Maciej Bartkowiak
Tomasz Żernicki
Politechnika Poznańska
Katedra Telekomunikacji Multimedialnej i Mikroelektroniki
ul. Polanka 3, 60-96, Poznań
mbartkow@multimedia.edu.pl

Wrocław, 09-11 kwietnia 2008

HYBRYDOWY PARAMETRYCZNO-TRANSFORMATOWY KODEK DŹWIĘKU

Streszczenie**: Referat opisuje nową technikę kompresji dźwięku szerokopasmowego do zastosowań wymagających bardzo małych prędkości transmisji. Zaproponowana technika jest hybrydowym połączeniem kodowania transformatowego i kodowania parametrycznego (sinusoidalno- szumowego). Istotą podejścia jest wąskopasmowe transformatowe kodowanie indywidualnych składowych tonalnych sygnału wzdłuż trajektorii częstotliwości tych składowych, wstępnie estymowanych za pomocą modelu sinusoidalnego.**

1. WSTĘP

Postęp w dziedzinie kompresji sygnałów fonicznych obejmuje coraz bardziej efektywne metody kodowania pozwalające na redukcję prędkości transmisji cyfrowego dźwięku szerokopasmowego (np. muzyki) do wartości do niedawna wykorzystywanych dla przesyłania mowy (np. 24kb/s), przy jednoczesnym zachowaniu dobrej jakości dźwięku. W tym zakresie przepływności prym wiodą techniki kodowania transformatowego, w szczególności MPEG-4 HE-AAC [1-3]. Warto zauważyć, że poważna redukcja prędkości transmisji w kodeku AAC (z 64kb/s do 24kb/s) została osiągnięta głównie dzięki uzupełnieniu tradycyjnego schematu wykorzystującego przekształcenie MDCT o mechanizmy kodowania parametrycznego: percepcyjnego zastąpienia szumu (PNS), poszerzania widma (SBR) i stereofonii parametrycznej (PS). Współczesny kodek MPEG-4 HE-AAC jest zatem kodekiem hybrydowym.

Kodowanie parametryczne stanowi odejście od zasady przybliżania oryginalnego przebiegu sygnału (gdzie stopień stratności można wyrazić miarą błędu, np. średniokwadratowego) na rzecz możliwie wiernego odwzorowania cech psychoakustycznych dźwięku [3-5]. Transmitowany program jest reprezentowany za pomocą zestawu parametrów modelu, pozwalającego zsyntezować równoważny sygnał w dekoderze. Działający według tej zasady standardowy kodek MPEG-4 SSC [4] pozwala na osiągnięcie bardzo efektywnej kompresji, gdzie dobra jakość sygnału zachowana jest przy prędkości transmisji około 24kb/s. Z drugiej strony bardzo trudne jest utrzymanie akceptowalnej jakości dźwięku przy dalszym zmniejszaniu przepływności, a także uzyskanie bardzo dobrej jakości przy zwiększaniu prędkości transmisji.

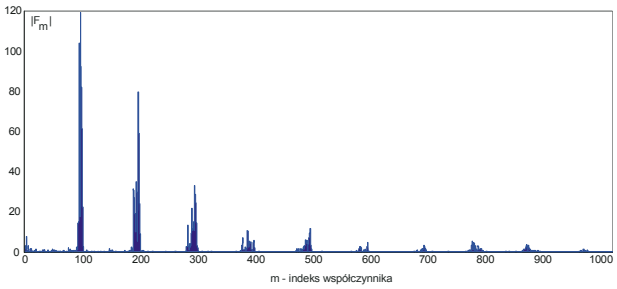
Proponowana nowa technika kompresji stanowi inną niż MPEG-4 HE AAC realizację idei połączenia kodowania transformatowego i parametrycznego. Głównym celem jest ominięcie wad obu technik standardowych i zachowanie dobrej jakości dźwięku dla bardzo małych prędkości transmisji. Nową technikę można interpretować jako kodowanie subpasmowe z adaptacyjną konfiguracją pasm, podążającą za zmianami częstotliwości składowych harmonicznych, lub jako zmodyfikowane kodowanie sinusoidalne, gdzie parametry trajektorii są zakodowane z wykorzystaniem MDCT.

