal of the Acoustical Society of America*, vol. 103, no. 5, pp. 2756–2757, 1998. [Online]. Available: http://link.aip.org/link/?JAS/103/2756/4

[50] B. Glasberg and B. C. J. Moore, "A model of loudness applicable to time-varying sounds," *Journal of the Audio Engineering Society*, vol. 50, pp. 331–342, May 2002.

[51] M. Goodwin, "Adaptive Signal Models: Theory, Algorithms, and Audio Applications," Ph.D. dissertation, Department of Electrical Engineering and Computer Science at the University Of California, Berkeley, 1997.

[52] GPSYCHO—A LGPL'd Psycho-Acoustic Model. [Online]. Available: http://lame.sourceforge.net/gpsycho.php

[53] S. Hainsworth, M. Macleod, and P. Wolfe, "Analysis of reassigned spectrograms for musical transcription," in *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, 2001, pp. 23–26.

[54] L. Hanzo, F. Somerville, and J. Woodard, *Voice and audio compression for wireless communications*. Wiley-IEEE Press, 2007.

[55] A. Härmä, M. Karjalainen, L. Savioja, V. Välimäki, U. Laine, and J. Huopaniemi, "Frequency-warped signal processing for audio applications," *Journal of the Audio Engineering Society*, vol. 48, no. 11, pp. 1011–1031, 2000.

[56] A. Harma and U. K. Laine, "A comparison of warped and conventional linear predictive coding," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 579–588, Jul. 2001.

[57] H.-W. Hsu, W.-C. Lee, and C.-M. Liu, "High Frequency Reconstruction by Linear Extrapolation," in *115th Convention of the Audio Engineering Society*, New York, US, 10 – 13 Oct. 2003.

[58] H.-W. Hsu, Y.-C. Yang, C.-M. Liu, , and W.-C. Lee, "Design for High Frequency Adjustment Module in MPEG-4 HE AAC Encoder based on Linear Prediction Method," in *120th Convention of the Audio Engineering Society*. Paris, France: PSPLab, Computer Science and Information Engineering, National Chiao-Tung University, Hsin-Chu, 33050, Taiwan, 20 – 23 May 2006.

[59] H. Hsu, C. Liu, W. Lee, and Z. Li, "Audio patch method in MPEG-4 HE AAC decoder," in *117th Convention of the Audio Engineering Society*, San Francisco, USA, 28 – 31 Oct. 2004.

[60] iBiquity Digital Corporation, "HD Digital Radio," www.hdradio.com/.

[61] IEC/CD 1672, *Electroacoustics-Sound Level Meters*, International Electrotechnical Commission (IEC) Std., Nov. 1996.

[62] International Standard ISO 532:1975, *Acoustics – Method for calculating loudness level*, International Organization for Standardization Std.

[63] International Standard ISO/IEC 11172-3, *Coding of moving pictures and associated audio for digital storage media at up to about 1,5 Mbit/s – Part 3: Audio*, International Organization for Standardization / International Electrotechnical Commission Std., 1993.

[64] International Standard ISO/IEC 13818-3, *Generic Coding of Moving Pictures and Associated Audio Information - Part 3: Audio*, International Organization for Standardization / International Electrotechnical Commission Std., 1998.

[65] International Standard ISO/IEC 13818-7, *Generic Coding of Moving Pictures and Associated Audio Information: – Part 7: Advanced Audio Coding (AAC)*, International Organization for Standardization / International Electrotechnical Commission Std., 2006.

[66] International Standard ISO/IEC 14496-3 Subpart 5, *Information Technology - Coding of Audiovisual Objects – Low Bitrate Coding of Multimedia Objects, Part 3: Audio, Subpart 5: Structured Audio*, International Organization for Standardization / International Electrotechnical Commission Std., 2009.

[67] International Standard ISO/IEC 14496-3:2001, *Coding of Audio-Visual Objects: Audio*, International Organization for Standardization / International Electrotechnical Commission Std., 2001.

[68] International Standard ISO/IEC 14496-3:2001/Amd.1:2003, *Bandwidth Extension*, International Organization for Standardization / International Electrotechnical Commission Std., 2003.

[69] International Standard ISO/IEC 14496-3:2001/Amd:2004, *Coding of Audio-Visual Objects. Parts 3: Audio, AMENDMENT 2" parametric coding of high quality audio*, International Organization for Standardization / International Electrotechnical Commission Std., Jul. 2004.

[70] International Standard ISO/IEC 15938-1:2002, *Information technology – Multimedia content description interface – Part 1: Systems*, International Organization for Standardization / International Electrotechnical Commission Std., 2002.

[71] B. Iser, W. Minker, and G. Schmidt, *Bandwidth Extension of Speech Signals*, ser. Lecture Notes in Electrical Engineering. Springer US, 2008, vol. 13.

[72] ISO/IEC JTC1/SC29/WG11, "Call for Proposals on Unified Speech and Audio Coding," International Organization for Standardization / International Electrotechnical Commission, Shenzhen, Doc. n9519, Oct. 2007.

[73] ISO/IEC JTC1/SC29/WG11, *Final Spatial Audio Object Coding Evaluation Procedures and Criterion*, International Organization for Standardization / International Electrotechnical Commission Doc. w9099, Apr. 2007.

[74] ISO/IEC JTC1/SC29/WG11, "Working Draft 6 of Unified Speech and Audio Coding," International Organization for Standardization / International Electrotechnical Commission, Kyoto, Japan, Doc. n11213, Jan. 2010.

[75] ISO/IEC 14496-10 (MPEG-4 AVC) / ITU-T Rec. H.264, *Advanced Video Coding for Generic Audiovisual Services*, International Organization for Standardization / International Electrotechnical Commission Std., 2003–2007.

[76] ITU-R Recommendation BS.1116-1, *Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems*, International Telecommunication Union Std., 1994.

[77] ITU-R Recommendation BS.1387-1, *method for Objective Measurements of Perceived Audio Quality*, International Telecommunication Union Std., 1999.

[78] ITU-R Recommendation BS.1534-1, *Method for the subjective assessment of intermediate quality levels of coding systems*, International Telecommunication Union Std., Jan. 2003.

[79] ITU-T Recommendation P.800, *Methods for subjective determination of transmission quality*, International Telecommunication Union Std., 1996.

[80] ITU-T Recommendation P.861, *Objective quality measurement of telephone-band (300-3400 Hz) speech codecs*, International Telecommunication Union Std., 1996.

[81] ITU-T Recommendation P.862, *Perceptual Evaluation Of Speech Quality (PESQ): An Objective Method For End-to-end Speech Quality Assessment Of Narrow-band Telephone Networks And Speech Codecs*, International Telecommunication Union Std., 2001.

[82] P. Jax and P. Vary, "Bandwidth extension of speech signals: a catalyst for the introduction of wideband speech coding?" *IEEE Commun. Mag.*, vol. 44, no. 5, pp. 106–111, 2006.

[83] N. Jayant, J. Johnston, and R. Safranek, "Signal compression based on models of human perception," *Proceedings of the IEEE*, vol. 81, no. 10, pp. 1385–1422, 1993.

[84] N. Jayant and P. Noll, *Digital coding of waveforms: principles and applications to speech and video*. Prentice Hall Professional Technical Reference, 1990.

[85] K. Jensen, "The timbre model," *Journal of the Acoustical Society of America*, vol. 112, no. 5, pp. 2238–2238, 2002.

[86] J. D. Johnston, "Estimation of perceptual entropy using noise masking criteria," in *International Conference on Acoustics, Speech, and Signal Processing, ICASSP-88*, vol. 5, 1988, pp. 2524–2527.

[87] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE Journal on selected areas in communications*, vol. 6, no. 2, pp. 314–323, 1988.

[88] J. D. Johnston, "Perceptual transform coding of wideband stereo signals," in *International Conference on Acoustics, Speech, and Signal Processing, ICASSP-89*, 1989, pp. 1993–1996.

[89] J. D. Johnston, D. Sinha, S. Dorward, and S. R. Quackenbush, "AT&T Perceptual Audio Coding (PAC)," in *AES Conference: Collected Papers on Digital Audio Bit-Rate Reduction*, May 1996.

[90] P. Kabal and R. P. Ramachandran, "The computation of line spectral frequencies using chebyshev polynomials," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 6, pp. 1419–1426, 1986.

[91] S. Kay, "Statistically/computationally efficient frequency estimation," in *International Conference on Acoustics, Speech, and Signal Processing, ICASSP-88*, vol. 4, Apr. 1988, pp. 2292–2295.

[92] F. Keiler and S. Marchand, "Survey on extraction of sinusoids in stationary sounds," in *Proc. DAFx*, 2002, pp. 51–58.

[93] A. Klapuri, "Wide-band pitch estimation for natural sound sources with inharmonicities," in *106th Convention of the Audio Engineering Society*, Munchen, 1999.

[94] B. Kostek, *Perception-based data processing in acoustics: applications to music information retrieval and psychophysiology of hearing*, ser. Studies in Computational Intelligence. Springer Verlag, 2005, vol. 3.

[95] E. Kruger and H. W. Strube, "Linear prediction on a warped frequency scale [speech processing]," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 9, pp. 1529–1531, 1988.

[96] M. Kulesza and A. Czyzewski, "Tonality Estimation and Frequency Tracking of Modulated Tonal Components," *Journal of the Audio Engineering Society*, vol. 57, no. 4, pp. 221–236, Apr. 2009.

[97] A. Laaksonen, "Bandwidth extension in high-quality audio coding," Master's thesis, Helsinki University Of Technology, 2005, praca powstała we współpracy z Centrum badawczym firmy Nokia. [Online]. Available: http://lib.tkk.fi/Dipl/2005/urn007914.pdf

[98] M. Lagrange, S. Marchand, M. Raspaud, and J. Rault, "Enhanced Partial Tracking Using Linear Prediction," in *Proceedings of the Digital Audio Effects (DAFx03) Conference*, vol. 141, 2003, p. 146.

[99] M. Lagrange, S. Marchand, and J.-B. Rault, "Enhancing the tracking of partials for the sinusoidal modeling of polyphonic sounds," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 5, pp. 1625–1634, July 2007.

[100] M. Lagrange and M. Raspaud, "Spectral Similarity Metrics for Sound Source Formation Based on the Common Variation Cue," in *Proceedings of the 2009 Seventh International Workshop on Content-Based Multimedia Indexing*. IEEE Computer Society, 2009, pp. 56–61.

[101] M. Lagrange, "A New Dissimilarity Metric for the Clustering of Partials Using the Common Variation Cue," in *Proceedings of the International Computer Music Conference (ICMC'05)*, 2005.

[102] U. Laine, M. Karjalainen, and T. Altosaar, "Warped linear prediction (WLP) in speech and audio processing," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-94*, vol. 3, 1994.

[103] E. Larsen and R. Aarts, *Audio Bandwidth Extension: Application of Psychoacoustics, Signal Processing and Loudspeaker Design*. John Wiley and Sons Ltd, 2004.

[104] E. Larsen, R. Aarts, and M. Danessis, "Efficient high-frequency bandwidth extension of music and speech," in *112th Convention of the Audio Engineering Society*, München, Germany, 10 – 13 May 2002.

[105] S. Levine, "Audio Representations for Data Compression and Compressed Domain Processing," Ph.D. dissertation, Department of Electrical Engineering of Stanford University, 1998.

[106] J. Lim and A. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[107] C.-M. Liu, H.-W. Hsu, and W.-C. Lee, "Compression Artifacts in Perceptual Audio Coding," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 4, pp. 681–695, 2008.

[108] C. Liu, W. Lee, and H. Hsu, "High frequency reconstruction for band-limited audio signals," in *6th International Conference on Digital Audio Effects (DAFx-03)*, Queen Mary, University of London, 8–11 Sep. 2003.

[109] N. Lyubimov and A. Lukin, "Audio Bandwidth Extension Using Cluster Weighted Modeling of Spectral Envelopes," in *127th Convention of the Audio Engineering Society*, ser. Preprint 7946, New York, NY, USA, Oct. 2009.

[110] H. S. Malvar, "Lapped transforms for efficient transform/subband coding," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 6, pp. 969–978, 1990.

[111] S. McAdams, "Segregation of concurrent sounds: I. Effects of frequency modulation coherence." *Journal of the Acoustical Society of America*, vol. 86, no. 6, pp. 2148–2159, 1989.

[112] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, no. 2, pp. 137–145, apr 1980.

[113] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 4, pp. 744–754, 1986.

[114] I. V. McLoughlin, "Line spectral pairs," *Signal Processing 88*, p. 448–467, 2008.

[115] R. Meddis and M. Hewitt, "Virtual pitch and phase sensitivity of a computer model of the auditory periphery: II. Phase sensitivity." *Journal of the Acoustical Society of America*, vol. 89, no. 6, pp. 2883–2894, 1991.

[116] S. Meltzer and G. Moser, "MPEG-4 HE-AAC v2 - audio coding for today's digital-media world," European Broadcasting Union (EBU), EBU Technical Review, Jan. 2006.

[117] Moving Picture Experts Group (MPEG), "ISO / IEC, WG11," http://www.chiariglione.org/mpeg/index.htm, 1998.

[118] M. G. MUZZI, "Amélioration d'un codeur paramétrique," Philips Digital Systems Laboratories, Ecole Nationale Superieure des Telecomunications Paris 6 / Philips Eindhoven, Tech. Rep. Report DEA ATIAM 2002 – 2003, 2003.

[119] F. Myburg, "Design of a scalable parametric audio coder," Ph.D. dissertation, Eindhoven University of Technology, 2004.

[120] F. Nagel and S. Disch, "A harmonic bandwidth extension method for audio codecs," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Taipei, Taiwan: IEEE Computer Society, 19-24 Apr. 2009, pp. 145–148.

[121] A. V. Oppenheim and D. H. Johnson, "Discrete representation of signals," *Proc. IEEE*, vol. 60, no. 6, pp. 681–691, Jun. 1972.

[122] E. Ozimek, *Dźwięk i jego percepcja. Aspekty fizyczne i psychoakustyczne*. Warszawa–Poznan: Wydawnictwo Naukowe PWN, 2002.

[123] T. Painter, A. Spanias *et al.*, "Perceptual coding of digital audio," *Proc. IEEE*, vol. 88, no. 4, pp. 451–513, 2000.

[124] M. Parfieniuk and A. Piotrowski, *Cyfrowe banki filtrów: analiza, synteza i implementacja dla systemów multimedialnych*, i ed. Wydawnictwo Politechniki Białostockiej, 2006.

[125] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the CUIDADO project." Ircam, Analysis/Synthesis Team, Tech. Rep., 2004.

[126] R. Plomp and H. Steeneken, "Effect of phase on the timbre of complex tones," *The Journal of the Acoustical Society of America*, vol. 46, p. 409, 1969.

[127] J. Princen and A. Bradley, "Analysis/synthesis filter bank design based on time domain aliasing cancellation," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 34, no. 5, pp. 1153–1161, 1986.

[128] J. Princen, A. Johnson, and A. Bradley, "Subband/transform coding using filter bank designs based on time domain aliasing cancellation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP'87*, vol. 12, 1987, pp. 2161–2164.

[129] Projekt międzynarodowej normy Pr PN-ISO/IEC 13818-7, *Technika informatyczna. Kodowanie obrazów ruchomych i towarzyszącej im informacji dźwiękowej. Część 7: Zaawansowane kodowanie dźwięku (AAC).*, Polski Komitet Normalizacyjny Std., 2009.

[130] H. Purnhagen, "Advances in parametric audio coding," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 1999, pp. 31–34.

[131] H. Purnhagen and B. Edler, "Error Protection and concealment fo HILN MPEG-4 parametric audio coding," in *110th Convention of the Audio Engineering Society*, ser. AES Preprint 5300, Amsterdam, 12-15 May 2001.

