Metody kodowania i przechowywania sygnałów dźwiękowych

Andrzej Majkowski
Politechnika Warszawska
amajk@ee.pw.edu.pl

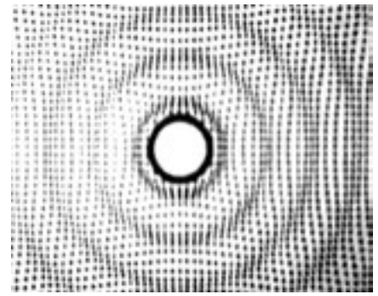
Streszczenie

Wykład jest poświęcony metodom przechowywania i kodowania sygnałów dźwiękowych. Na wstępie przedstawiamy, w jaki sposób człowiek odbiera dźwięki i podajemy podstawowe informacje dotyczące sygnałów dźwiękowych, jak również w jaki sposób oceniamy jakość sygnału dźwiękowego. Następnie przedstawiamy różne formaty zapisu dźwięku. W dalszej części wykładu opisujemy pewne właściwości sygnałów dźwiękowych wykorzystywane podczas kodowania, a szczególnie w kompresji sygnałów dźwiękowych. Poznamy, co to jest psychoakustyka i efekty maskowania. Przedstawiamy również etapy kompresji dźwięku w standardzie MP3. Szczególną uwagę zwracamy na elementy, które znacząco wpływają na jakość kodowania MP3 oraz ułatwiają znalezienie właściwego kompromisu pomiędzy stopniem kompresji a jakością nagrania.

Spis treści

1 DŹWIĘK

Fala dźwiękowa rozchodzi się jako podłużna fala akustyczna w danym ośrodku sprężystym: gazie, płynie (rys. 1). W ciałach stałych, takich jak metale, występuje również fala poprzeczna. Najczęściej mówimy o rozchodzeniu się dźwięku w powietrzu. Dźwięk, jako drgania cząsteczek, charakteryzuje się tym, że cząsteczka pobudzona przekazuje energię cząstce sąsiedniej, a sama drga wokół własnej osi. Skutkiem tego są lokalne zmiany ciśnienia ośrodka rozchodzące się falowo. Co ciekawe, w wodzie dźwięk rozchodzi się znacznie szybciej niż w powietrzu, a w próżni oczywiście nie rozchodzi się w ogóle. W potocznym znaczeniu **dźwięk** to każde rozpoznawalne przez człowieka pojedyncze wrażenie słuchowe.



Rysunek 1.
Fala dźwiękowa [źródło: http://sound.eti.pg.gda.pl/student/elearning/fd.htm]

Zakres częstotliwości od 20 Hz do 20 kHz jest zakresem częstotliwości słyszalnych (fonicznych, audio). Dźwięki o częstotliwości mniejszej od 20 Hz są nazywane **infradźwiękami**, zaś o częstotliwości większej od 20 kHz – **ultradźwiękami.**

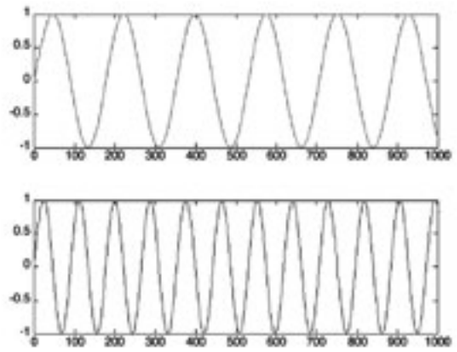
1.1 JAK ODBIERAMY DŹWIĘKI

Elementarnym rodzajem dźwięku, dla którego fala dźwiękowa ma postać sinusoidy (rys. 2) jest **ton**. **Wysokość tonu** to atrybut wrażenia słuchowego, umożliwiający uszeregowanie dźwięków na skali niskie-wysokie. Przez **wysokość dźwięku** rozumie się częstotliwość drgań fali akustycznej – im wyższa częstotliwość drgań tym wyższy dźwięk. Na rysunku 2 częstotliwość drugiego sygnału jest dwa razy większa niż pierwszego, zatem dźwięk o takim przebiegu będzie odbierany jako wyższy. Dźwięki są najczęściej sygnałami złożonymi (występuje w nich wiele składowych sinusoidalnych o różnych amplitudach i częstotliwościach). Wysokość dźwięku, często utożsamiana z częstotliwością, w dużym stopniu zależy od niej, ale nie wyłącznie. Innymi czynnikami wpływającymi na wrażenia wysokości są m.in. natężenie dźwięku czy współobecność innych tonów. Występują też różnice w postrzeganiu wysokości dźwięku między lewym i prawym uchem.