2. WADY I OGRANICZENIA TECHNIK KOMPRESJI

2.1. Koder transformatowy AAC

Tradycyjne kodowanie transformatowe reprezentuje widmo chwilowe kolejnych bloków próbek sygnału za pomocą współczynników transformaty kosinusowej, które podlegają kwantowaniu zgodnie z liczbą bitów przydzieloną przez algorytm alokacji sterowany modelem psychoakustycznym [1]. Sterowanie kwantyzacją odbywa się w tzw. pasmach skalowania, zatem w przypadku wystąpienia zarówno pojedynczej składowej harmonicznej jak i grupy prążków zniekształcenie ma charakter pasmowy.

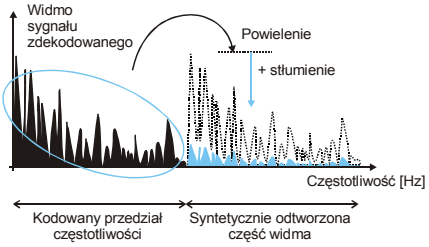
W przypadku dźwięków o niestabilnej częstotliwości podstawowej (zjawisko często obserwowane w mowie, śpiewie i instrumentach o swobodnej intonacji) składowe tonalne są reprezentowane nie przez pojedyncze współczynniki transformaty, lecz całe grupy takich współczynników (rys. 1), a ich kompresja jest mało efektywna. Co więcej, ze względu na konieczność sygnalizacji, kodowanie entropijne wartości współczynników wybiera tablicę kodu Huffmana, która pozostaje wspólna dla wszystkich współczynników w paśmie skalowania lub grupie takich pasm. W konsekwencji duża dynamika wartości współczynników w ramach jednego pasma również obniża efektywność kompresji.



Rys. 1. Przykładowy rozkład współczynników MDCT dla sygnału o silnej składowej tonalnej

Konstrukcja modelu psychoakustycznego w kodeku AAC faworyzuje składowe niskoczęstotliwościowe, co przy silnej kwantyzacji objawia się dramatycznym pogorszeniem jakości górnej części widma sygnału.

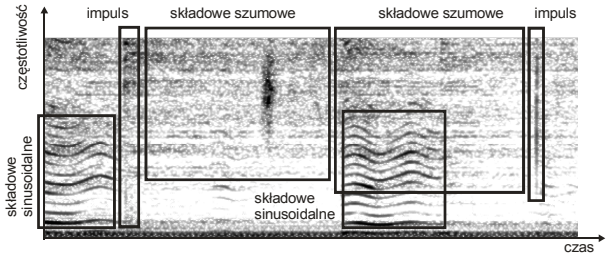
Mechanizm SBR stosowany dla małych prędkości transmisji pozwala wprawdzie zakodować sygnał o mocno ograniczonej szerokości pasma i zrekonstruować brakujące składowe wyższych częstotliwości przez replikację dolnej części widma (rys. 2), jednak ta rekonstrukcja wymaga dobrej jakości sygnału odniesienia (warunek trudny do spełnienia przy prędkości transmisji rzędu 16kb/s, gdy wymuszona granica podziału wynosi około 3kHz). Ponadto, użyta w SBR metoda generacji brakującej części widma przez proste kopiowanie z dolnego pasma uniemożliwia prawidłowe odtworzenie proporcjonalnych zmian częstotliwości składowych sinusoidalnych i ich harmonicznych relacji (rys. 7)



Rys. 2. Zasada działania mechanizmu SBR

2.2. Koder parametryczny SSC

W kompresji parametrycznej w celu uzyskania dużej wierności konieczne okazuje się wykorzystanie zestawu dedykowanych modeli dla różnych klas składowych dźwięku: o charakterze tonalnym, transientowym oraz stochastycznym (rys. 3). Dlatego koder MPEG-4 SSC generuje trzy strumienie bitowe, zawierające parametry wykrytych w sygnale sinusoid, impulsów oraz szumu [3,6].



Rys. 3. Fragment spektrogramu sygnału fonicznego z wyróżnionymi klasami składowych poddawanych modelowaniu

Głównym (generującym około 80% całkowitego strumienia bitowego i mającym największe znaczenie psychoakustyczne) elementem kodera SSC jest model deterministyczny, który przybliża tonalne składowe dźwięku sumą sinusoid o zmiennych w czasie amplitudach i częstotliwościach chwilowych (1).

$$x_{\mathrm{det}}(t) = \mathrm{Re}\left\{\sum_{k=1}^{K} A_k(t)\exp\left(j\varphi_k + j2\pi\int_{-\infty}^{t} f_k(\tau)\,d\tau\right)\right\} \quad (1)$$