[132] H. Purnhagen and N. Meine, "HILN-the MPEG-4 parametric audio coding tools," in *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS)*, vol. 3, Geneva, May 2000, pp. 201–204.

[133] A. Roebel, M. Zivanovic, and X. Rodet, "Signal decomposition by means of classification of spectral peaks," in *Proceedings of the International Computer Music Conference*, 2004, pp. 446–449.

[134] S.-U. Ryu and K. Rose, "Enhanced Accuracy of the Tonality Measure and Control Parameter Extraction Modules in MPEG-4 HE-AAC," in *119th Convention of the Audio Engineering Society*, New York, NY, USA, 7 – 10 Oct. 2005.

[135] S.-U. Ryu, K. Rose, and J.-H. Chang, "Effective High Frequency Regeneration Based On Sinusoidal Modeling For MPEG-4 HE-AAC," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, 16 –19 Oct. 2005.

[136] B. Scharf, "Critical bands," *Foundations of modern auditory theory*, vol. 1, pp. 159–202, 1970.

[137] E. Scheirer, "The MPEG-4 structured audio standard," in *IEEE International Conference on Acoustics Speech and Signal Processing*, vol. 6, 1998.

[138] E. Schuijers, W. Oomen, B. Brinker, and J. Breebaart, "Advances in parametric coding for high-quality audio," *PREPRINTS-AUDIO ENGINEERING SOCIETY*, 2003.

[139] X. Serra, "A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition," Ph.D. dissertation, Stanford University, Department of Music, 1989.

[140] X. Serra, *Musical Signal Processing*. Swets & Zeitlinger, 1997, ch. Musical Sound Modeling with Sinusoids plus Noise, pp. 91–122.

[141] X. Serra and J. Smith III, "Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Computer Music Journal*, vol. 14, no. 4, pp. 12–24, 1990.

[142] D. Sinha, A. Ferreira, and D. Sen, "A Fractal Self-Similarity Model for the Spectral Representation of Audio Signals," in *118th Convention of the Audio Engineering Society*, Barcelona, Spain, May 2005.

[143] D. Sinha, A. Ferreira, and H. E. V., "A Novel Integrated Audio Bandwidth Extension Toolkit (ABET)," in *120th Convention of the Audio Engineering Society*, Paris, France, May 2006.

[144] K. Skretting, J. Husøy, and S. Aase, "Improved Huffman Coding Using Recursive Splitting," in *Proceedings of Norwegian Signal Processing, NORSIG*, Asker, Norway, Sep. 1999.

[145] K. Skretting, J. Husøy, and S. Aase, "Arithmetic coding and huffman coding in MATLAB," [Online] Available: http://www.ux.uis.no/~karlsk/proj99/index.html, http://www.mathworks.com/matlabcentral/fileexchange/2818-huffman-coding-and-arithmetic-coding, Jun. 2001.

[146] J. Smith and X. Serra, "PARSHL: A program for the analysis/synthesis of inharmonic sounds based on a sinusoidal representation," in *International Computer Music Conference*, 1987.

[147] G. Stoll and F. Kozamernik, "EBU listening tests on Internet audio codecs," European Broadcasting Union (EBU), EBU Technical Review, Jun. 2000.

[148] H. Strube, "Linear prediction on a warped frequency scale," *The Journal of the Acoustical Society of America*, vol. 68, p. 1071, 1980.

[149] R. Tadeusiewicz, *Sygnał mowy*, ser. Problemy Elektroniki i Telekomunikacji. Wydawnictwo Komunikacji i Łączności, Warszawa 1988.

[150] D. Thomson, "Spectrum estimation and harmonic analysis," *Proceedings of the IEEE*, vol. 70, no. 9, pp. 1055–1096, 1982.

[151] University of Iowa Electronic Music Studios, "Musical Instrument Samples Database," www: http://theremin.music.uiowa.edu/, 1997. [Online]. Available: http://theremin.music.uiowa.edu/

[152] T. Verma and T. Meng, "A 6kbps to 85kbps scalable audio coder," in *IEEE International Conference On Acoustics Speech And Signal Processing*, vol. 2, 2000.

[153] *PlusV Specification*, VLSI Solution, Dec. 2001.

[154] *PlusV Reference Code Filters*, VLSI Solution, Oct. 2001.

[155] M. Walesiak and E. Gatnar, *Statystyczna analiza danych z wykorzystaniem programu R.* Warszawa: Wydawnictow naukowe PWN, 2009.

[156] S. Wang, A. Sekey, A. Gersho, T. Syst, and C. Berkeley, "An objective measure for predicting subjective quality of speechcoders," *IEEE J. Sel. Areas Commun.*, vol. 10, no. 5, pp. 819–829, 1992.

[157] XM Satellite Radio, http://www.xmradio.com/.

[158] W. Xue and M. Sandler, "Sinusoid modeling in a harmonic context," in *Proc. 10th Int. Conference on Digital Audio Effects*, Bordeaux, 2007.

[159] W. Yang, M. Benbouchta, and R. Yantorno, "Performance of the modified Bark spectral distortion as anobjective speech quality measure," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, 1998.

[160] Tomasz Zieliński, *Od teorii do cyfrowego przetwarzania sygnałów*, I ed. Wydział EAIiE AGH, 2002.

[161] E. Zwicker and H. Fastl, *Psychoacoustics: facts and models.* Springer-Verlag New York Inc, 2007.

[162] R. Zieliński and W. Zieliński, *Tablice statystyczne.* PWN, 1972.

[163] Żernicki, T. and Bartkowiak, M., "Audio bandwidth extension by frequency scaling of sinusoidal partials," in *125th Convention of the Audio Engineering Society*, ser. AES Preprint 7622, San Francisco, USA, 2-5 Oct. 2008.

[164] Żernicki, T. and Bartkowiak, M., "Zastosowanie modelowania sinusoidalnego do regeneracji wysokoczęstotliwościowych składowych tonalnych kodera MPEG-4 HE-AAC", *Krajowa Konferencja Radiokomunikacji, Radiofonii i Telewizji, KKRRiT*, Wrocław, Polska, 9-11 Kwiecień 2008, s. 511–514.

[165] Żernicki, T. and Domański, M., "Improved coding of tonal components in MPEG-4 AAC with SBR," in *Proceedings of IX European Signal Processing Conference, EUSIPCO'08*, Lausanne, Switzerland, Aug. 25-29 2008.

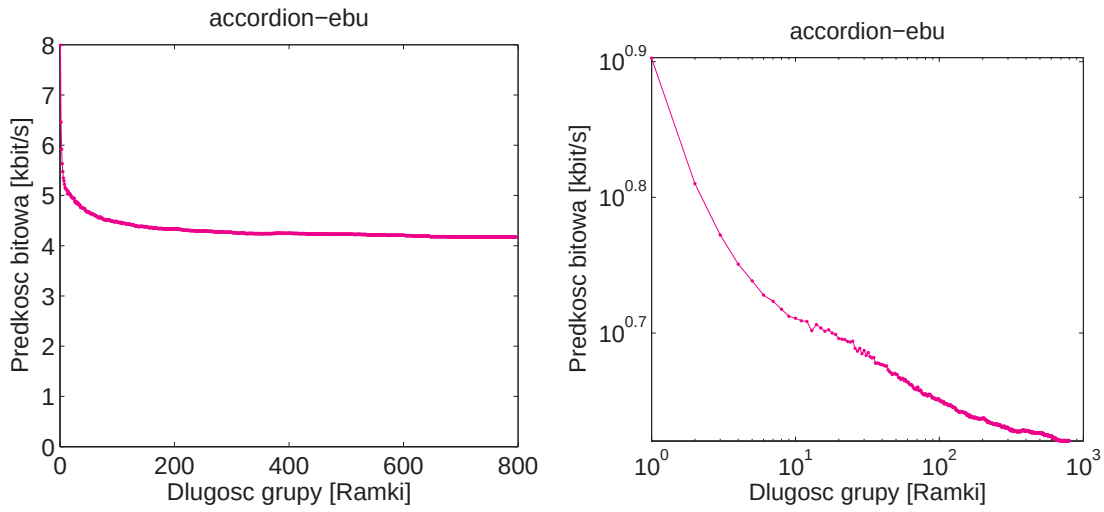
Dodatki

Dodatek A

Kodowanie parametrów modelu sinusoidalnego

A.1. Badanie zależność prędkości bitowej od długości bloku ramek objętych kodowaniem tą samą tablicą Huffmana

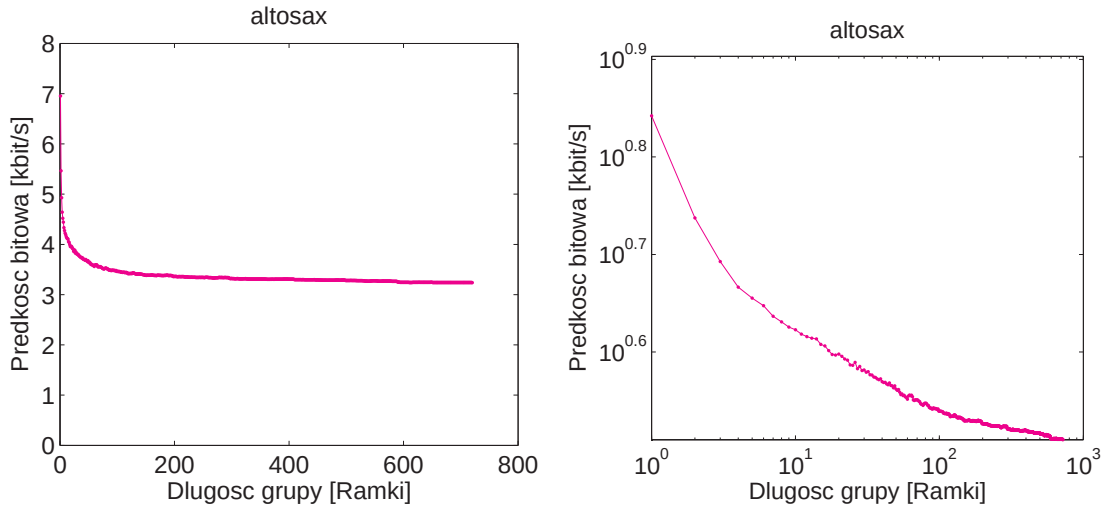
Dokładny opis eksperymentu został przedstawiony w punkcie 3.7.4. Długość okna analizy wynosi 2048 próbek, natomiast skok 512 próbek. Częstotliwość próbkowania f_S



(a) Skala liniowa

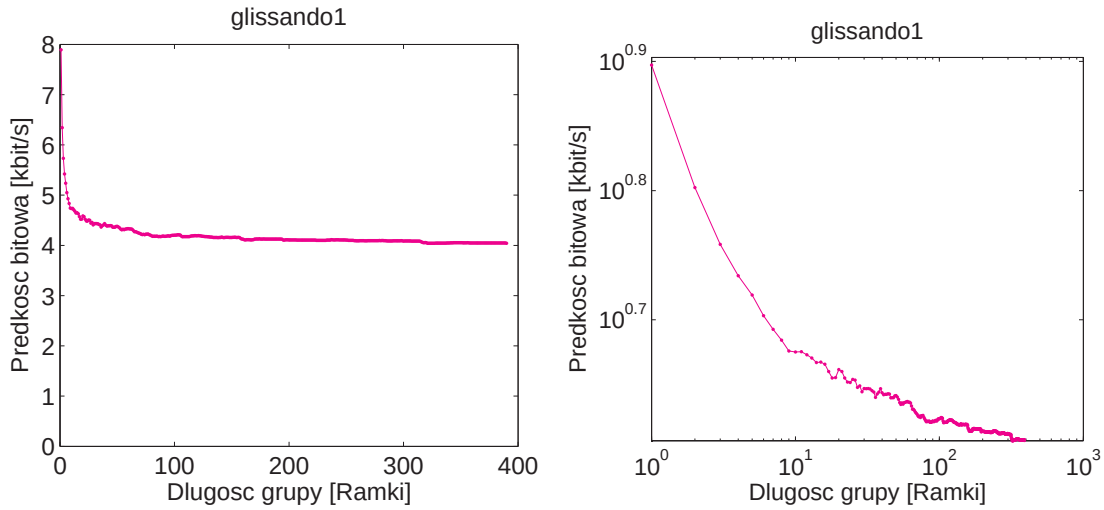
(b) Skala logarytmiczna loglog

Rysunek A.1: Zależność prędkości bitowej od długości bloku ramek objętych kodowaniem tą samą tablicą Huffmana



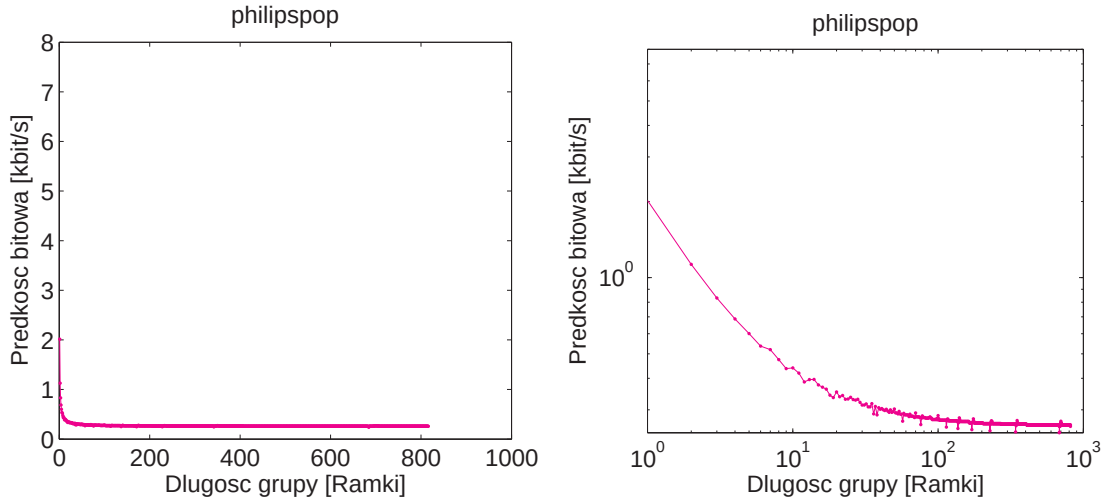
(a) Skala liniowa (b) Skala logarytmiczna loglog

Rysunek A.2: Zależność prędkości bitowej od długości bloku ramek objętych kodowaniem tą samą tablicą Huffmana



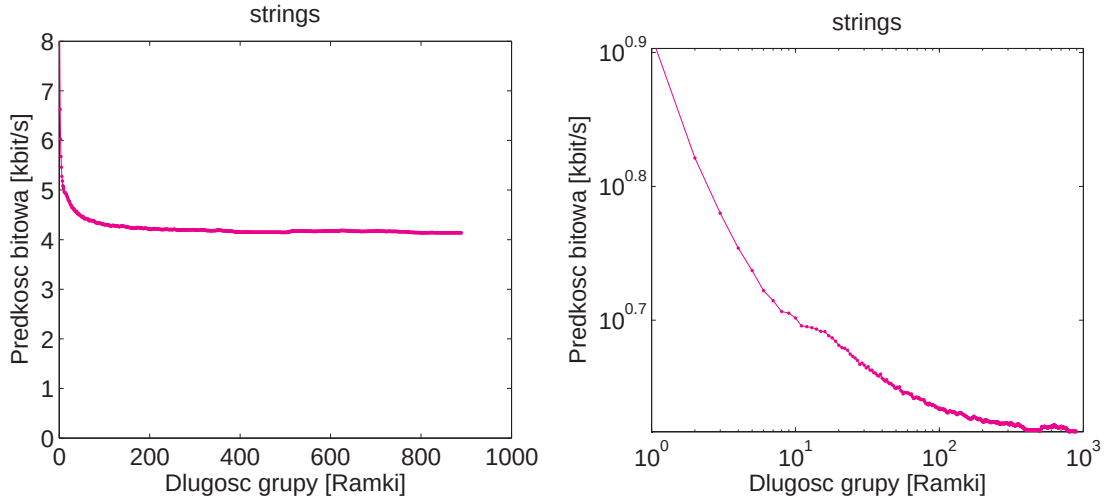
(a) Skala liniowa (b) Skala logarytmiczna loglog