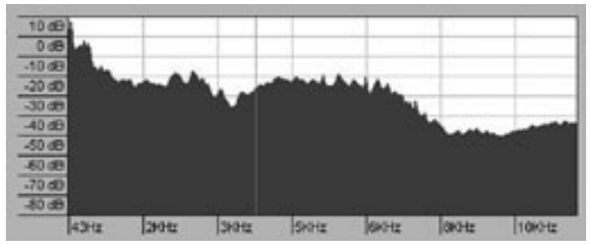
Z pojęciem wysokości dźwięku są związane **interwały muzyczne,** czyli odległości między dźwiękami na skali muzycznej. Określone są stosunkiem częstotliwości sygnałów. Oktawa jest to interwał określający dźwięki, których stosunek częstotliwości jest równy 2:1. Człowiek jest w stanie interpretować poprawnie interwały muzyczne dla tonów o częstotliwości max. ok. 5 kHz. Powyżej 2,5 kHz występują znaczne błędy. Natomiast powyżej częstotliwości 5 kHz występuje brak wrażenia melodii chociaż spostrzegane są różnice częstotliwości.

Bardzo często w analizie sygnału dźwiękowego korzysta się z jego częstotliwościowej reprezentacji. Mówimy wtedy o tzw. **widmie** sygnału dźwiękowego. Widmo sygnału dźwiękowego umożliwia zobrazowanie, jakie składowe sinusoidalne, będące funkcjami czasu, i o jakich częstotliwościach i amplitudach, tworzą dany dźwięk. Rysunek 3 przedstawia przykładowe widmo sygnału dźwiękowego. Oś *Ox* reprezentuje częstotliwość składowych sinusoidalnych, w tym przypadku w zakresie od 43 Hz do 12 000 Hz. Na osi *Oy* można odczytać pośrednio informacje o amplitudach składowych sinusoidalnych.

Barwa dźwięku to cecha wrażenia słuchowego, dzięki której rozróżniamy dźwięki o tej samej głośności i częstotliwości. Barwa dźwięku zależy głównie od jego struktury widmowej, natężenia dźwięku i przebiegu czasowego dźwięku. I tak, interesujące eksperymenty pokazują, że w przypadku niektórych instrumentów ważniejszą rolę odgrywa struktura widmowa (klarnet, trąbka), a innych – czasowa (flet). Kluczową rolę odgrywa też proces narastania i trwania dźwięku.



Rysunek 2.
Dwa sygnały sinusoidalne o tych samych amplitudach, przy czym częstotliwość pierwszego sygnału jest dwa razy mniejsza niż drugiego



Rysunek 3.
Widmo sygnału dźwiękowego

Słuch ludzki charakteryzuje pewna niesymetryczność w odbiorze wysokości dźwięków w uchu lewym i prawym. U zdrowego człowieka różnice nie przekraczają zwykle 3%. Osoby o słuchu muzycznym potrafią określić wysokość dźwięku z dokładnością do 0,3-1%.

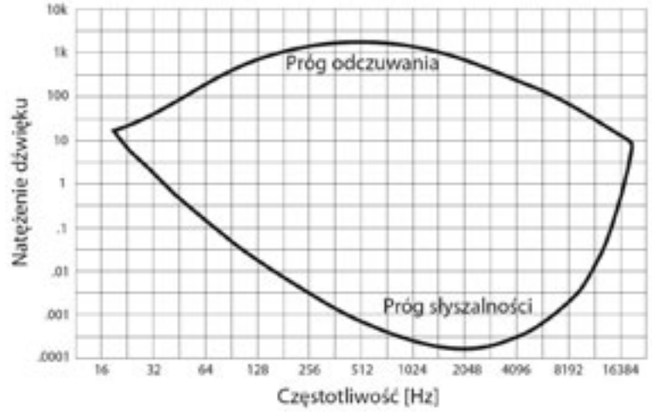
1.2 ZAKRES SŁYSZALNOŚCI

Głośność to taka cecha wrażenia słuchowego, która umożliwia uszeregowanie dźwięków na skali głośno-cicho. Teoretycznie ucho ludzkie potrafi odebrać i przetworzyć drgania o częstotliwości 16 Hz do 20 kHz. Jest to jednak duże uproszczenie niemające wiele wspólnego z rzeczywistością. Okazuje się, że powyższy zakres jest słyszalny tylko wtedy, gdy energia dźwięku jest duża. Przy cichych dźwiękach czułość ucha drastycznie maleje w obszarze częstotliwości poniżej 200 Hz oraz powyżej 8 kHz. W tych zakresach trudniej jest również rozróżniać wysokość dźwięku. Zakres częstotliwościowy percepcji dźwięków maleje też wraz z wiekiem.