Podstawowym założeniem wynikającym z obserwacji spektrogramów rzeczywistych sygnałów jest wolnozmienny przebieg funkcji $A_k(t)$ oraz $f_k(t)$. Przebiegi te są aproksymowane w dekoderze odcinkami liniowymi lub wielomianami niskiego rzędu. Jakość zrekonstruowanego sygnału w dużej mierze zależy od dokładności tej aproksymacji oraz od prawidłowej detekcji składników (1) na podstawie widma chwilowego sygnału oryginalnego. Zasadnicze znaczenie ma również dokładność estymacji [7] i poprawne śledzenie zmian częstotliwości chwilowej i amplitudy, pozwalające na skonstruowanie prawidłowych *trajektorii* $A_k(t)$ i $f_k(t)$ [8,9].

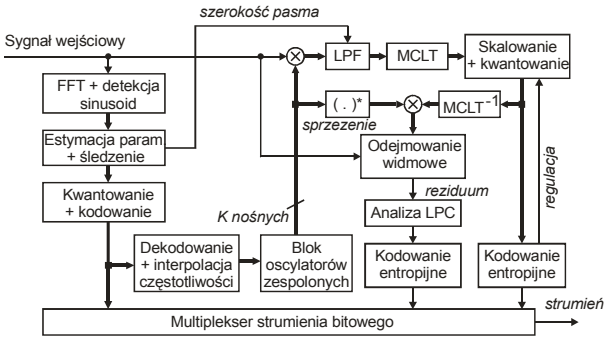
Pozostałe składowe sygnału modelowane są przez analizę reziduum uzyskanego w wyniku odjęcia zrekonstruowanej części deterministycznej (1) od sygnału oryginalnego. Obecne w reziduum składowe impulsowe kodowane są przez parametryczny opis ich kształtu w dziedzinie czasu, a składniki stochastyczne – poprzez współczynniki modelu typu AR aproksymującego obwiednię widmową.

Podstawowy problem ograniczający efektywność kodowania techniką SSC to stosunkowo duży wpływ kwantyzacji amplitud i częstotliwości na degradację zdekodowanego sygnału. Dlatego też kwantowanie w standardowym koderze MPEG-4 SSC nie jest wykorzystywane do regulacji przepływności a raczej jest ustalone w ten sposób, że błąd kwantyzacji amplitudy i częstotliwości nie przekracza odpowiednio ±0,18dB i ±0,5%. Sterownie prędkością transmisji odbywa się poprzez rezygnację z przesyłania mniej istotnych trajektorii sinusoidalnych, przy czym o kolejności wyboru decyduje model psychoakustyczny.

3. NOWA METODA KOMPRESJI

3.1. Schemat kodera

Schemat blokowy proponowanego kodera przedstawiono na rys. 4. W koderze tym indywidualne składowe harmoniczne dźwięku są reprezentowane przez trajektorię częstotliwości oraz sygnał wąskopasmowej obwiedni zespolonej kodowanej transformatowo w paśmie podstawowym. Zaletą takiego podejścia jest bardziej realistyczny opis natury składowych tonalnych, które w rzeczywistym sygnale podlegają drobnym fluktuacjom. Istotnym składnikiem obwiedni jest składowa wolnozmienna oscylacyjna wynikająca z błędów estymacji oraz kwantowania częstotliwości.

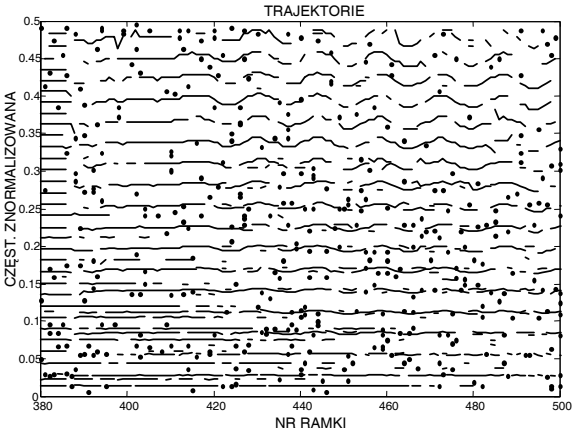


Rys. 4. Schemat blokowy kodera

3.2. Reprezentacja składowych tonalnych

Pierwszym etapem kodowania jest analiza sinusoidalna, w trakcie której dokonuje się detekcji składowych harmonicznych oraz śledzenia zmian częstotliwości tych składowych w czasie (rys. 5). Analiza wykonywana jest w blokach o długości N=2048 próbek, przy

czym kolejne bloki pobierane są co H=512 próbek. Tak duża nadmiarowość pozwala wykorzystać podobieństwo widma chwilowego w śledzeniu trajektorii, nie jest jednak powodem nadmiarowej ich reprezentacji, bowiem aktualizacja wartości przesyłanych parametrów odbywa się z większym krokiem (wielokrotność H).