Rysunek A.3: Zależność prędkości bitowej od długości bloku ramek objętych kodowaniem tą samą tablicą Huffmana



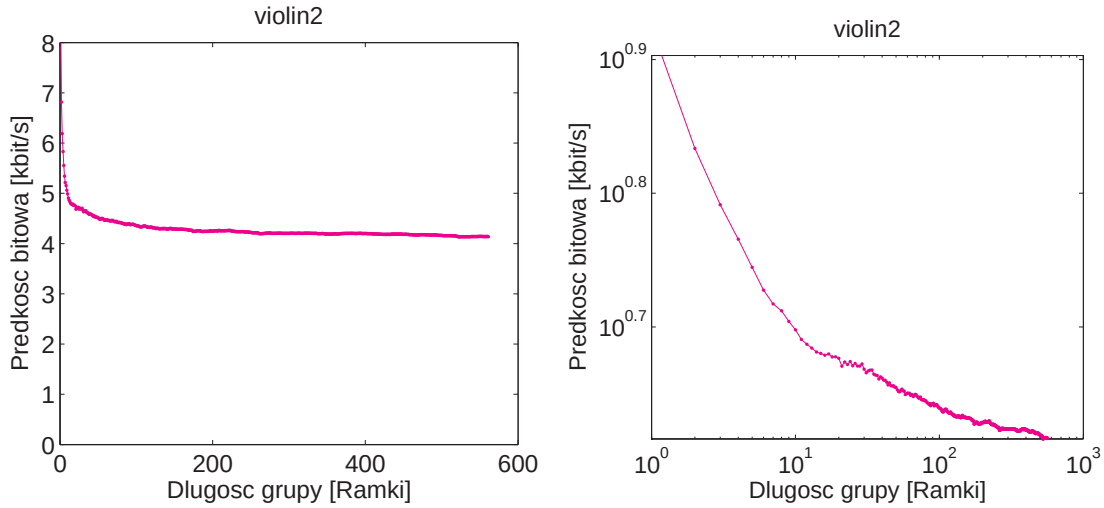
(a) Skala liniowa (b) Skala logarytmiczna loglog

Rysunek A.4: Zależność prędkości bitowej od długości bloku ramek objętych kodowaniem tą samą tablicą Huffmana



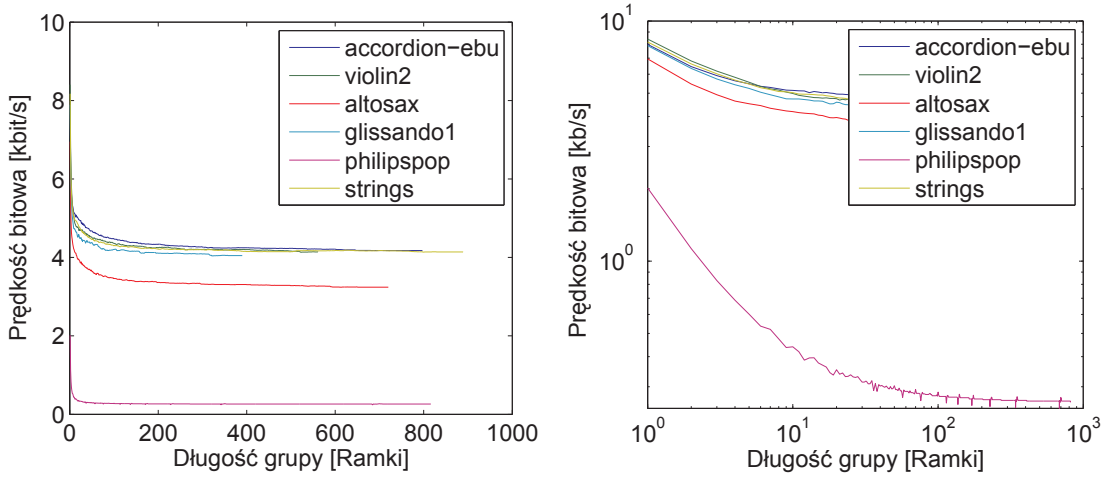
(a) Skala liniowa (b) Skala logarytmiczna loglog

Rysunek A.5: Zależność prędkości bitowej od długości bloku ramek objętych kodowaniem tą samą tablicą Huffmana



(a) Skala liniowa (b) Skala logarytmiczna loglog

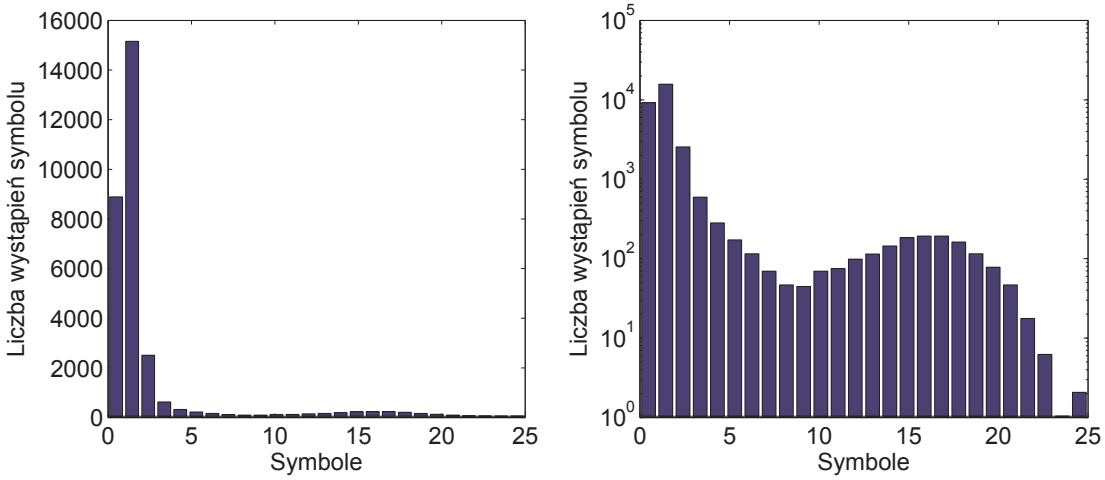
Rysunek A.6: Zależność prędkości bitowej od długości bloku ramek objętych kodowaniem tą samą tablicą Huffmana



(a) Skala liniowa (b) Skala logarytmiczna loglog

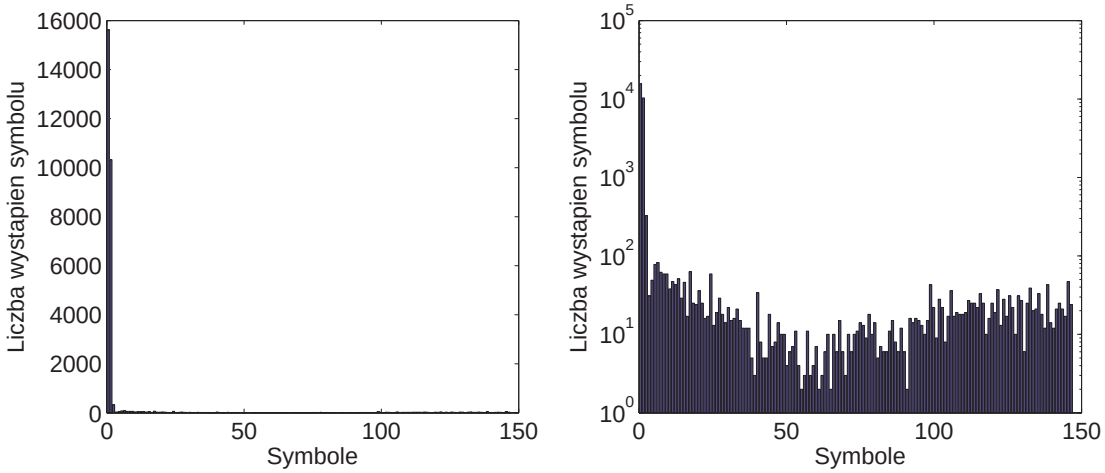
Rysunek A.7: Zależność prędkości bitowej od długości bloku ramek objętych kodowaniem tą samą tablicą Huffmana. Porównanie dla wszystkich sygnałów fonicznych biorących udzał w badaniu.

A.2. Badanie rozkładu funkcji prawdopodobieństwa dla skwantowanych parametrów modelu sinusoidalnego



(a) Skala liniowa (b) Skala logarytmiczna loglog

Rysunek A.8: Liczba wystąpień symboli dla skwantowanych wartości amplitudy.



(a) Skala liniowa (b) Skala logarytmiczna loglog

Rysunek A.9: Liczba wystąpień symboli dla skwantowanych wartości częstotliwości.

Dodatek B

Implementacja modelowania sinusoidalnego

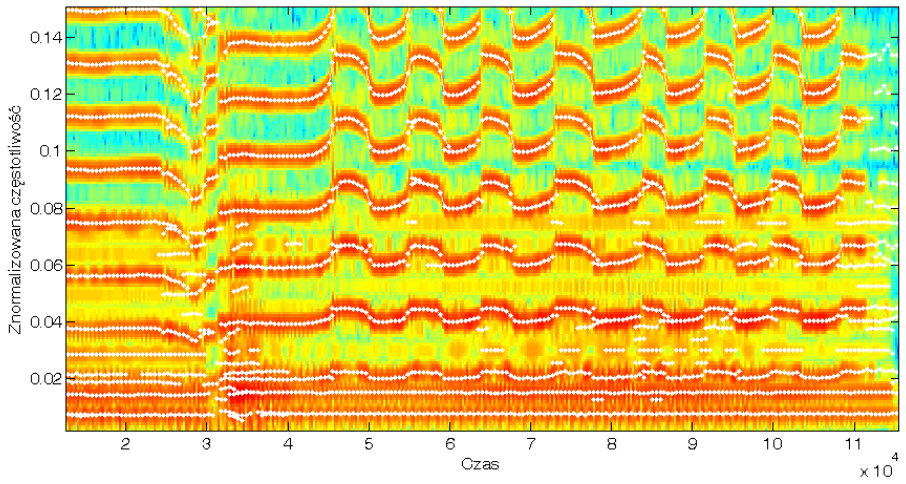
Opisany poniżej narzędzie modelowania sinusoidalnego oraz kodek sinusoidalny powstał we współpracy z Maciejem Bartkowiakiem w ramach projektu badawczego MNiSW 3T11D 01730, prowadzonego w Katedra Telekomunikacji Multimedialnej i Mikroelektroniki na Politechnice Poznańskiej.

B.1. Analiza sinusoidalna

Schemat modelu sinusoidalnego przedstawiony jest na rysunku 2.3. W modelowaniu sinusoidalnym w celu selekcji składowych tonalnych zastosowano przekształcenie STFT, dla którego długość okna analizy wynosi $N = 2048$ próbek, natomiast przesunięcie pomiędzy sąsiednimi oknami analizy wynosi $H = 512$ próbek. Parametry N i H zostały dobrane doświadczalnie i stanowią kompromis pomiędzy dokładnością estymacji parametrów sinusoid, a czasem obliczeń niezbędnym do ich wykonania. Parametry sinusoid - amplituda A i częstotliwość f wyznaczane są na podstawie analizy QIFFT opisanej w punkcie 3.4.2 w danym oknie STFT.

B.2. Śledzenie trajektorii sinusoidalnych

Wyznaczone podczas analizy sekwencje parametrów chwilowych amplitudy i częstotliwości łączone są w trajektorie sinusoidalne (rysunek B.1). Sposób tworzenia trajektorii zależy



Rysunek B.1: Spektrogram sygnału z zaznaczonymi białą linią trajektoriami sinusoidalnymi.

od *miary dopasowania połączenia*, które określa w jaki sposób mają być łączone sinusoidy pochodzące z widm sąsiednich ramek czasowych. Połączenie, którego wartość miary jest najwyższa uznawane za najbardziej prawdopodobne:

$$Q^f_{p,q} = \frac{1}{1 + \frac{|f_{LPC,p,k} - f_{q,k}|}{f_{p,k-1} + f_{q,k}}} \tag{B.1}$$

gdzie:

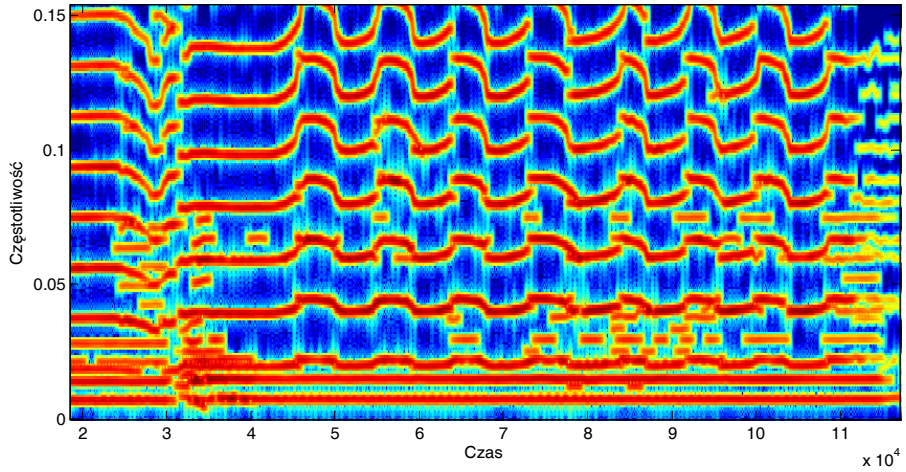
k – indeks ramki,

p – indeks sinusoidy w ramce bieżącej (k),

q – indeks sinusoidy w ramce poprzedniej (k-1),

$f_{LPC,p,k}$ – jest wartością częstotliwości przewidzianą na podstawie predykcji liniowej metodą Burga opisanej w pracy [98]. Predyktor 6-tego rzędu wyznacza nową wartość na podstawie 20 poprzednich punktów trajektorii.

Algorytm śledzenia wyznacza wartości miar dopasowania dla wszystkich możliwych połączeń trajektorii sinusoidalnych w ramce $k-1$ z sinusoidami w bieżącej ramce k. Punkty o największych wartościach miary $Q^f_{p,q}$ są ze sobą łączone tworząc trajektorię sinusoidalną. Dodatkowo algorytm śledzenia nie dopuszcza do łączenia więcej niż jednej trajektorii w jednym punkcie. W ten sposób otrzymywany jest zbiór trajektorii sinusoidalnych P zawierający trj trajektorii sinusoidalnych o pogrupowanych wartościach amplitud i częstotliwości.



Rysunek B.2: Synteza sinusoid na podstawie wyznaczonych trajektorii sinusoidalnych.

B.3. Synteza sygnału

Na wejściu algorytmu syntezy sinusoid znajdują się parametry chwilowe amplitudy i częstotliwości, które opisują stan trajektorii w krótkim odcinku czasu. Należy więc interpolować wartości próbek występujących między węzłami trajektorii, aby uzyskać sygnał o częstotliwości próbkowania oryginału. W tym celu wykorzystano interpolację funkcjami sklejanymi 3-go stopnia (ang. *cubic spline*). Metoda ta pozwala na lepsze odwzorowanie zmian w sygnale w porównaniu do interpolacji liniowej. Sygnał syntezowany jest na podstawie wzoru:

$$s(n) = \sum_n P_k(n) = \sum a_k(n) \cos\left(2\pi \sum_{m=0}^{m} f_k(m)\right), \tag{B.2}$$

gdzie:

k – indeks trajektorii,

$a_k(n), f_k(n)$ – parametry k-tej trajektorii,

$n = 0, \ldots, H-1$ - numer próbki w ramce syntezy,

H – długość ramki syntezy (skok okna analizy w algorytmie STFT).