Na wrażenie głośności dźwięku wpływa wiele dodatkowych czynników, np. czas trwania dźwięku. Dla krótkich czasów trwania dźwięków występuje efekt czasowego sumowania głośności. Natomiast dla czasów od ok. 1 s do ok. 3 min, dla dźwięków o niskim poziomie lub wysokiej częstotliwości, głośność maleje ze wzrostem czasu trwania. Jest to efektem adaptacji głośności. W wyniku efektu sumowania głośności powiększenie szerokości pasma częstotliwościowego szumu białego powoduje wzrost głośności. Głośność szumu (i dźwięków złożonych) jest wyższa niż tonów (sinusoidalnych) o takim samym natężeniu dźwięku.

Próg słyszalności (próg absolutny, próg detekcji sygnału) jest to najmniejszy poziom ciśnienia akustycznego dźwięku, który wywołuje zaledwie wyczuwalne wrażenie słuchowe wobec braku innych dźwięków. Najniższa wartość ciśnienia akustycznego (przy częstotliwości 1000 Hz) wykrywanego przez ucho ludzkie wynosi średnio 20μPa (rys. 4). **Próg bólu** jest to wartość ciśnienia akustycznego, przy której ucho odczuwa wrażenie bólu. Jest ono prawie niezależne od częstotliwości i wynosi 140 dB dla dźwięków sinusoidalnych oraz 120 dB dla szumów. Wrażenie bólu wywołane jest reakcją mięśni bębenka i kosteczki ucha środkowego na impulsy wysokiego ciśnienia akustycznego. Reakcja ta ma na celu ochronę aparatu słuchowego przed ewentualnymi uszkodzeniami.

Okazuje się, że człowiek nie wszystkie dźwięki o tym samym poziomie głośności słyszy jednakowo dobrze. Dźwięki bardzo niskie i bardzo wysokie są słyszane słabo, za to tony o częstotliwościach od 1 kHz do 5 kHz (mniej więcej zakres mowy ludzkiej) są słyszane wyjątkowo dobrze. Na przykład ton 10 dB mający częstotliwość 1000 Hz będzie przez większość ludzi świetnie słyszalny, ale ton 10 dB o częstotliwości 25 Hz chyba wszyscy odbierzemy jako ciszę. Uświadomienie sobie faktu, że nie wszystkie dźwięki o tej samej energii są przez ludzkie ucho rozpoznawane jako tak samo głośne, to dopiero początek problemów związanych z pojęciem głośności. Następnym problemem jest to, że ucho działa nieliniowo. Oznacza to, że dwa razy większe natężenie dźwięku wcale nie jest przez nas odbierane jako dwa razy głośniejszy dźwięk. Ucho dokonuje silnego spłaszczenia odczuwania głośności – dźwięk, który odczuwamy jako kilka razy głośniejszy od początkowego, ma w rzeczywistości energię dziesiątki, a nawet setki razy większą.



Rysunek 4.
Zakres słyszalności człowieka

1.3 OCENA JAKOŚCI DŹWIĘKU

Układ słuchowy, tak jak wzrokowy, jest instrumentem nieliniowym, a odbierane przez niego dźwięki są interpretowane w różny sposób przez różne osoby. Wpływ na sklasyfikowanie odbieranego dźwięku mają między innymi wspomnienia, wiedza, doświadczenie i uszkodzenia narządu słuchowego. Ocena jakości dźwięku przeprowadzona przez dwie osoby może dać zatem bardzo różne wyniki.

2 FORMATY ZAPISU I PRZECHOWYWANIA PLIKÓW MULTIMEDIALNYCH

Pliki przechowujące materiały multimedialne często muszą umożliwić zapis i przechowywanie różnego rodzaju danych: dźwięków, obrazów, filmów, napisów itp. Potrzebny jest do tego specjalny format zapisu danych, który będzie umożliwiał poprawne wyświetlenie lub synchronizację danych w celu ich jednoczesnego odtworzenia. Taki format zapisu nazywa się **kontenerem multimedialnym**. Istnieją 3 typy kontenerów multimedialnych:

- kontenery audio;
- kontenery audio-wideo;
- kontenery obrazkowe.

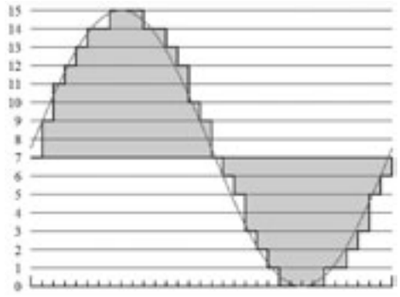
Przykładami kontenerów multimedialnych są:

AVI (ang. *Audio Video Interleave*) jest kontenerem multimedialnym stworzonym przez firmę Microsoft w roku 1992 jako część projektu Video for Windows. W kontenerze tym mogą być zawarte zarówno strumienie audio-wizualne, jak i dane służące do ich synchronizacji.
OGG jest bezpłatnym otwartym kontenerem dla multimediów wysokiej jakości. Wyróżniamy następujące rozszerzenia plików OGG, które są związane o określonym typem danych multimedialnych: .oga – pliki zawierające muzykę, .ogv – pliki zawierające wideo, .ogx – pliki zawierające aplikacje, .ogg – pliki zawierające muzykę w formacie Vorbis.