Rys. 5. Przykładowe trajektorie częstotliwości składowych sinusoidalnych estymowane w pliku testowym (fragment)

Częstotliwości poszczególnych węzłów trajektorii sinusoidalnych są reprezentowane w skali logarytmicznej i bardzo efektywnie kodowane z wykorzystaniem mechanizmu predykcji adaptacyjnej (rząd predyktora jest zmienny i rośnie wraz z upływem czasu od początku trajektorii). Błąd predykcji jest stosunkowo silnie kwantowany (szerokość przedziału kwantyzacji odpowiada błędowi względnemu około 1,5%). Tak duży błąd częstotliwości w zwykłym koderze parametrycznym (np. MPEG-4 SSC) byłby źródłem dokuczliwych zniekształceń, jednak w przypadku proponowanej techniki następuje częściowa kompensacja tego błędu przez oscylacyjną składową zespolonej obwiedni.

Częstotliwości trajektorii sinusoidalnych są w koderze rekonstruowane i interpolowane w sposób zapewniający ciągłość zmian częstotliwości i jej pierwszej pochodnej (w implementacji eksperymentalnej wykorzystano wielomian Hermite'a 3. rzędu). Interpolowane częstotliwości służą do wygenerowania zestawu zespolonych nośnych, co pozwala na indywidualną demodulację (przesunięcie w dziedzinie częstotliwości) do pasma podstawowego każdej ze składowych sinusoidalnych sygnału. Jedną z głównych zalet demodulacji z interpolowaną częstotliwością jest „wyprostowanie" zmian częstotliwości chwilowej poszczególnych harmonicznych, co powoduje zawężenie ich widma.

W celu wyodrębnienia pożądanej składowej wolnozmiennej z produktów demodulacji zastosowano filtr dolnoprzepustowy o symetrycznej odpowiedzi impulsowej. Szerokość pasma filtru jest dobierana automatycznie tak, aby uwzględnić lokalne właściwości widmowe danej składowej. W tym celu w bloku estymacji parametrów modelu sinusoidalnego szacowana jest szerokość głównej wstęgi widma odpowiadającego poszczególnym harmonicznym. Otrzymana informacja determinuje dobór szerokości pasma filtru.

3.3. Kompresja przebiegów obwiedni

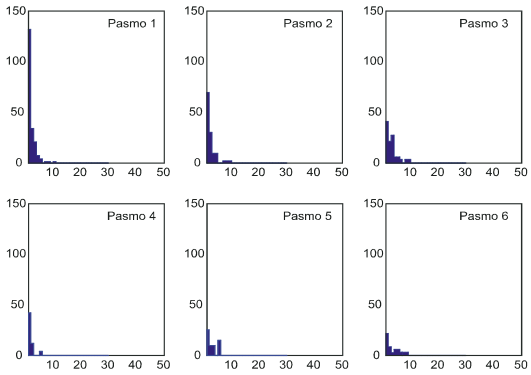
Po odfiltrowaniu niepożądanych produktów koder dysponuje grupą zespolonych obwiedni, które podlegają wąskopasmowemu kodowaniu transformatowemu. Kodowanie transformatowe sygnałów akustycznych zazwyczaj wykorzystuje zmodyfikowane przekształcenie kosinusowe. Można wykazać, że dla sygnałów o wartościach zespolonych optymalnym rozszerzeniem tej metody jest wykorzystanie zespolonej transformacji MCLT,

$$X_m = \sum_{n=0}^{2N-1} x(n)\, w(n) \exp\left[- j \tfrac{\pi}{N}\left(n + \tfrac{N+1}{2}\right)\left(m + \tfrac{1}{2}\right)\right], \quad (2)$$

gdzie $x(n)$ to sygnał w dziedzinie czasu, $w(n)$ oznacza okno Hanna, a m jest indeksem współczynnika [10].