Przykład syntezy pokazano na rysunku B.2.

B.4. Kompresja parametrów modelu sinusoidalnego

B.4.1. Kodowanie trajektorii sinusoidalnych

Parametry trajektorii sinusoidalnych (amplitudy i częstotliwości) kodowane są przy użyciu tego samego algorytmu (rysunek B.3), który składa się z następujących etapów:

- Niezależne kodowanie każdej trajektorii przy pomocy techniki ADPCM. Parametry LPC są takie same jak dla algorytmu śledzącego trajektorie sinusoidalne B.2 (predyktor 6 rzędu i 20 pamiętanych próbek). Kodowany jest błąd predykcji e_n, gdzie n jest indeksem parametru w danej trajektorii. W zastosowanej implementacji predyktor jest zerowany co 250 ramek, w celu zapewnienia dostępu do strumienia bitowego.

- Początki trajektorii kodowane są różnicowo, rozpoczynając od trajektorii reprezentujących najniższe częstotliwości. Trajektorie uszeregowane są według rosnącej średniej częstotliwości.

- Błąd predykcji e_n podany jest kwantyzacji równomiernej w skali logarytmicznej.

- Ostatnim etapem jest kodowanie entropijne. W tym celu, wykorzystano algorytm Huffmana z książką kodową, wyznaczaną osobno dla każdego sygnału fonicznego (rozdział 3.7.4).

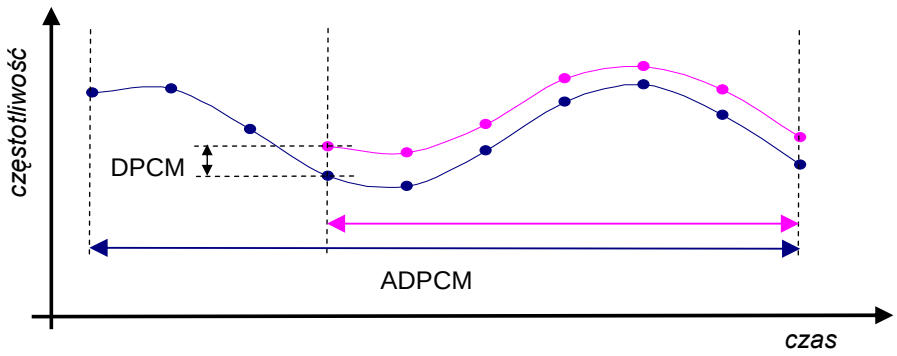
B.4.2. Kwantyzacja parametrów modelu sinusoidalnego

Częstotliwości sinusoid analizowane są w skali centowej, która odpowiada muzycznej skali wysokości dźwięku:

$$[cent] = \sqrt[1200]{2} \approx 1,0006. \tag{B.3}$$

Przeliczenie na skalę centową odbywa się na podstawie wzoru:

$$f_a = \log_{cent}\left(\frac{f}{f_{ref}}\right), \tag{B.4}$$



Rysunek B.3: Kodowanie trajektorii modelu sinusoidalnego. Ciągła linia pokazuje przebieg trajektorii sinusoidalnych. Punkty przedstawiają kodowane parametry, wyznaczone w kolejnych ramkach.

gdzie:

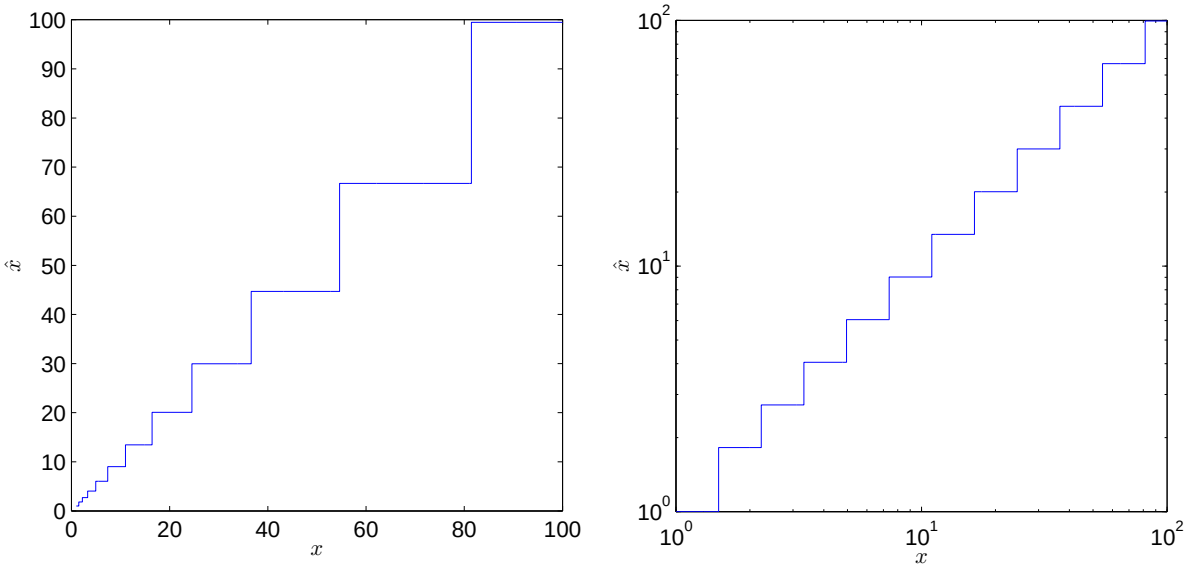
f_a – częstotliwość w skali centowej,

f – wartość częstotliwości w znormalizowanej skali liniowej ($f_s = 1$),

f_{ref} – częstotliwość odniesienia.

Częstotliwości w skali centowej poddawane są kwantyzacji równomiernej z przedziałem kwantyzacji równym 40 centów. Przedział kwantyzacji został dobrany na podstawie testów eksperymentalnych przedstawionych w rozdziale 3.9.3.

Parametry amplitud przeliczane są na skalę logarytmiczną, a następnie podlegają kwantyzacji równomiernej. Przedział kwantyzacji stosowany w implementacji modelu sinusoidalnego wynosi 3 dB (3.9.3).



(a) Skala liniowa

(b) Skala podwójnie logarytmiczna

Rysunek B.4: Zastosowany model kwantyzacji.

Zastosowanie kwantyzatora nierównomiernego jest uzasadnione właściwościami percepcyjnymi ludzkiego ucha. Właściwości te obrazują izofony [161], czyli krzywe jednakowego poziomu dźwięku. Dana krzywa izofoniczna ma tę samą wartość, na przestrzeni różnych częstotliwości, liczoną w fonach, ale różną liczoną w decybelach. W związku z tym, że zależności te mają charakter logarytmiczny, w pracy zastosowano kwantyzator równomierny w skali logarytmicznej.

Dodatek C

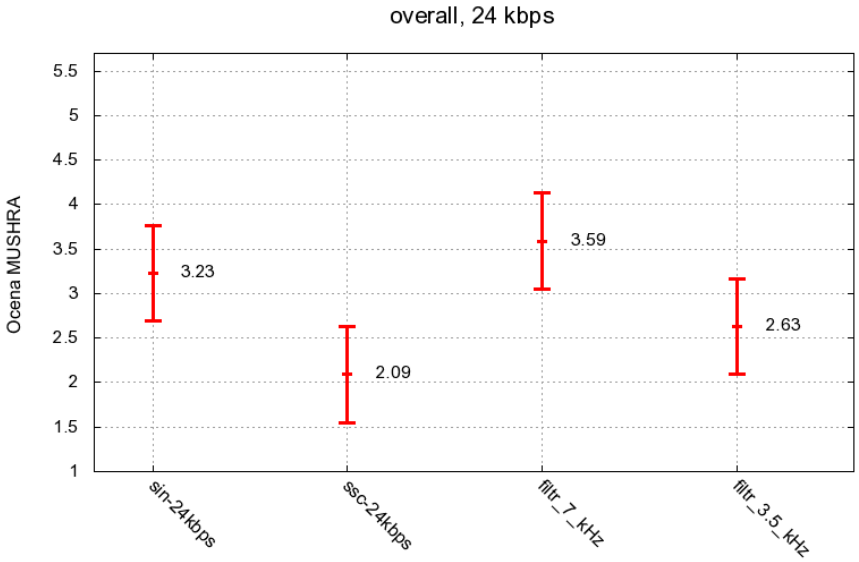
Testy odsłuchowe dla koderów sinusoidalnych

W bieżącym dodatku przedstawione zostały wyniki badania efektywności kodera sinusoidalnego skonstruowanego dla potrzeb rozprawy doktorskiej. Zaprezentowane wykresy przedstawiają wyniki badań eksperymentalnych dla kodeków sinusoidalnych:

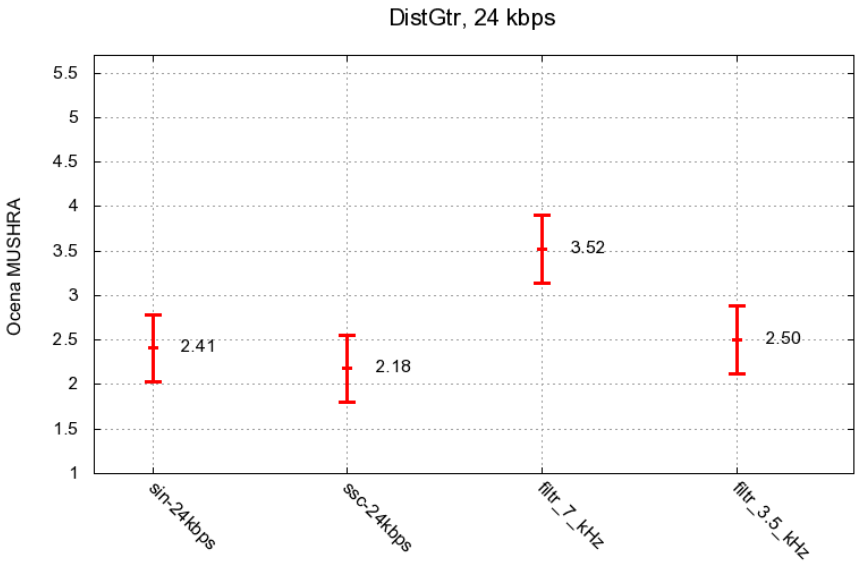
- „sin" – opisany w rozprawie oraz w dodatku B,
- „ssc" – implementacja [118] kodera MPEG-4 SSC (ang. *Sinusoidal Coding*) [69], opracowana przez firmę Philips.

Porównanie dotyczy sygnałów zakodowanych dla prędkości bitowej 24 kb/s. W testach odsłuchowych wzięło udział 15 osób. Badaniu podlegały 4 sygnały testowe pochodzące z różnych źródeł [42, 73, 151].

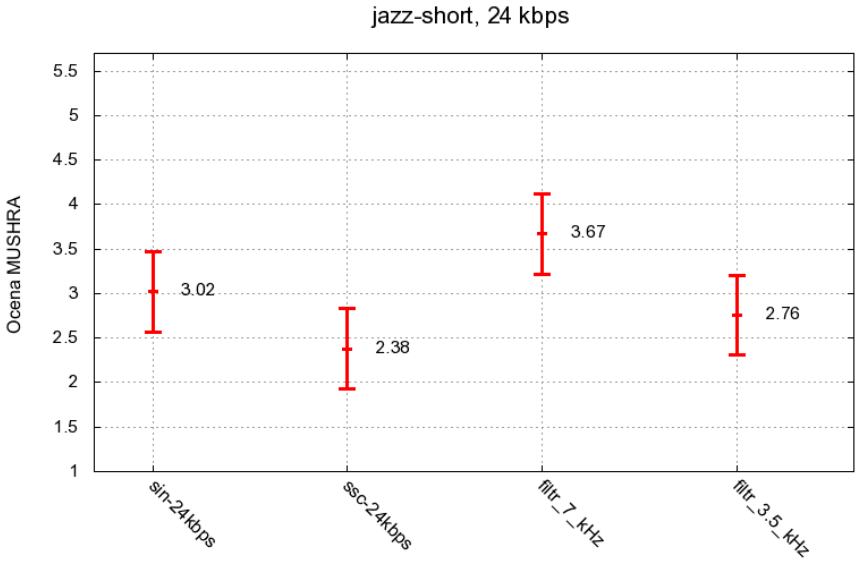
Wyniki badań eksperymentalnych pokazują, że opracowano efektywną implementację analizy i kodowania parametrów modelu sinusoidalnego. W większości przypadków własny koder sinusoidalny otrzymał wyższe oceny w teście subiektywnym niż koder MPEG-4 SSC [69] opracowany przez firmę Philips [118].



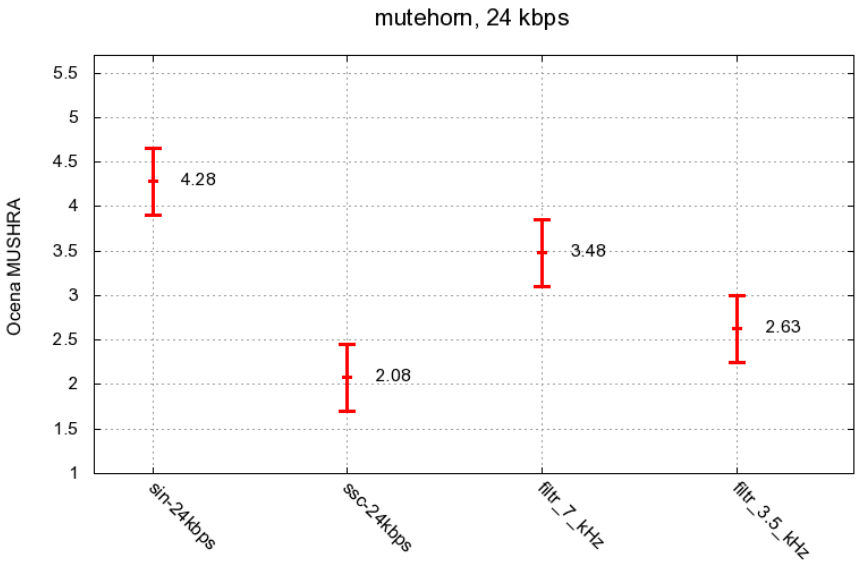
Rysunek C.1: Porównanie modelowania sinusoidalnego. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 24 kb/s z 95% przedziałem ufności.



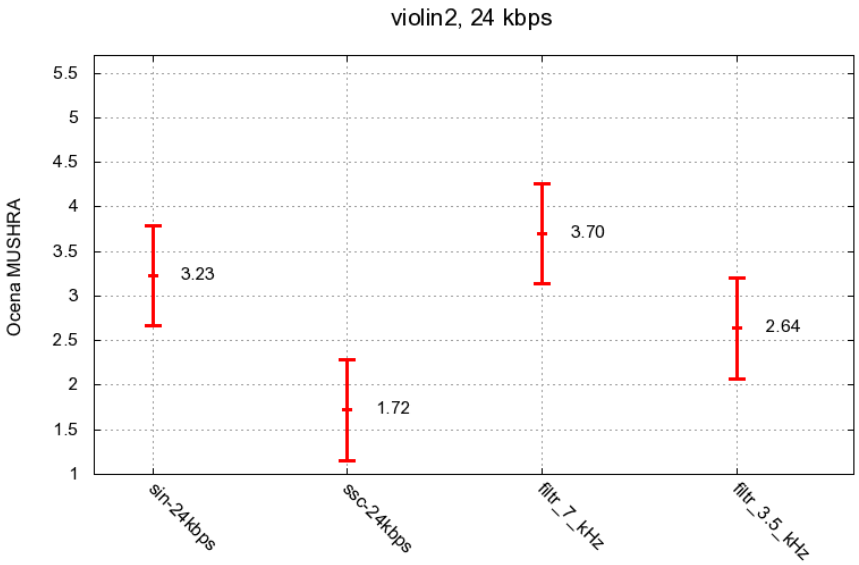
Rysunek C.2: Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 24 kb/s z 95% przedziałem ufności dla sygnału DistGtr.