MPEG-4, wprowadzony pod koniec roku 1998, jest oznaczeniem grupy standardów kodowania audio i wideo wraz z pokrewnymi technologiami, opracowanej przez grupę MPEG (ang. *Moving Picture Experts Group*). Główne zastosowania MPEG-4 to: media strumieniowe w sieci Web (technika dostarczania informacji multimedialnej na życzenie, najpopularniejsze media strumieniowe opierają się na transmisji skompresowanych danych multimedialnych poprzez Internet), dystrybucja CD, DVD, wideokonferencje, telewizja. Oficjalne rozszerzenie pliku to .mp4. MPEG-4 może przechowywać zarówno dane audio-wideo, jak i teksty i obrazki. Może przechowywać dane zachowane praktycznie w każdym formacie.

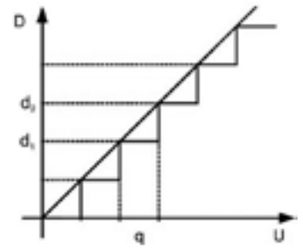
Dźwięk przechowywany w kontenerze multimedialnym musi być zapisany w postaci cyfrowej. Jedną z najpopularniejszych metod zapisu sygnału dźwiękowego jest **PCM** (ang. *Pulse Code Modulation*). Ta metoda jest używana w telekomunikacji, w cyfrowej obróbce sygnału (np. w procesorach dźwięku), do zapisu na płytach CD (CD-Audio) i w wielu zastosowaniach przemysłowych.

Metoda PCM polega na reprezentacji wartości chwilowej sygnału (**próbkowaniu**) w określonych (najczęściej równych) odstępach czasu (rys. 5), czyli z określoną częstością (tzw. **częstotliwością próbkowania**).



Rysunek 5.
Ilustracja zapisu dźwięku w formacie PCM przy 4-bitowym kodowaniu [źródło: http://pl.wikipedia.org/wiki/PCM]

Wartość chwilowa sygnału jest przedstawiana za pomocą słowa kodowego, którego wartości odpowiadają wybranym przedziałom kwantyzacji sygnału wejściowego. Przydział zakresu wartości analogowej jednej wartości cyfrowej jest nazywany **kwantyzacją sygnału,** prowadzi to do pewnej niedokładności (błąd kwantyzacji). Ilustracja kwantyzacji jest przedstawiona na rysunku 6. Z konkretnego przedziału kwantyzacji *q* wartości analogowe z przedziału od d_1 do d_2 zostaną zastąpione jedną wartością zapisaną cyfrowo, najbliższą liczbie d_1. Liczba poziomów kwantyzacji jest zazwyczaj potęgą liczby 2 (ponieważ do zapisu próbek używane są słowa binarne) i wyraża się wzorem 2^n, gdzie *n* to liczba bitów przeznaczona na pojedynczą próbkę. Im większa częstotliwość próbkowania i im więcej bitów słowa kodowego reprezentuje każdą próbkę, tym dokładność reprezentacji jest większa, a tak zapisany sygnał jest wierniejszy oryginałowi. Dobór częstotliwości próbkowania w taki sposób, aby połowa częstotliwości próbkowania (częstotliwość Nyquista) była większa od najwyższej częstotliwości składowej sinusoidalnej występującej w sygnale dźwiękowym (analiza widmowa), umożliwia bezstratną informacyjnie zamianę sygnału ciągłego na dyskretny.



Rysunek 6.
Kwantyzacja sygnału

Dźwięk w formacie PCM może być zapisywany z różną częstotliwością próbkowania, najczęściej jest to 8 kHz (niektóre standardy telefonii), 44,1 kHz (płyty CD-Audio) oraz różną rozdzielczością, najczęściej 8, 16, 20 lub 24 bity na próbkę, może reprezentować 1 kanał (dźwięk monofoniczny), 2 kanały (stereofonia dwukanałowa) lub więcej (stereofonia dookólna). Reprezentacja dźwięku próbkowana z częstotliwością 44,1 kHz i w rozdzielczości 16 bitów na próbkę (65 536 możliwych wartości amplitudy fali dźwiękowej na próbkę) jest uważana za bardzo wierną swemu oryginałowi, ponieważ z matematycznych wyliczeń wynika, iż pokrywa cały zakres pasma częstotliwości słyszalnych przez człowieka oraz prawie cały zakres rozpiętości dynamicznej słyszalnych dźwięków. Taki format kodowania zastosowano na płytach CD-Audio.