Dobór długości okna transformacji odbywa się według jednoznacznej reguły: dla każdej trajektorii początkowe próbki kodowane są z krótkim oknem (N=256), po czym następuje przełączenie na okna długie (2048 próbek, N=1024). Taka strategia wynika z obserwacji, że początki trajektorii często występują w okolicach transientów, co wiąże się z bardziej dynamicznym zmianami w obwiedni, które wymagają lepszej dokładności czasowej.

Kolejnymi etapami kodowania są: kwantowanie współczynników transformaty oraz kodowanie entropijne. Ze względu na wolnozmienny charakter obwiedni i wynikające z niego bardzo skupione rozkłady współczynników transformaty (rys. 6), efektywność tej kompresji jest dużo wyższa niż pełnopasmowego kodowania transformatowego.



Rys. 6. Przykład rozkładu współczynników MDCT w pierwszych sześciu składowych sinusoidalnych (bez znaku)

Kwantowanie i kodowanie entropijne wartości współczynników poprzedzone jest uszeregowaniem według rosnących indeksów (naprzemiennie dla części rzeczywistej i urojonej każdego współczynnika i niezależnie w grupach odpowiadających obwiedniom poszczególnych składowych sinusoidalnych). Proces kwantyzacji jest wzorowany na technice AAC, to znaczy wykorzystuje nieliniowy kwantyzator skalarny, a stopień stratności jest regulowany przez współczynniki skalujące (globalny gsf i wspólny dla grupy scf_k),

$$\overline{X_m} = \mathrm{sgn}(X_m)\, floor\left[2^{(scf_k - gsf)/4} \left| X_m \right|^{3/4} + 0.0946\right]. \quad (3)$$

Globalny współczynnik skalowania gsf występuje we wszystkich operacjach kwantowania (dla każdego

współczynnika MDCT i każdej składowej harmonicznej). Jego głównym przeznaczeniem jest łączne sterowanie stopniem stratności oraz liczbą bitów generowaną przez koder.

Współczynniki scf_k są jednakowe dla próbek transformaty MDCT opisujących pojedynczy sygnał obwiedni, co powoduje, że energia błędu kwantyzacji jest rozłożona równomiernie w dziedzinie częstotliwości w wąskim paśmie. W zdekodowanym sygnale błąd kwantyzacji ulega powrotnemu przesunięciu w dziedzinie częstotliwości i w efekcie końcowych stanowi wąskie pasmo szumu podążające za zmianami częstotliwości danej składowej harmonicznej.

3.4. Kodowanie pozostałej części sygnału

Zakodowane składowe harmoniczne muszą zostać odtworzone w koderze w celu wyznaczenia sygnału rezidualnego. W tym celu dokonywane jest odwrotne przekształcenie MCLT, a otrzymany sygnał podlega modulacji przy pomocy sprzężonego sygnału nośnej. W ten sposób widma poszczególnych składowych sinusoidalnych (wraz z szumem kwantyzacji) ulegają przesunięciu na właściwe miejsce. Sygnał reziduum uzyskiwany jest w procesie odejmowania widmowego sygnału oryginalnego $x(n)$ i sumy zrekonstruowanych sinusoid $s(n)$

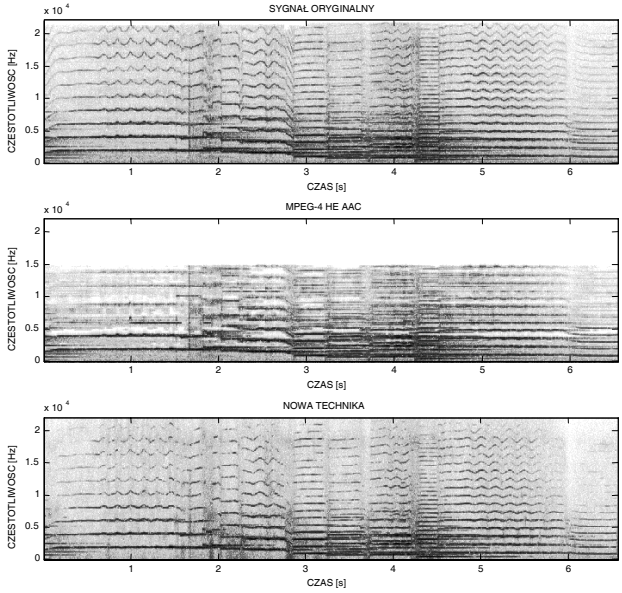
$$R_m(k) = thr\left[\left|X_m(k)\right| - \left|S_m(k)\right|\right]\exp\left[\arg\left\{X_m(k)\right\}\right], \quad (4)$$

gdzie *thr*[ ] jest funkcją progującą (odcinającą wartości ujemne), a w operacja wykonywana jest na reprezentacji czasowo-częstotliwościowej, $X_m(k)=\mathrm{FFT}\{x(n)\, w(n-m)\}$.