Rysunek C.3: Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 24 kb/s z 95% przedziałem ufności dla sygnału jazz-short.



Rysunek C.4: Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 24 kb/s z 95% przedziałem ufności dla sygnału mutehorn.

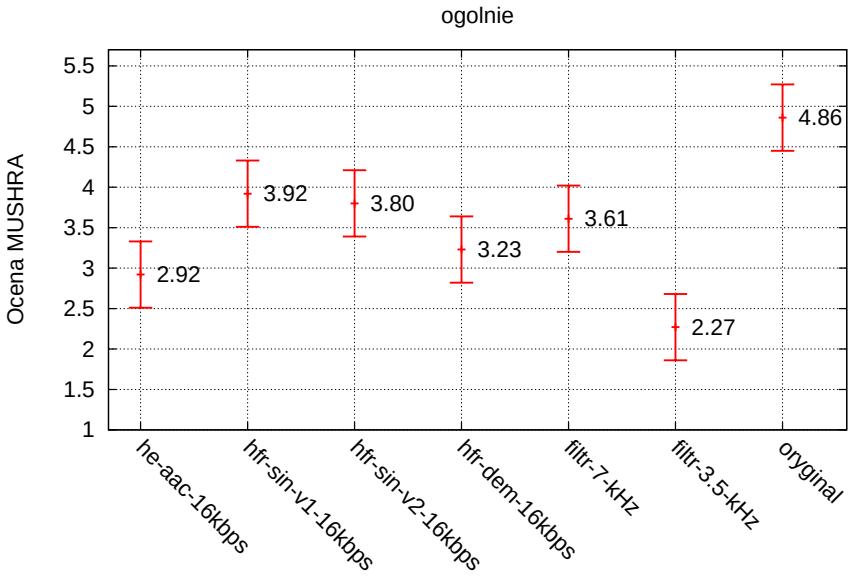


Rysunek C.5: Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 24 kb/s z 95% przedziałem ufności dla sygnału violin2.

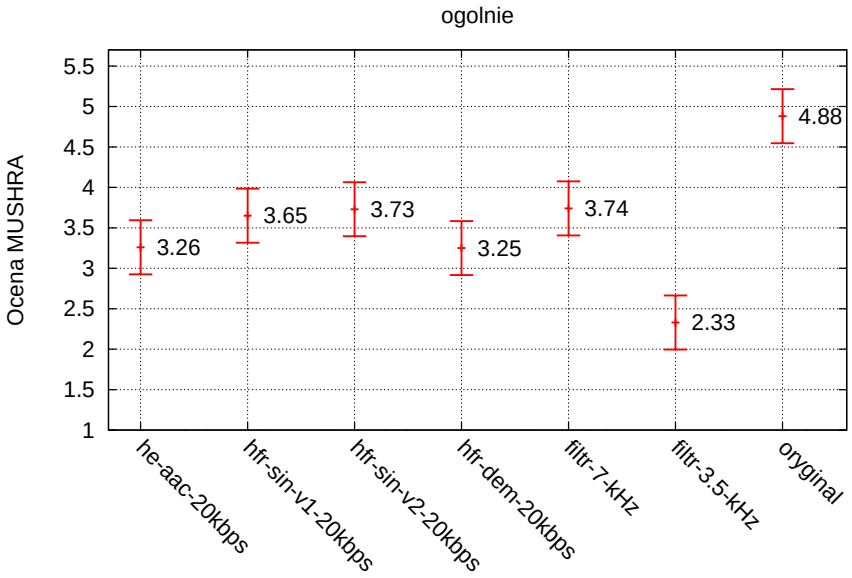
Dodatek D

Testy odsłuchowe dla proponowanych technik rozszerzania widma

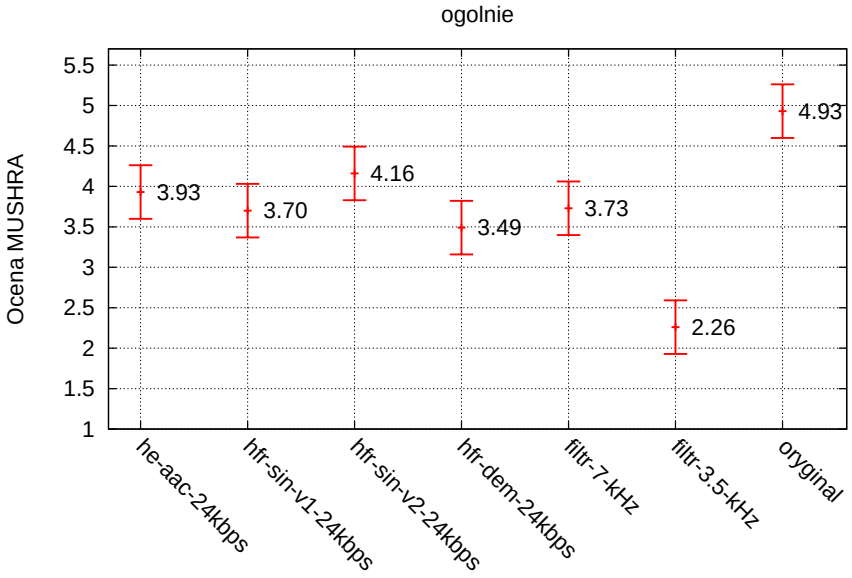
D.1. Porównanie kodeków wykorzystujących techniki rozszerzania widma dla sygnałów muzycznych



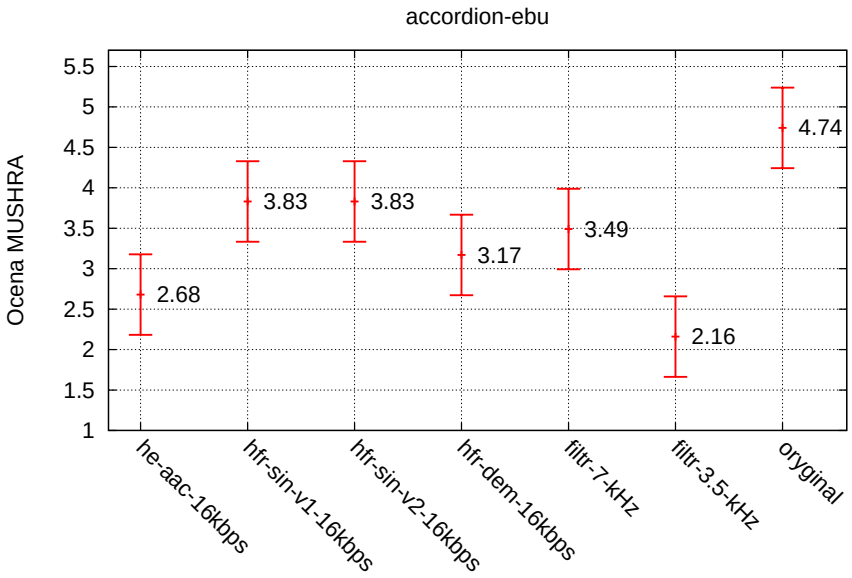
Rysunek D.1: Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 16 kb/s z 95% przedziałem ufności.



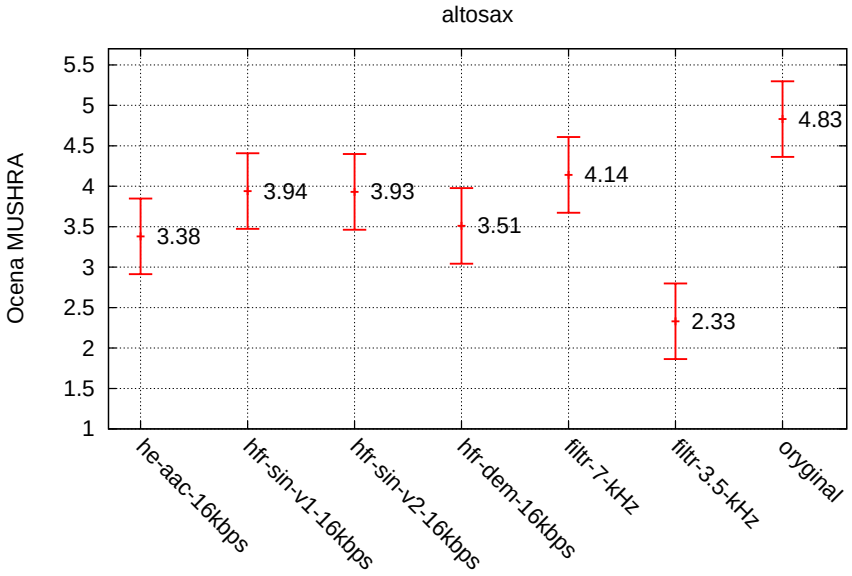
Rysunek D.2: Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 20 kb/s z 95% przedziałem ufności.



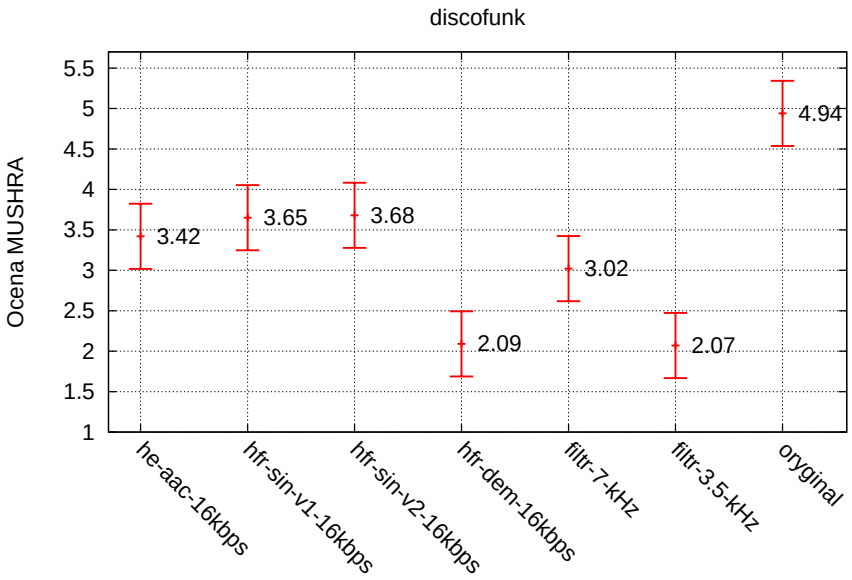
Rysunek D.3: Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 24 kb/s z 95% przedziałem ufności.



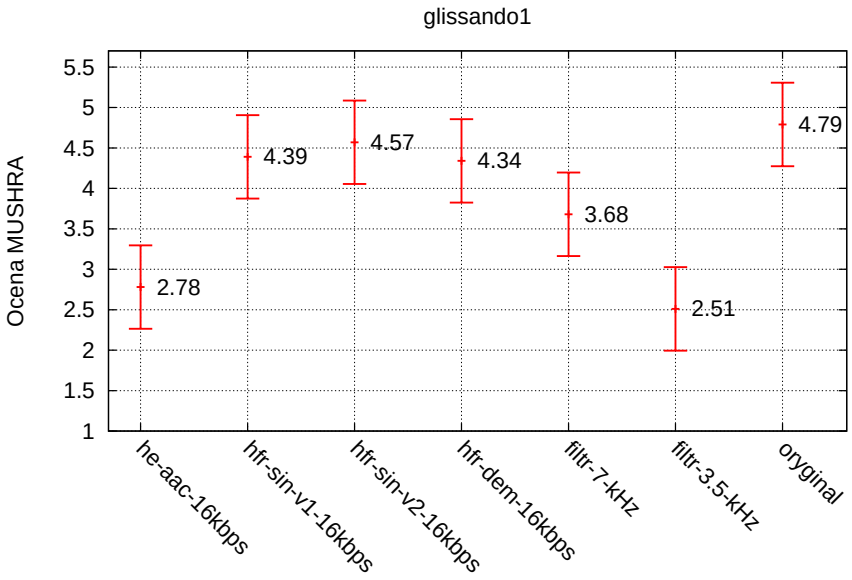
Rysunek D.4: Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 16 kb/s z 95% przedziałem ufności dla sygnału accordion-ebu.



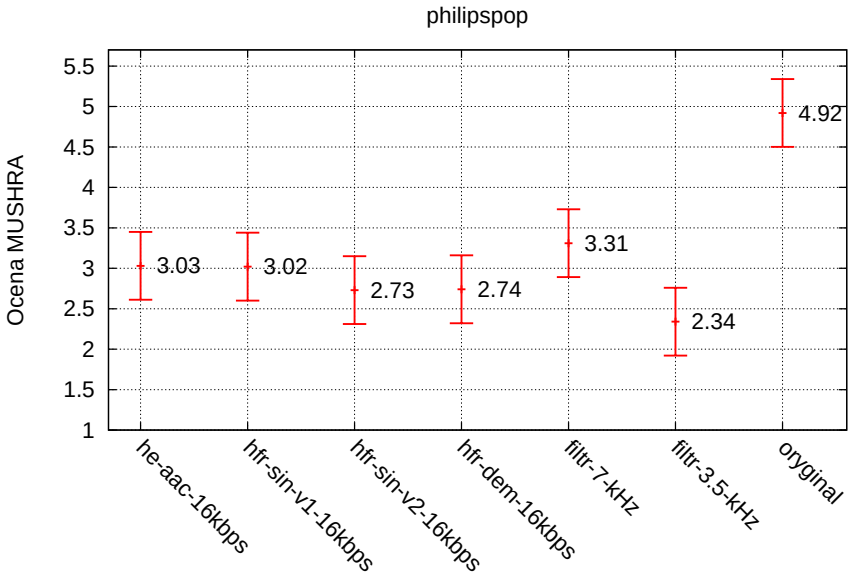
Rysunek D.5: Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 16 kb/s z 95% przedziałem ufności dla sygnału altosax.



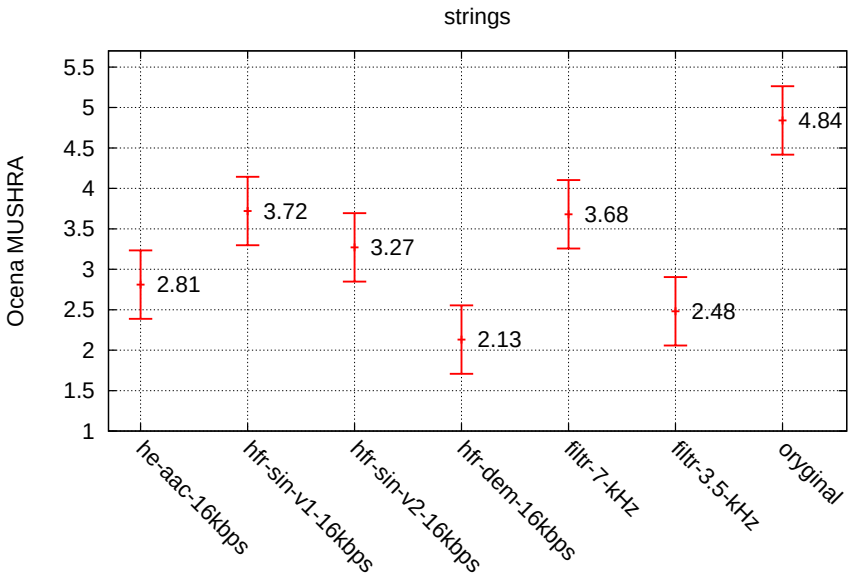
Rysunek D.6: Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 16 kb/s z 95% przedziałem ufności dla sygnału discofunk.