Inne formy cyfrowego kodowania dźwięku są zazwyczaj dużo bardziej złożone. Często wykorzystują różne metody kompresji danych w celu zredukowania ich liczby. Istnieją 2 rodzaje kompresji:

- **kompresja bezstratna** – algorytm upakowania informacji do postaci zawierającej mniejszą liczbę bitów w taki sposób, aby informację dało się odtworzyć do postaci identycznej z oryginałem,
- **kompresja stratna** – algorytm zmniejszania liczby bitów potrzebny do wyrażenia danej informacji, przy czym nie ma gwarancji, że odtworzona informacja będzie identyczna z oryginałem. Dla niektórych danych algorytm kompresji stratnej może odtworzyć informację prawie idealnie.

Przetworzenie pliku dźwiękowego do określonego formatu cyfrowego wymaga specjalnego programu, tzw. **kodeka**, w którym zaimplementowane są zaawansowane algorytmy cyfrowego przetwarzania sygnałów dźwiękowych. Poniżej krótko opisano najpopularniejsze kodeki dźwięku. W dalszej części szerzej będzie opisany sposób kodowania MP3.

Ogg Vorbis jest kodekiem ogólnego zastosowania. Najlepiej sprawdza się w tworzeniu plików o dużym stopniu kompresji (od 48 do 128 kbps). Uznaje się, że średnia jakość dźwięku zakodowanego w formacie Ogg Vorbis jest porównywalna do AAC i wyższa niż MP3 o tej samej **przepływności** (czyli szybkości transmisji danych

mierzonej w bitach na jednostkę czasu). W odróżnieniu od MP3, format Ogg Vorbis nie jest opatentowany i pozostaje bezpłatny, zarówno do celów prywatnych, jak i komercyjnych. Dekodowanie plików zapisanych w tym formacie wymaga większego zapotrzebowania na moc obliczeniową procesora niż MP3 (w przenośnych odtwarzaczach szczególnie uwidacznia się to poprzez skrócenie czasu pracy). Jest kodekiem z natury typu VBR (czyli dźwięk jest kodowany ze zmienną w czasie szybkością przepływu danych).

MPEG-4 Part 14 został utworzony w oparciu o format kontenera Apple QuickTime i jest właściwie identyczny z formatem MOV, ale wspiera wszystkie właściwości standardu MPEG. Pliki z zakodowanym dźwiękiem mają często rozszerzenie .mp4, nie istnieje natomiast coś takiego jak format kompresji dźwięku MP4.

AAC (ang. *Advanced Audio Coding*) to z kolei algorytm stratnej kompresji danych dźwiękowych, którego specyfikacja została opublikowana w 1997 roku. Format AAC zaprojektowany został jako następca MP3, oferujący lepszą jakość dźwięku przy podobnym rozmiarze danych.

Kompresja AAC jest modularna i oferuje standardowo cztery profile:

- Low Complexity (LC) – najprostszy, najszerzej stosowany i odtwarzany przez wszystkie odtwarzacze obsługujące format AAC;
- Main Profile (MAIN) – rozszerzenie LC;
- Sample-Rate Scalable (SRS) lub Scalable Sample Rate (AAC-SSR) – zakres częstotliwości dzielony jest na cztery kompresowane niezależnie pasma, jakość jest przez to nieco niższa niż pozostałych profili;
- Long Term Prediction (LTP) – rozszerzenie MAIN wymagające mniejszej liczby obliczeń.

Usprawnienia AAC w stosunku do poprzednich algorytmów kompresji dźwięku:

- próbkowanie 8–96 kHz (MP3 16–48 kHz);
- do 48 kanałów (MP3 – 2 kanały w standardzie MPEG-1 i 5,1 w standardzie MPEG-2);
- skuteczniejszy i wydajniejszy;
- lepsze przenoszenie częstotliwości ponad 16 kHz;
- lepszy tryb kompresji sygnału stereofonicznego joint stereo.

3 PSYCHOAKUSTYKA I PODSTAWY KOMPRESJI SYGNAŁÓW DŹWIĘKOWYCH

Psychoakustyka to współczesna dziedzina wiedzy zajmująca się związkiem obiektywnych (fizycznych) cech dźwięku z jego cechami subiektywnymi, z wrażeniem jakie w mózgu słuchacza wywołują bodźce dźwiękowe. Psychoakustyka próbuje przewidzieć zachowanie się słuchu człowieka w określonych warunkach fizycznych.