Sygnał reziduum zawiera głównie składowe szumowe kodowanego programu. Składowe te są modelowane przy pomocy klasycznej techniki LPC, z długością ramki N=2048 na odcinkach stacjonarnych sygnału oraz N=256 w okolicach transientów. Obwiednia widmowa szumu reprezentowana jest przez 16 współczynników predyktora adaptacyjnego w postaci PARCOR. Współczynniki podlegają kwantyzacji w skali logarytmicznej z rozdzielczością 8 bitów.

4. WYNIKI EKSPERYMENTALNE

Opisany schemat kodowania został zaimplementowany w systemie MATLAB z wykorzystaniem całkowicie autorskiego opracowania modelu sinusoidalnego i pozostałych bloków kodeka. W celu zweryfikowania efektywności kompresji przeprowadzono serię eksperymentów symulacyjnych z wykorzystaniem fragmentów nagrań muzyki z udziałem instrumentów o widmie bogatym w składowe harmoniczne. Wszystkie fragmenty zakodowano kodekiem Nero Digital, będącym komercyjną implementacją standardu MPEG-4 HE AAC z prędkościami transmisji 16kb/s, 20kb/s i 24 kb/s. Te same sygnały zakodowano przy pomocy kodeka eksperymentalnego, przy ustawieniu parametru *gsf* w taki sposób, aby uzyskać podobne przepływności. W każdym z badanych przypadków uczestnicy ślepego testu odsłuchowego wskazywali wyraźną przewagę jakości nowej techniki nad techniką HE AAC. Przewagę tę w sposób oczywisty pokazuje również porównanie spektrogramów (rys. 7).



Rys. 7. Porównanie spektrogramów(w kolejności) sygnału oryginalnego, zakodowanego techniką MPEG-4 AAC przy prędkości transmisji 24kb/s i zakodowanego nową techniką

PODSUMOWANIE

W artykule przedstawiono propozycję nowej hybrydowej techniki kodowania dźwięku łączącej zalety kodowania transformatowego i parametrycznego. Przeprowadzone testy dowodzą, że nowa technika wyraźnie przewyższa pod względem jakości zdekodowanego sygnału technikę standardową MPEG-4 HE AAC.

SPIS LITERATURY

[1] ISO/IEC JTC1/SC29/WG11 MPEG, „Int. Standard ISO/IEC 14496-3/Amd1, Coding of Audio-Visual Objects: Audio", 1999.

[2] ISO/IEC JTC1/SC29/WG11 MPEG, „Int. Standard ISO/IEC 14496-3:2001/Amd1, Bandwidth Extension", 2003.

[3] M. Bartkowiak, „Metody parametryczne we współczesnych technikach kodowania dźwięku szerokopasmowego", *KKRRiT*, Poznań, 2-7 czerwca 2006.

[4] X. Serra, „Musical sound modeling with sinusoids plus noise", w: *Musical Signal Processing*, C Roads (red.), Sweets & Zeitlinger, 1997.

[5] W. Oomen, A. den Brinker, "Sinusoids plus noise modeling of audio signals", *AES 17th Int. Conf. on High-Quality Audio Coding*, 1999.

[6] ISO/IEC JTC1/SC29/WG11 MPEG, „Int. Standard ISO/IEC 14496-3:2005/Dcor2, Sinusoidal Coding", 2005.

[7] S. Marchand, F. Keiler, „Survey on Extraction of Sinusoids in Stationary Sounds", *Digital Audio Effects* (*DAFx'02*) *Conference*, Hamburg, 2002.

[8] M. Lagrange, S. Marchand, J.-B. Rault, „Enhancing the Tracking of Partials for the Sinusoidal Modeling of Polyphonic Sounds", *IEEE Trans. Audio, Speech and Language Proc*., Vol. 15, No. 5, July 2007.

[9] M. Bartkowiak, T. Żernicki, „Improved partial tracking technique for sinusoidal modeling of speech and audio", *Poznańskie Warsztaty Telekomunikacyjne, Poznań, 2007*.

[10] H.S. Malvar, „A modulated complex lapped transform and its applications to audio processing", *Int. Conf. ICASSP'99*, Phoenix, 1999.