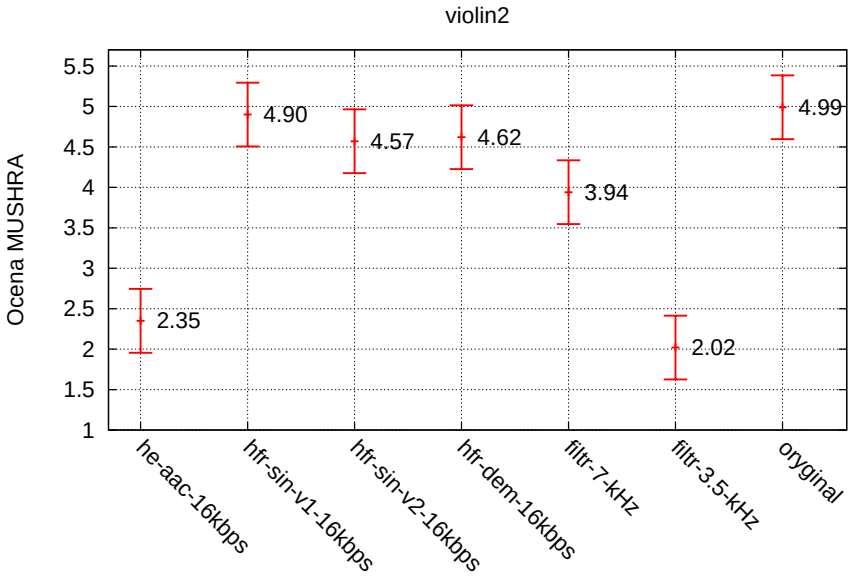
Rysunek D.7: Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 16 kb/s z 95% przedziałem ufności dla sygnału glissando1.



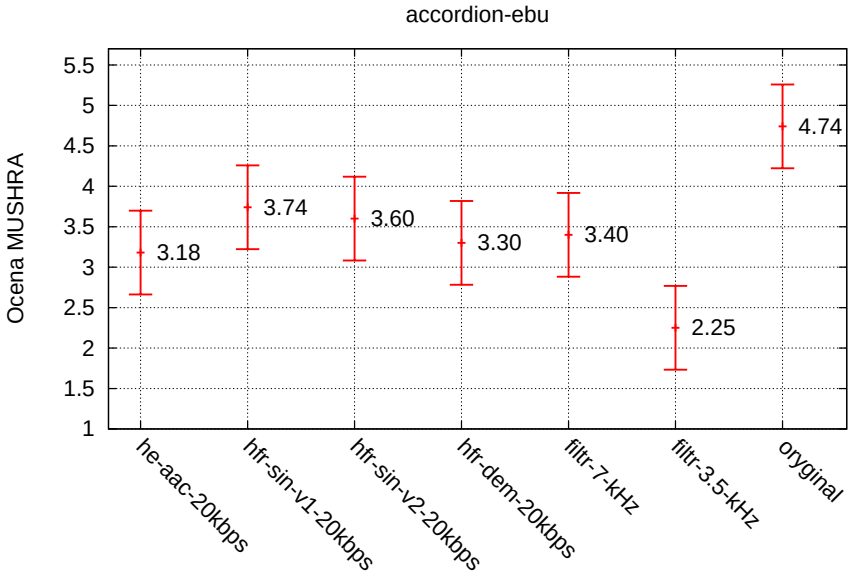
Rysunek D.8: Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 16 kb/s z 95% przedziałem ufności dla sygnału philipspop.



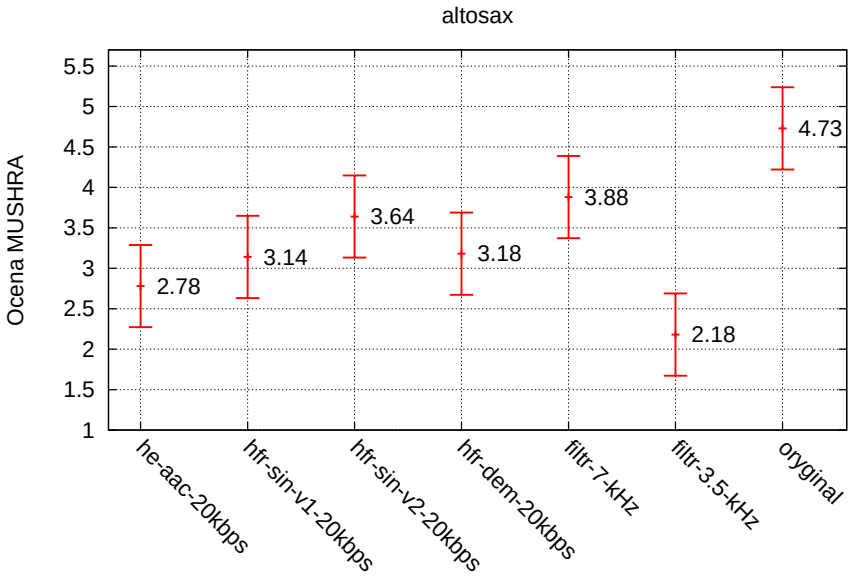
Rysunek D.9: Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 16 kb/s z 95% przedziałem ufności dla sygnału strings.



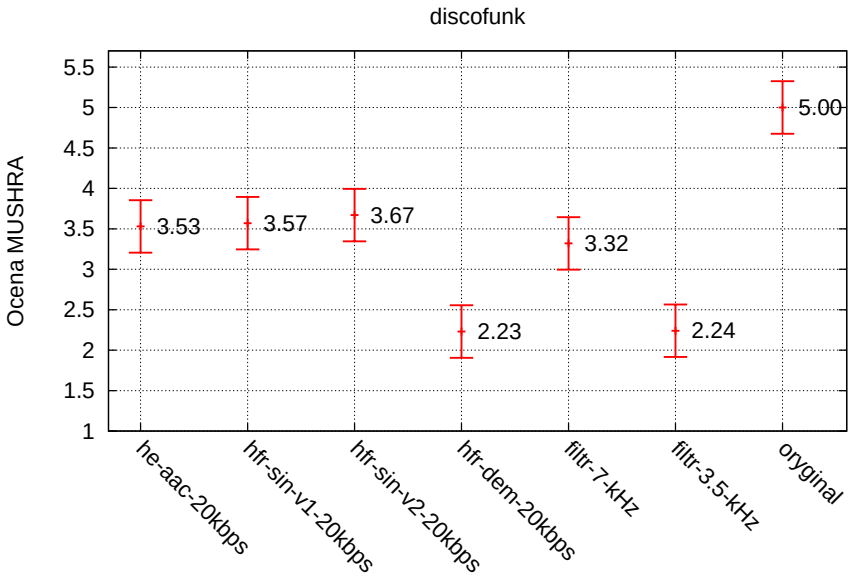
Rysunek D.10: Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 16 kb/s z 95% przedziałem ufności dla sygnału violin2.



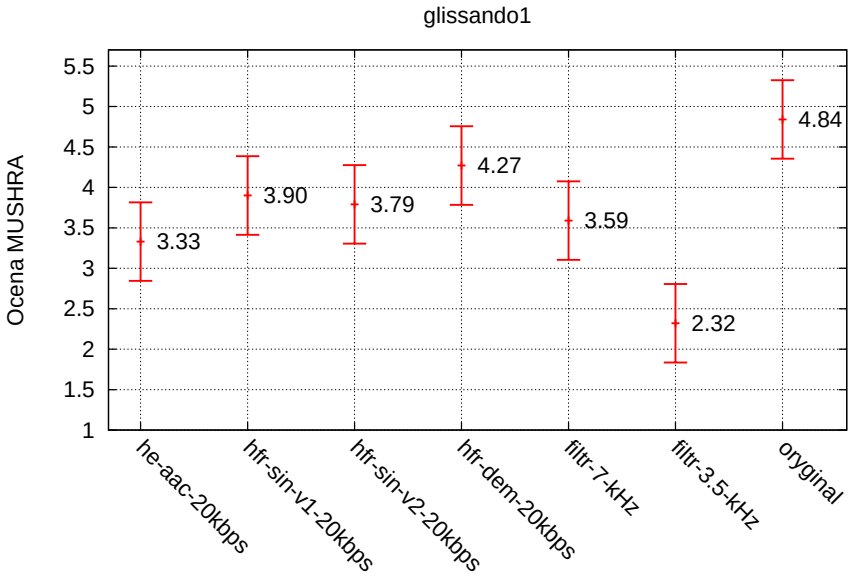
Rysunek D.11: Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 20 kb/s z 95% przedziałem ufności dla sygnału accordion-ebu.



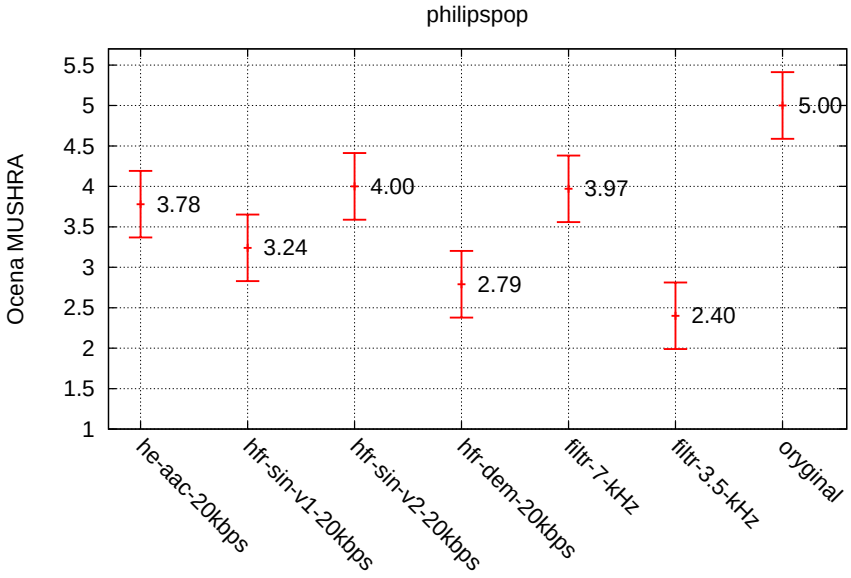
Rysunek D.12: Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 20 kb/s z 95% przedziałem ufności dla sygnału altosax.



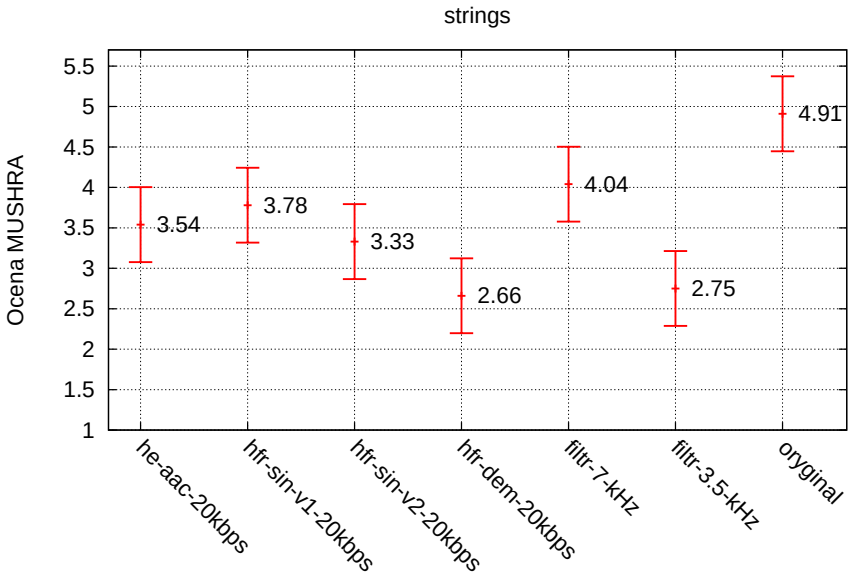
Rysunek D.13: Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 20 kb/s z 95% przedziałem ufności dla sygnału discofunk.



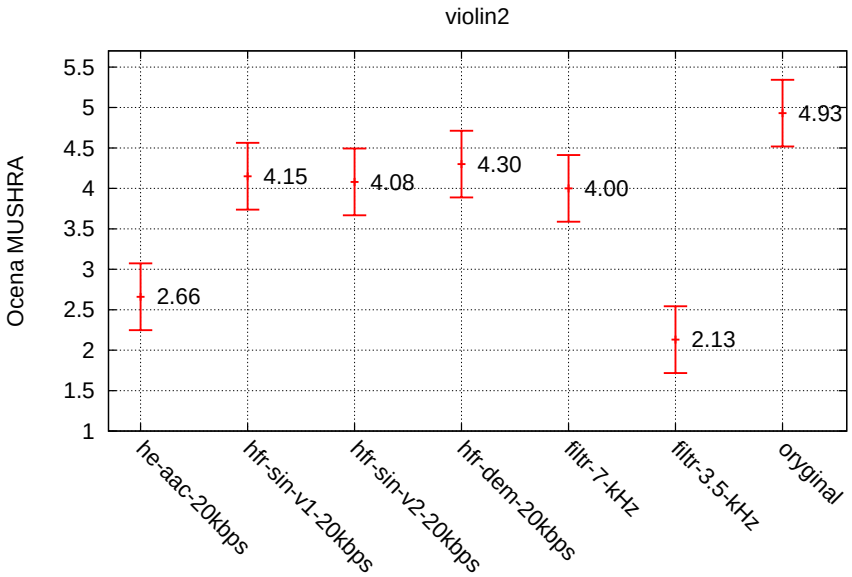
Rysunek D.14: Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 20 kb/s z 95% przedziałem ufności dla sygnału glissando1.



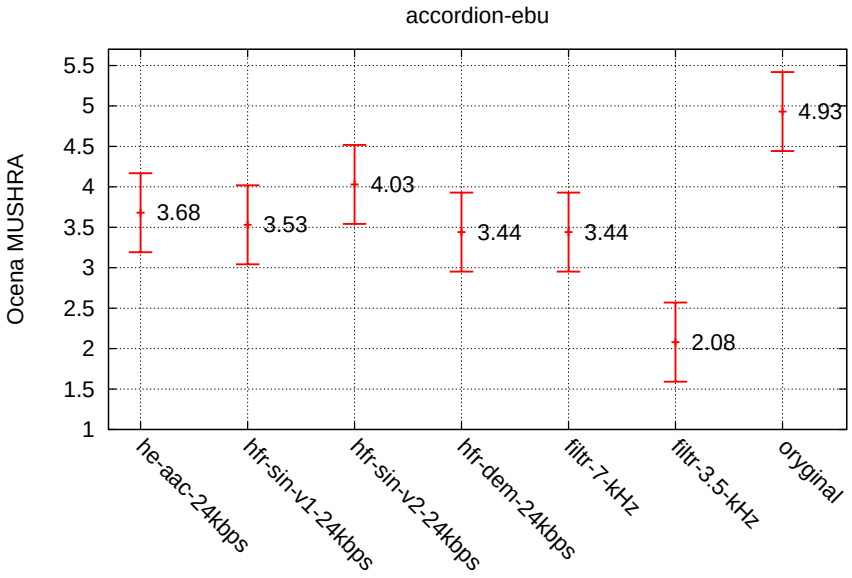
Rysunek D.15: Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 20 kb/s z 95% przedziałem ufności dla sygnału philipspop.



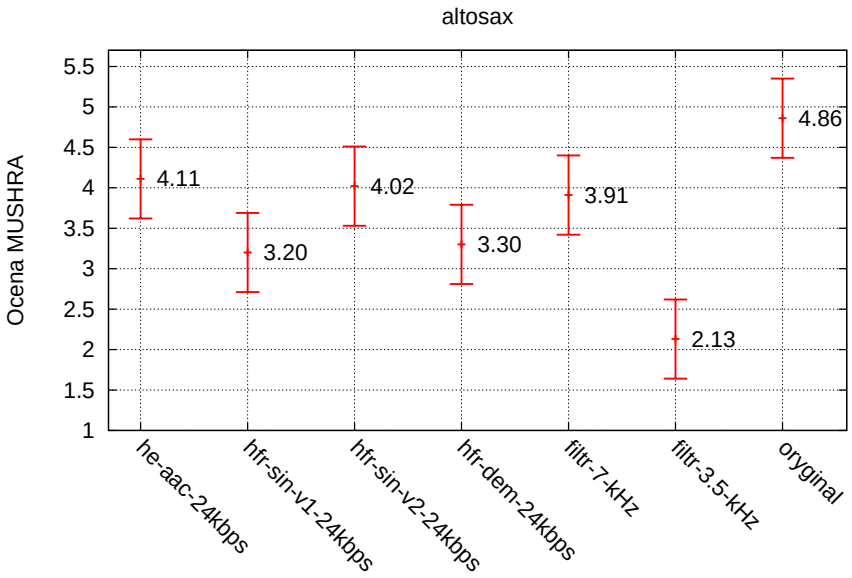
Rysunek D.16: Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 20 kb/s z 95% przedziałem ufności dla sygnału strings.