Modelami psychoakustycznymi nazywamy modele systemu słyszenia, które uwzględniają ograniczenia i tolerancje mechanizmów percepcji przeciętnego słuchacza, są to modele matematyczne mówiące, jakie dźwięki są rozpoznawalne przez ludzkie ucho, jakie natomiast nie są. Modele psychoakustyczne są podstawą między innymi kompresji dźwięku, algorytmów oceny jakości transmisji mowy, systemów automatycznie rozpoznających mowę oraz systemów rozpoznających mówców.

Wytyczne do modelowania pochodzą z pomiarów psychoakustycznych (odsłuchowych), w których słuchacze oceniają wrażenia wywołane różnymi sygnałami testowymi prezentowanymi w określonym kontekście (np. czy słyszą ton sinusoidalny prezentowany na tle szumu). Model przetwarza sygnał w taki sposób, aby jego wyjście stanowiło predykcję subiektywnych ocen słuchaczy. Najprostszym faktem psychoakustycznym jest różna czułość ludzkiego ucha na dźwięki o różnych częstotliwościach (niektórych częstotliwości np. bardzo wysokich lub bardzo niskich nie słyszymy w ogóle). Modele psychoakustyczne przewidują zwykle zakres

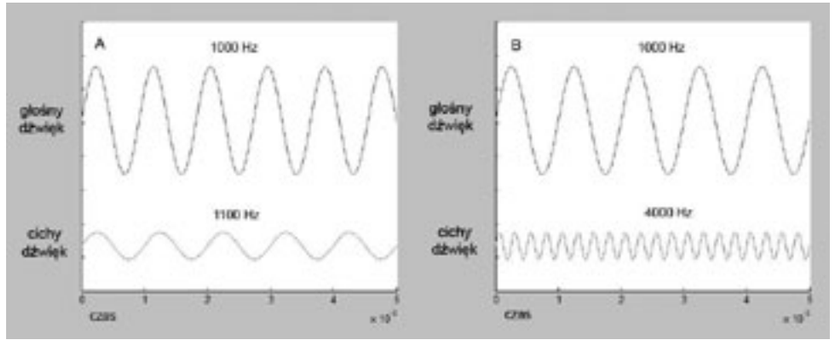
słyszalności od 20 Hz do 20 kHz (dlatego właśnie większość współczesnych odtwarzaczy muzyki zapisanej cyfrowo ma takie pasmo przenoszenia) i maksymalną czułość w zakresie od 2 kHz do 4 kHz.

Innym szeroko stosowanym faktem psychoakustycznym jest **maskowanie dźwięków**. Najogólniej, maskowanie polega na przysłanianiu sygnałów słabszych sąsiadujących z sygnałami znacznie głośniejszymi, które je zagłuszają.

Rozróżniamy 2 rodzaje maskowania:

- maskowanie równoczesne,
- maskowanie czasowe.

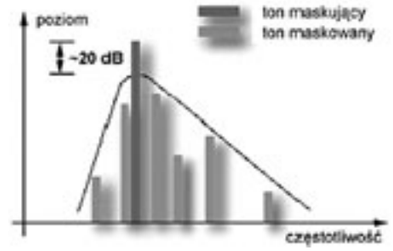
Efekt maskowania równoczesnego opiera się na tym, że człowiek nie jest w stanie odróżnić dwóch dźwięków o zbliżonej częstotliwości, jeśli jeden z nich jest znacznie głośniejszy od drugiego (rys. 7, przypadek A). Możliwe jest to dopiero wtedy, gdy sygnały mają zupełnie różne częstotliwości (przypadek B).



Rysunek 7.
Efekt maskowania równoczesnego

Najprościej mówiąc, maskowanie równoczesne polega na tym, że ciche dźwięki o częstotliwościach zbliżonych do częstotliwości dźwięku głośnego nie są słyszalne. Wszystkie standardy MPEG audio (a więc również MP3) wykorzystują tę właściwość ucha ludzkiego, bazują one na usuwaniu słabszych dźwięków, które nie docierają do mózgu człowieka.

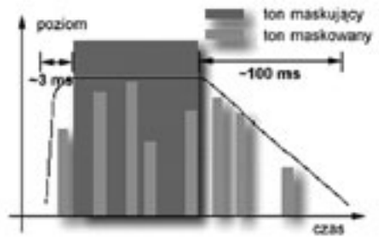
Maskowanie czasowe polega na eliminacji składowych o mniejszym natężeniu, które mają zbliżoną częstotliwość do dźwięku o większym natężeniu i występują razem w pewnym przedziale czasu.