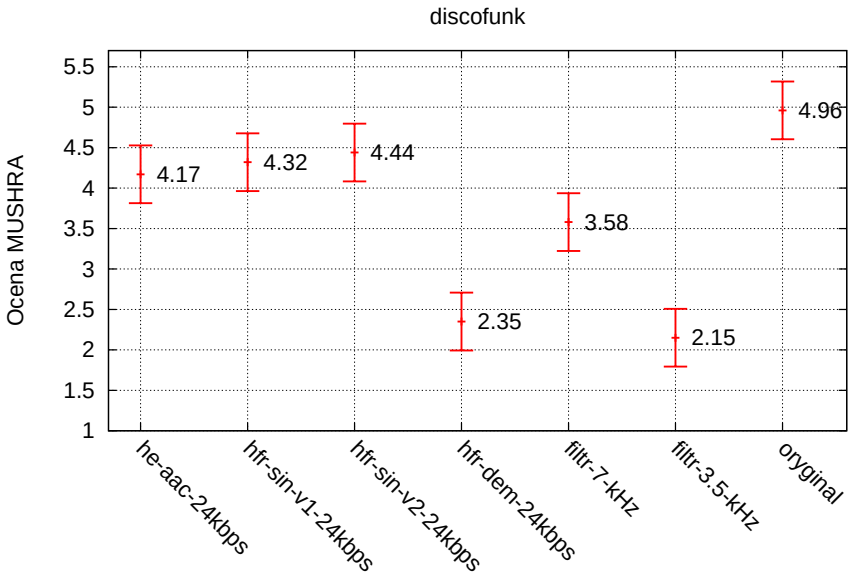
Rysunek D.17: Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 20 kb/s z 95% przedziałem ufności dla sygnału violin2.



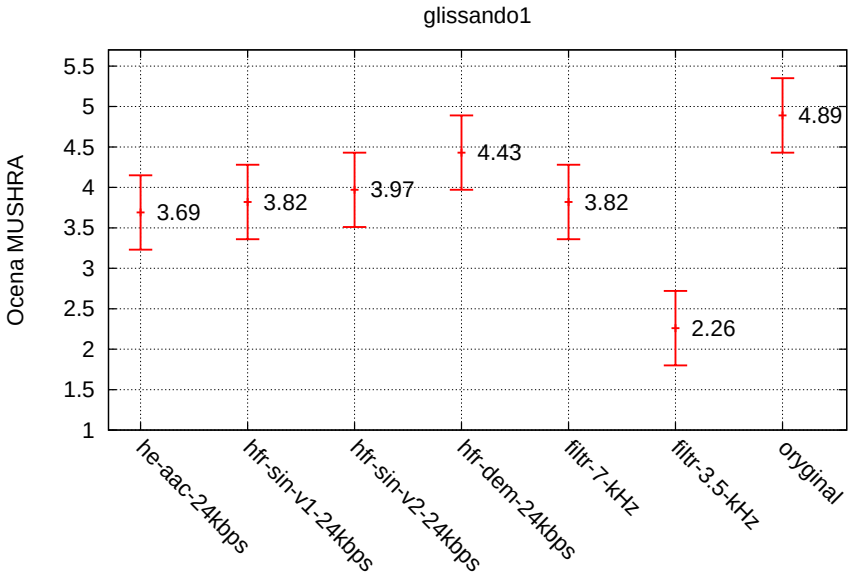
Rysunek D.18: Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 24 kb/s z 95% przedziałem ufności dla sygnału accordion-ebu.



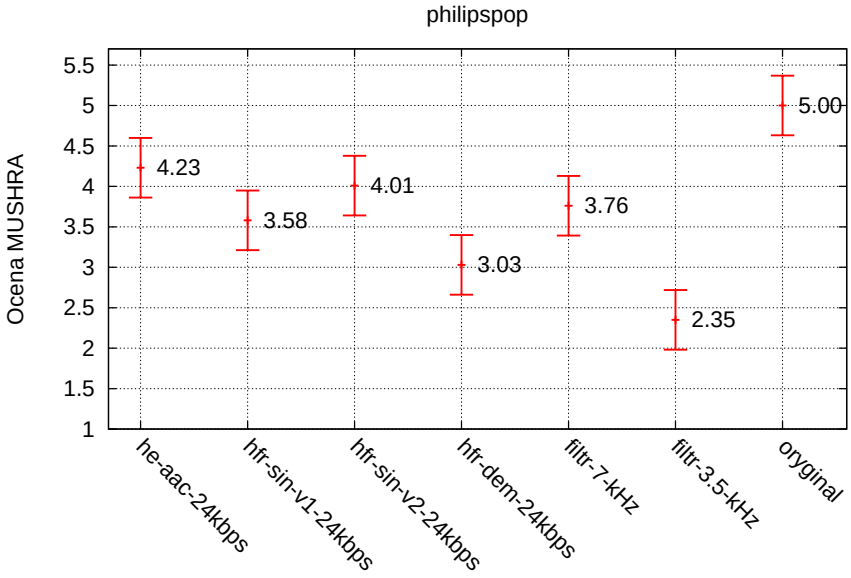
Rysunek D.19: Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 24 kb/s z 95% przedziałem ufności dla sygnału altosax.



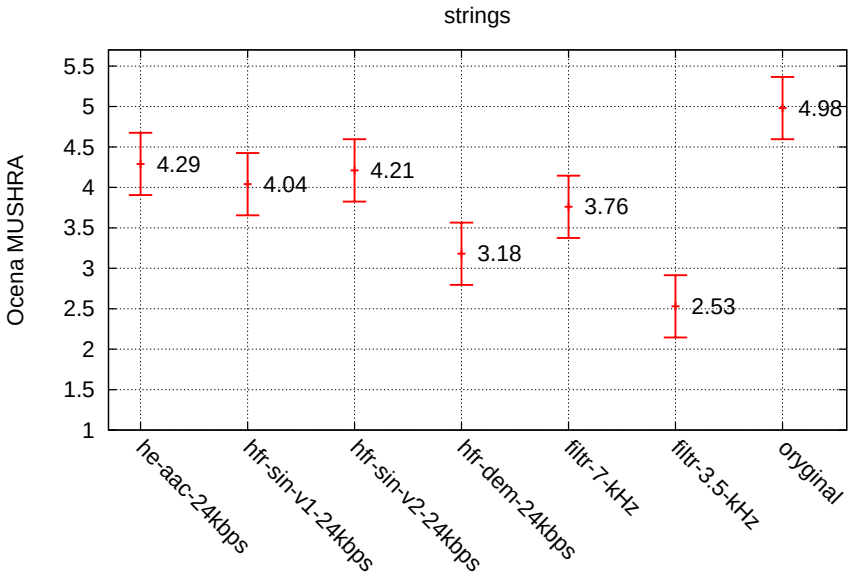
Rysunek D.20: Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 24 kb/s z 95% przedziałem ufności dla sygnału discofunk.



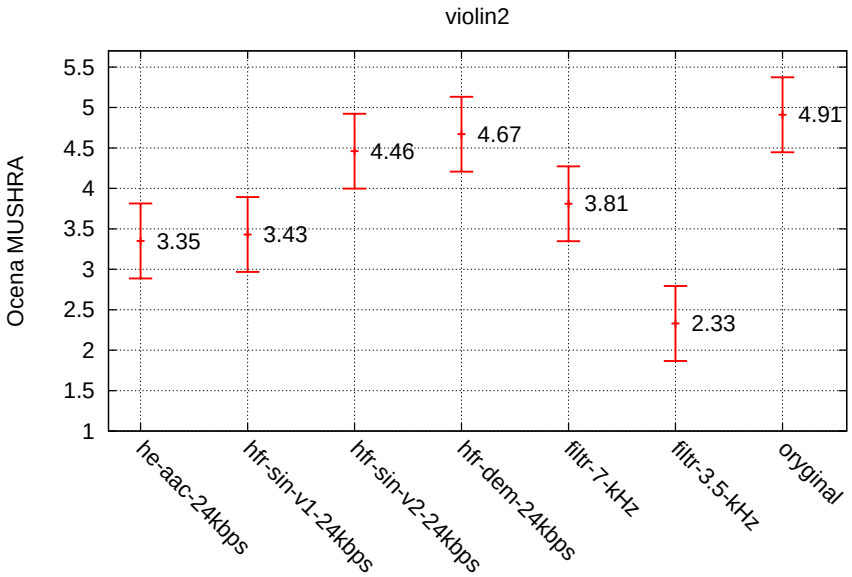
Rysunek D.21: Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 24 kb/s z 95% przedziałem ufności dla sygnału glissando1.



Rysunek D.22: Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 24 kb/s z 95% przedziałem ufności dla sygnału philipspop.

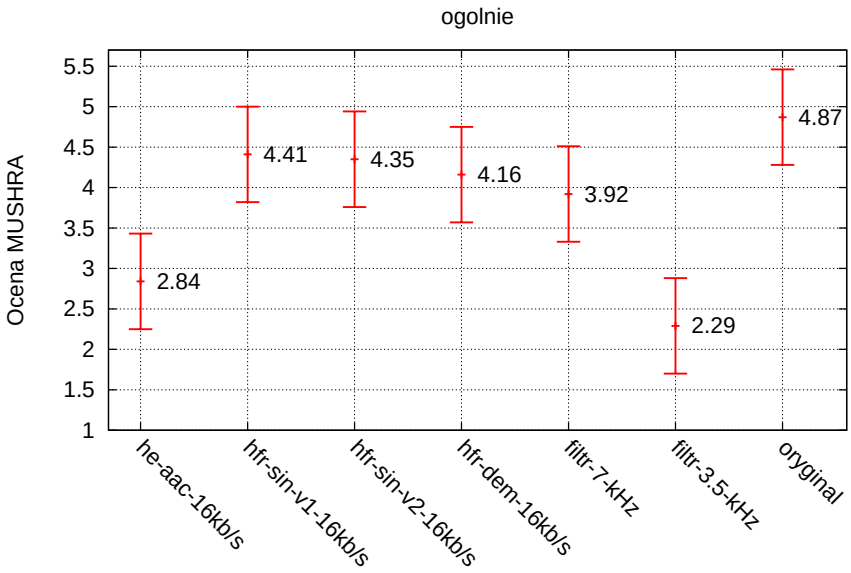


Rysunek D.23: Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 24 kb/s z 95% przedziałem ufności dla sygnału strings.

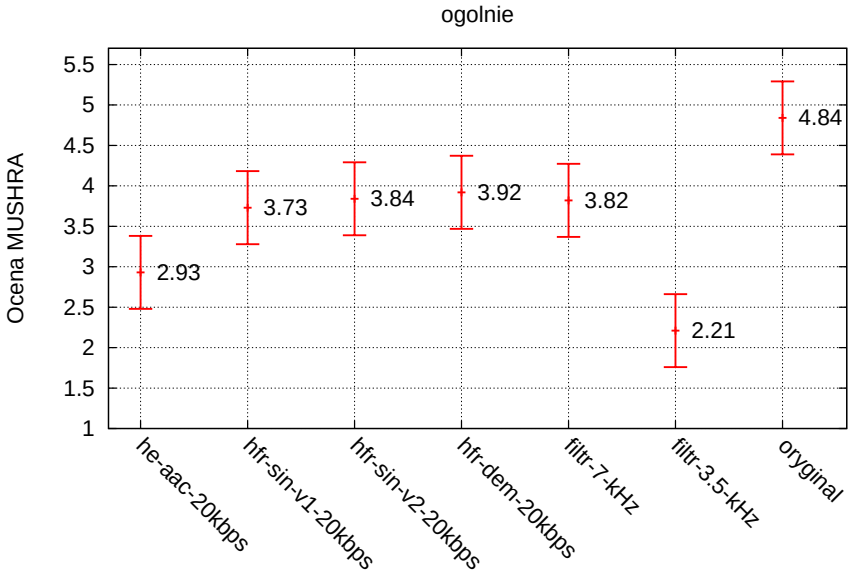


Rysunek D.24: Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 24 kb/s z 95% przedziałem ufności dla sygnału violin2.

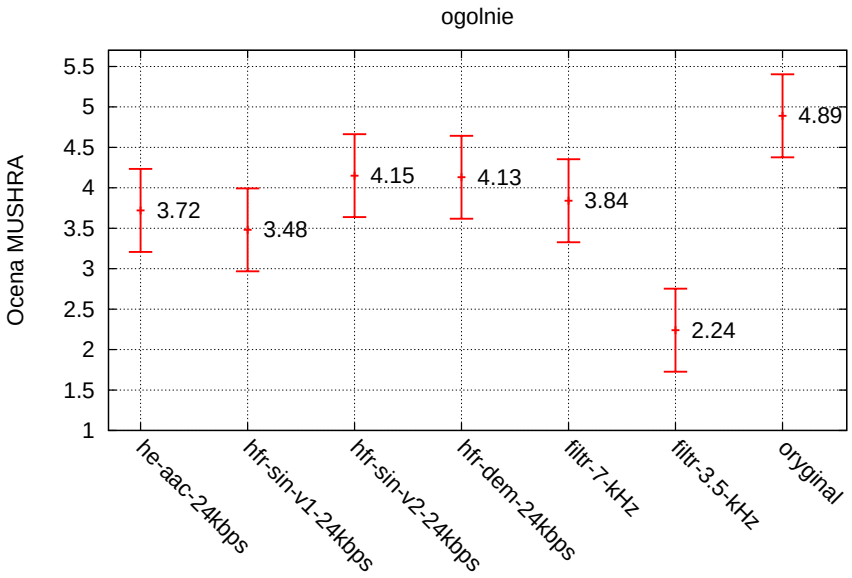
D.2. Porównanie kodeków wykorzystujących techniki rozszerzania widma dla sygnałów zawierających silne składowe tonalne



Rysunek D.25: Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 16 kb/s z 95% przedziałem ufności.



Rysunek D.26: Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 20 kb/s z 95% przedziałem ufności.



Rysunek D.27: Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 24 kb/s z 95% przedziałem ufności.

D.3. Sygnały foniczne wykorzystane w badaniach eksperymentalnych - płyta CD

Sygnały wykorzystane w badaniach eksperymentalnych znajdują się na dołączonej do rozprawy płycie CD.

Spis rysunków

1.1. Ogólny schemat kodera perceptualnego. 6
1.2. Porównanie subiektywnej oceny zniekształcenia wprowadzanego przez kodery kształtu fali oraz parametryczne w funkcji prędkości bitowej (na podstawie pracy [119]). 7
1.3. Schemat kompresji z rozszerzaniem widma. W standardzie MPEG-4 AAC HE składowe LF kodowane są za pomocą techniki AAC (koder podstawowy), natomiast składowe HF - z wykorzystaniem techniki SBR. 9
1.4. Interfejs graficzny aplikacji wykorzystanej podczas testów subiektywnych MUSHRA. 16

2.1. Wynik analizy sinusoidalnej dla widma amplitudowego pojedynczej ramki sygnału. 23
2.2. Śledzenie trajektorii sinusoidalnych. 24
2.3. Schemat modelowania sinusoidalnego 24
2.4. Model hybrydowy sygnału. Spektrogram pokazujący detekcję poszczególnych elementów modelu. 25
2.5. Widmo amplitudowe powstałe po zastosowaniu dwóch różnych metod rozszerzania widma. 29
2.6. Idea rekonstrukcji składowych dużych częstotliwości w technice SBR. 30
2.7. Ogólny schemat techniki SBR. 32
2.8. Schemat kodera SBR. 33

2.9. Rysunek przedstawia siatkę czasowo-częstotliwościową (białe przerywane linie) nałożoną na spektrogram sygnału. Siatka ilustruje granice segmentów energetycznych. Można zauważyć zmienną rozdzielczość częstotliwościową dla dużych częstotliwości oraz zmienną rozdzielczość czasową (krótkie okna w okolicy transjentów). 35
2.10. Schemat blokowy dekodera SBR. 38
2.11. Powstawanie efektu dudnienia oraz nieharmoniczności podczas procesu kopiowania widmowego. 44
2.12. Przykład rekonstrukcji składowych tonalnych z wykorzystaniem MPEG-4 AAC HE (SBR). 45
2.13. Przykład rekonstrukcji składowych tonalnych z wykorzystaniem MPEG-4 AAC HE (SBR). 45
2.14. Przykład rekonstrukcji składowych tonalnych z wykorzystaniem MPEG-4 AAC HE (SBR). 46
2.15. Rysunek przedstawiający spktrogram funkcji sweep oraz pojedynczą ramkę analizy FFT. 48
2.16. Schemat rekonstrukcji składowych HF wykorzystujących układ nieliniowy. . . 51

3.1. Rozszerzenie techniki SBR o dodatkowy blok modelowania sinusoidalnego. . . . 55
3.2. Ogólny schemat proponowanej techniki. 56
3.3. Schemat zmodyfikowanego kodeka MPEG-4 AAC HE uwzględniający dodatkowe bloki modelowania sinusoidalnego. Dodatkowe bloki kodera i dekodera służą kodowaniu tonalnych składowych HF. 57
3.4. Paraboliczna interpolacja prążków widma – QIFFT. 61
3.5. Rysunek poglądowy przedstawiający zastosowanie tłumienia widma. 66
3.6. Przykład zastosowania tłumienia widma. a) spektrogram sygnału oryginalnego x, b) spektrogram sygnału tłumiącego s - tylko składowe tonalne, c) wynik tłumienia widma y - sygnał ma charakter szumowy. 68

3.7. Ilustracja problemu błędnej rekonstrukcji składowych tonalnych. U góry przedstawione są spektrogramy sygnałów oryginalnych oraz trajektorie sinusoidalne (biała linia). Na dole przedstawione są spektrogramy sygnałów zsyntezowanych na podstawie parametrów modelu sinusoidalnego. a) szybkozmienne składniki tonalne nie zostały poprawnie zidentyfikowane przez model sinusoidalny, b) spektrogramy przedstawiają transjent szumowy - model sinusoidalny zaklasyfikował szum jako składowe tonalne. 69

3.8. Schemat umieszczenia dodatkowego modułu detektora tonalności oraz przełącznika wyboru pracy proponowanej techniki BWE. 70

3.9. Spektrogram sygnału. Białą linią zaznaczona wartość współczynnika tonalności dla danej chwili czasu. Dla potrzeb rysunku wartość współczynnika tonalności została przeskalowana – podzielona przez 2. 71

3.10. Schemat pracy kodera sinusoidalnego bez petli sterownia przepływnością – wariant I. 75

3.11. Schemat pracy kodera sinusoidalnego z pętlą sterowania przepływnością – wariant II. 76

3.12. Wykres krzywej ważenia A w zakresie 1 kHz – 22 kHz, dla częstotliwości próbkowania 44100 Hz. 78

3.13. Zależność prędkości bitowej od długości grupy ramek, objętych kodowaniem tą samą tablicą Huffmana. Jest to porównanie dla wszystkich sygnałów fonicznych biorących udział w badaniu, zgodnie z tabelą 1.1. Tablica kodów entropijnych przesyłana jest raz na każdą grupę ramek, np. dla długości grupy równej 100, tablica jest umieszczana w strumieniu bitowym co 100 ramek. Szczegółowe wyniki znajdują się w dodatku A. 81

3.14. Liczba wystąpień symboli dla skwantowanych wartości dla amplitudy i częstotliwości. Wykresy są w skali logarytmicznej. 83

3.15. Rekonstrukcja sygnału – sygnał wyjściowy jest sumą sygnału $\hat{m}$ odtworzonego w dekoderze MPEG-4 AAC HE oraz sygnału $\hat{s}$ odtworzonego w dekoderze sinusoidalnym. Zrekonstruowany sygnał $\hat{x}$ posiada poprawną strukturę harmoniczną składowych tonalnych oraz szersze pasmo w stosunku do sygnału dekodowanego bezpośrednio w dekoderze MPEG-4 AAC HE. 85

3.16. Spektrogramy przedstawiające poszczególne etapy kodowania sygnału z zastosowaniem proponowanej techniki: a) sygnał wejściowy, b) zsyntezowane składowe tonalne s, c) sygnał y po procesie tłumienia, d) zsyntezowane składowe tonalne po stronie dekodera - sygnał $\hat{s}$, e) zrekonstruowany sygnał - proponowana technika, f) zrekonstruowany sygnał za pomocą MPEG-4 AAC HE. Kodowaniu podlegał sygnał syntetyczny o harmonicznej strukturze. Sygnały zakodowano dla prędkości bitowej równej 16 kb/s. 86

3.17. Spektrogramy przedstawiające poszczególne etapy kodowania sygnału z zastosowaniem proponowanej techniki: a) sygnał wejściowy, b) zsyntezowane składowe tonalne s, c) sygnał y po procesie tłumienia, d) zsyntezowane składowe tonalne po stronie dekodera - sygnał $\hat{s}$, e) zrekonstruowany sygnał - proponowana technika, f) zrekonstruowany sygnał za pomocą MPEG-4 AAC HE. Kodowaniu poddany został fragment nagrania skrzypiec - sygnał *violin2*. Sygnały zakodowano dla prędkości bitowej równej 16 kb/s. 87

3.18. Przykładowe spektrogramy sygnałów testowych wykorzystanych w badaniu. . 91

3.19. Wyniki eksperymentów dla kwantyzacji amplitud i częstotliwości. Objaśnienie: $f_{xx} - a_y$, gdzie xx - przedział kwantyzacji częstotliwości w centach, y - przedział kwantyzacji amplitudy w dB. Na przykład $f_{20} - a_3$ oznacza składowe kodowane były z przedziałem kwantyzacji częstotliwości równym 20 centów i amplitudy równym 3 dB. 93

3.20. Wyniki badań odsłuchowych pozwalające oszacować niezbędną liczbę sinusoid w górnej części pasma. Objaśnienie: sin_{ss} - gdzie ss to średnia liczba sinusoid na ramkę $\overline{N}_{Sin}$. 4kHz i 7kHz to sygnały ograniczone pasmowo odpowiednio do 4 i 7 kHz. 94

4.1. Widmo sygnału sinusoidalnego a) i widmo składowej tonalnej b). c) i d) Spektrogramy sygnałów składających się odpowiednio z sygnałów z rysunków a) i b). 97

4.2. Proces analizy i syntezy sygnału w proponowanej technice BWE: a) spektrogram sygnału ograniczonego pasmowo rekonstruowanego w dekoderze, b) wąskopasmowa składowa tonalna, c) zrekonstruowane składowe tonalne HF, d) zrekonstruowany sygnał stanowiący sumę składowych HF oraz sygnału zdekodowanego w dekoderze podstawowym. 101

4.3. Schemat kodera 103
4.4. Schemat dekodera 103
4.5. Wyznaczanie współczynników skalujących - przykład obwiedni energetycznej dla pojedynczego szeregu harmonicznego w obrębie jednej ramki analizy. 107
4.6. Przykład skalowania energii dla pojedynczego okna analizy. Linia o kolorze niebieskim reprezentuje $G_{LPC_m}(f)$, linia o kolorze czarnym $G_D G_{LPC_m}(f)$, a przerywana linia o kolorze czerwonym oznacza połączone wartości amplitud chwilowych sinusoid wyznaczonych przez analizę sinusoidalną. 108
4.7. Przykład zmiany współczynników LSF w kolejnych oknach analizowanego sygnału. 112
4.8. Wykrywanie częstotliwości podstawowej 114
4.9. Trajektoria wykorzystania w algorytmie skalowania częstotliwości, spełniająca kryterium (4.27). 117
4.10. Przykład rekonstrukcji sygnału violin2 dla prędkości bitowej 16 kb/s - rysunki a), c), e) oraz rekonstrukcji sygnału strings dla prędkości bitowej 20 kb/s - rysunki b), d), f). Spektrogramy przedstawiają: a), b) sygnał oryginalny, c), d) sygnał zakodowany przy pomocy techniki MPEG-4 AAC HE d), e) sygnał zdekodowany przy użyciu proponowanej techniki BWE. Należy zauważyć, że maksymalny zakres rekonstruowanych częstotliwości przez technikę MPEG-4 AAC HE jest ograniczony przez mechanizm sterowania prędkości bitowej. Tego typu efekt nie występuje w proponowanej technice BWE - rekonstruowany jest pełen zakres częstotliwości. 118

5.1. Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 16 kb/s z 95% przedziałem ufności. 125
5.2. Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 20 kb/s z 95% przedziałem ufności. 125
5.3. Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 24 kb/s z 95% przedziałem ufności. 126
5.4. Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 16 kb/s z 95% przedziałem ufności. 127
5.5. Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 20 kb/s z 95% przedziałem ufności. 128

5.6. Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 24 kb/s z 95% przedziałem ufności. 128
5.7. Odległość widmowa dla prędkości bitowej 16 kb/s. 129
5.8. Odległość widmowa dla prędkości bitowej 20 kb/s. 129
5.9. Odległość widmowa dla prędkości bitowej 24 kb/s. 130
5.10. Zależność subiektywnej oceny od prędkości bitowej dla wszystkich badanych sygnałów. „hfr-sin" oznacza technikę opisaną w rozdziale 3 z pętlą sterowania przepływnością. 131
5.11. Zależność subiektywnej oceny od prędkości bitowej dla nagrań zawierających silne składowe tonalne. „hfr-sin" oznacza technikę opisaną w rozdziale 3 z pętlą sterowania przepływnością. 132
5.12. Zależność obiektywnej oceny (odległości widmowej) od prędkości bitowej dla wszystkich badanych sygnałów. 132

A.1. Zależność prędkości bitowej od długości bloku ramek objętych kodowaniem tą samą tablicą Huffmana 152
A.2. Zależność prędkości bitowej od długości bloku ramek objętych kodowaniem tą samą tablicą Huffmana 153
A.3. Zależność prędkości bitowej od długości bloku ramek objętych kodowaniem tą samą tablicą Huffmana 153
A.4. Zależność prędkości bitowej od długości bloku ramek objętych kodowaniem tą samą tablicą Huffmana 154
A.5. Zależność prędkości bitowej od długości bloku ramek objętych kodowaniem tą samą tablicą Huffmana 154
A.6. Zależność prędkości bitowej od długości bloku ramek objętych kodowaniem tą samą tablicą Huffmana 155
A.7. Zależność prędkości bitowej od długości bloku ramek objętych kodowaniem tą samą tablicą Huffmana. Porównanie dla wszystkich sygnałów fonicznych biorących udział w badaniu. 155
A.8. Liczba wystąpień symboli dla skwantowanych wartości amplitudy. 156
A.9. Liczba wystąpień symboli dla skwantowanych wartości częstotliwości. 156

B.1. Spektrogram sygnału z zaznaczonymi białą linią trajektoriami sinusoidalnymi. 158

B.2. Synteza sinusoid na podstawie wyznaczonych trajektorii sinusoidalnych. 159
B.3. Kodowanie trajektorii modelu sinusoidalnego. Ciągła linia pokazuje przebieg trajektorii sinusoidalnych. Punkty przedstawiają kodowane parametry, wyznaczone w kolejnych ramkach. 160
B.4. Zastosowany model kwantyzacji. 161

C.1. Porównanie modelowania sinusoidalnego. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 24 kb/s z 95% przedziałem ufności. 163
C.2. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 24 kb/s z 95% przedziałem ufności dla sygnału DistGtr. 163
C.3. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 24 kb/s z 95% przedziałem ufności dla sygnału jazz-short. 164
C.4. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 24 kb/s z 95% przedziałem ufności dla sygnału mutehorn. 164
C.5. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 24 kb/s z 95% przedziałem ufności dla sygnału violin2. 165

D.1. Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 16 kb/s z 95% przedziałem ufności. 166
D.2. Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 20 kb/s z 95% przedziałem ufności. 167
D.3. Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 24 kb/s z 95% przedziałem ufności. 167
D.4. Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 16 kb/s z 95% przedziałem ufności dla sygnału accordion-ebu. 168
D.5. Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 16 kb/s z 95% przedziałem ufności dla sygnału altosax. 168
D.6. Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 16 kb/s z 95% przedziałem ufności dla sygnału discofunk. 169

D.7. Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 16 kb/s z 95% przedziałem ufności dla sygnału glissando1. 169
D.8. Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 16 kb/s z 95% przedziałem ufności dla sygnału philipspop. 170
D.9. Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 16 kb/s z 95% przedziałem ufności dla sygnału strings. 170
D.10. Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 16 kb/s z 95% przedziałem ufności dla sygnału violin2. 171
D.11. Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 20 kb/s z 95% przedziałem ufności dla sygnału accordion-ebu. 171
D.12. Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 20 kb/s z 95% przedziałem ufności dla sygnału altosax. 172
D.13. Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 20 kb/s z 95% przedziałem ufności dla sygnału discofunk. 172
D.14. Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 20 kb/s z 95% przedziałem ufności dla sygnału glissando1. 173
D.15. Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 20 kb/s z 95% przedziałem ufności dla sygnału philipspop. 173
D.16. Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 20 kb/s z 95% przedziałem ufności dla sygnału strings. 174

D.17. Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 20 kb/s z 95% przedziałem ufności dla sygnału violin2. 174
D.18. Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 24 kb/s z 95% przedziałem ufności dla sygnału accordion-ebu. 175
D.19. Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 24 kb/s z 95% przedziałem ufności dla sygnału altosax. 176
D.20. Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 24 kb/s z 95% przedziałem ufności dla sygnału discofunk. 176
D.21. Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 24 kb/s z 95% przedziałem ufności dla sygnału glissando1. 177
D.22. Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 24 kb/s z 95% przedziałem ufności dla sygnału philipspop. 177
D.23. Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 24 kb/s z 95% przedziałem ufności dla sygnału strings. 178
D.24. Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 24 kb/s z 95% przedziałem ufności dla sygnału violin2. 178
D.25. Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 16 kb/s z 95% przedziałem ufności. 179
D.26. Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 20 kb/s z 95% przedziałem ufności. 179
D.27. Porównanie technik rozszerzania widma. Wyniki testów odsłuchowych MUSHRA dla prędkości transmisji 24 kb/s z 95% przedziałem ufności. 180

Spis tabel

1.1. Sygnały testowe. 14
1.2. Subiektywna skala ocen – MUSHRA. 16

2.1. Charakterystyczne częstotliwości techniki SBR dla kodeka MPEG-4 AAC HE – częstotliwość podziału f_{SBR} i górna granica częstotliwości $f_{SBR_{max}}$ dla f_s = 44100 Hz. Wartości pochodzą z implementacji opracowanej przez konsorcjum 3GPP[2] oraz firmę Coding Technologies [116]. Przedstawione prędkości bitowe odpowiadają zakresowi badań ujętych w rozprawie. 34

4.1. Parametry kwantyzacji ustalone dla proponowanej techniki BWE. Skwantowane wartości podlegają kodowaniu ADPCM. 119
4.2. Prędkości bitowe strumieni opisujących poszczególne parametry proponowanego kodeka. 119