Rysunek 8.
Efekt maskowania czasowego [źródło: 6]

Na rysynku 8 jest pokazany efekt maskowania czasowego, czarną linią zaznaczono próg słyszalności. Można w tym przypadku wyróżnić dwa typy maskowania:

- maskowanie dźwięków następujących (maskowanie pobodźcowe) – głośny dźwięk potrafi zagłuszyć cichsze dźwięki następujące zaraz po nim,
- maskowanie dźwięków poprzedzających (maskowanie wsteczne) – cichy dźwięk poprzedzający w krótkim czasie dźwięk głośny nie jest słyszalny. Ta własność układu słuchowego jest szczególnie ciekawa, gdyż nie da się jej wyjaśnić na gruncie adaptacji krótkoterminowej układu słuchowego. Równocześnie pokazuje ona, że układ słuchowy nosi pewne cechy układu nieprzyczynowego (tzn. skutek wywołany przez jakiś bodziec występuje przed wystąpieniem bodźca).



Rysunek 9.
Całkowity efekt maskowania [źródło: 6]

Na rysunku 9 zilustrowano efekt maskowania równoczesnego i czasowego jednocześnie. Czarna linia oznacza próg słyszalności. Słabe dźwięki (tony maskowane), które są maskowane przez dźwięk silniejszy, mogą zostać podczas kompresji usunięte. Pozostanie tylko dźwięk słyszalny (ton maskujący).

4 IDEA KOMPRESJI MP3

W 1987 roku w niemieckim instytucie Fraunhofera rozpoczęto prace nad radiofonią cyfrową. Jednym z kluczowych elementów było opracowanie systemu kompresji danych umożliwiającego skuteczny zapis sygnałów dźwiękowych. Algorytmy tam opracowane stały się później podstawą systemu MP3.

Należy zaznaczyć, że algorytm stosowany przy kompresji MP3 wykorzystuje kompresję stratną – przy odtwarzaniu, dźwięk nie odpowiada dokładnie dźwiękowi sprzed kompresji. Kompresja powoduje nawet ponad dziesięciokrotne zmniejszenie wielkości miejsca na dysku w stosunku do objętości dźwięku, który kompresji nie podlegał. Osoby z bardziej wrażliwym słuchem odbierają dźwięk skompresowany jako gorszy pod względem jakości.

W roku 1991 ukończone zostały prace w instytucie Fraunhofera nad algorytmem kodowania MPEG-1 – Layer3. Opracowany algorytm stał się najbardziej optymalnym sposobem kodowania sygnałów audio w rodzinie określanej przez międzynarodowe normy ISO-MPEG. Używając tego algorytmu – znanego powszechnie w Internecie jako **MP3**, ze względu na rozszerzenie – do kodowania plików audio, jakość „prawie CD", tj. stereo, 44 kHz, 16 bitów, można uzyskać przy przepływności 112–128 kbps (stopień kompresji 11:1–13:1).

Kompresja MP3 jest oparta na matematycznym modelu psychoakustycznym ludzkiego ucha.

- Idea kompresji MP3 polega na wyeliminowaniu z sygnału tych danych, które są dla człowieka niesłyszalne lub które słyszymy bardzo słabo.
- Kompresja MP3 jest połączeniem metody kompresji stratnej z kompresją bezstratną.
- Etap 1 – koder eliminuje z sygnału składowe słabo słyszalne i niesłyszalne dla człowieka (kompresja stratna).
- Etap 2 – uzyskane dane są poddawane dodatkowej kompresji w celu eliminacji nadmiarowości (kompresja bezstratna).

Algorytm operuje na dźwięku próbkowanym z jakością: 16; 22,5; 24; 32; 44,1 oraz 48 kHz. Jest optymalizowany pod wyjściową przepustowość 128 kbps dla sygnału stereo, aczkolwiek dostępne są przepustowości od 32 kbps do 320 kbps.

Algorytm kodowania MP3 może operować na 4 rodzajach dźwięku wejściowego:

- mono;
- stereo – kompresja dwóch oddzielnych strumieni;
- joint stereo – badane jest podobieństwo sygnałów w obu kanałach; jeśli w obu kanałach jest ten sam sygnał, to koder przełącza się do trybu mono; umożliwia to kodowanie dźwięku z większą dokładnością;
- dual channel – zawiera dwa niezależne kanały, jest stosowany np. przy tworzeniu kilku różnych wersji językowych dla filmu.

W procesie kodowania MP3 występuje kilka procesów, które wymagają dodatkowego wyjaśnienia. Należą do nich dyskretna transformacja kosinusowa, kwantyzacja oraz kodowanie Huffmana.

Dyskretna transformacja kosinusowa (DCT) pomaga rozdzielić sygnał na części, przekształcając dane do postaci umożliwiającej zastosowanie efektywnych metod kompresji. DCT przetwarza sygnał określony w dziedzinie czasu na sygnał określony w dziedzinie częstotliwości. W wyniku działania transformaty na sygnale wejściowym powstają odpowiadające mu współczynniki transformaty. Transformata kosinusowa jest odwracalna, to znaczy, że dysponując tylko współczynnikami transformaty można odtworzyć odpowiadający im sygnał bez żadnych strat. Zaletą transformaty DCT jest to, że większość współczynników jest zwykle bliska zeru, a zatem po procesie kwantyzacji współczynniki te można pominąć, co umożliwia lepszą kompresję danych.

Kwantyzacja jest to proces ograniczenia zbioru wartości sygnału w taki sposób, aby można go było zapisać na skończonej liczbie bitów. Polega na przypisaniu wartości analogowych do najbliższych poziomów reprezentacji, co oznacza nieodwracalną utratę informacji (rys. 6). Kwantyzacja polega na przeskalowaniu współczynników DCT poprzez podzielnie ich przez właściwy współczynnik znajdujący się w tabeli kwantyzacji, a następnie zaokrągleniu wyniku do najbliższej liczby całkowitej. Tablice kwantyzacji dobierane są doświadczalnie.

Kodowanie Huffmana to bezstratna metoda kodowania, przedstawiona przez Davida Huffmana w roku 1952. Kodowanie Huffmana stanowi jedną z najprostszych i jednocześnie łatwych w implementacji metod kompresji bezstratnej. W algorytmie jest wykorzystywany fakt, że pewne wartości danych występują częściej niż inne. Jeżeli zatem zakodujemy częściej występujące wielkości za pomocą krótszych słów kodowych, a rzadziej występujące – za pomocą dłuższych, to sumarycznie długość zakodowanych danych będzie krótsza niż przed kodowaniem.

4.1 KODOWANIE DŹWIĘKU W STANDARDZIE MP3

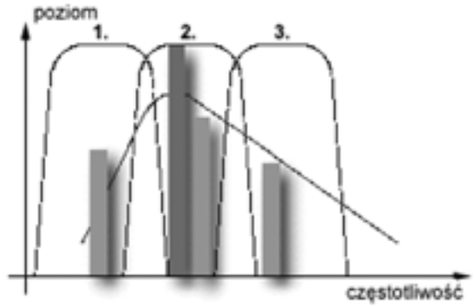
Poniżej podane są najważniejsze etapy kodowania dźwięku w standardzie MP3.

- Sygnał wejściowy jest dzielony na mniejsze fragmenty zwane **ramkami** o czasie trwania ułamka sekundy.
- Dla sygnału dźwiękowego określonego w czasie jest wyliczana jego reprezentacja częstotliwościowa, czyli wyliczane jest widmo sygnału dźwiękowego.
- Widmo sygnału dla każdej ramki jest porównywane z matematycznym modelem psychoakustycznym. W wyniku tego porównania koder określa, które ze składowych dźwięku jako najlepiej słyszalne muszą zostać odwzorowane najwierniej, a które można zakodować w przybliżeniu lub w ogóle pominąć.
- Ustalany jest optymalny przydział bitów na poszczególne częstotliwości pasma akustycznego tak, aby zapewnić możliwie najwierniejsze zakodowanie sygnału.
- Strumień bitów jest poddawany ponownej kompresji poprzez kodowanie Huffmana. Celem tej operacji jest usunięcie nadmiarowości z danych przetworzonych w pierwszym etapie, czyli dodatkowa kompresja bezstratna.

Kolejne ramki poprzedzone nagłówkami są składane w pojedynczy ciąg bitów (strumień bitowy). Nagłówki zawierają metainformacje określające parametry poszczególnych ramek.

Kompresja MP3 rozpoczyna się rozdzieleniem sygnału wejściowego na małe fragmenty (ramki) trwające ułamek sekundy, a następnie ramki są dzielone według pasma na 576 części – najpierw 32 w wielofazowym banku filtrów, a następnie podpasma przekształcane są dyskretną transformatą kosinusową, która generuje 18 współczynników dla każdego podpasma. Zwiększa to szanse na usunięcie niepotrzebnych informacji, sygnał może też być lepiej kontrolowany w celu śledzenia progów maskowania (rys. 11).

Na rysunku 10 zobrazowano ideę działania banku filtrów. Linie pod numerami 1, 2 i 3 oznaczają podział sygnału dźwiękowego na pasma częstotliwościowe 1, 2 i 3. Czwarta linia wyznacza poziom progu słyszalności wyliczony na podstawie modelu psuchoakustycznego. Dwa sygnały oznaczone słupkami po prawej stronie znajdują się poniżej poziomu słyszalności, można więc usunąć sygnał w trzecim podzakresie. Sygnał najbardziej z lewej strony jest słyszalny, można jednak podnieść dopuszczalny poziom szumów, czyli zapisać go mniejszą liczbą bitów. Jeśli kwantowany dźwięk da się utrzymać poniżej progu maskowania, to efekt kompresji powinien być nieodróżnialny od oryginalnego sygnału.



Rysunek 10.
Idea działania banku filtrów [źródło: 6]

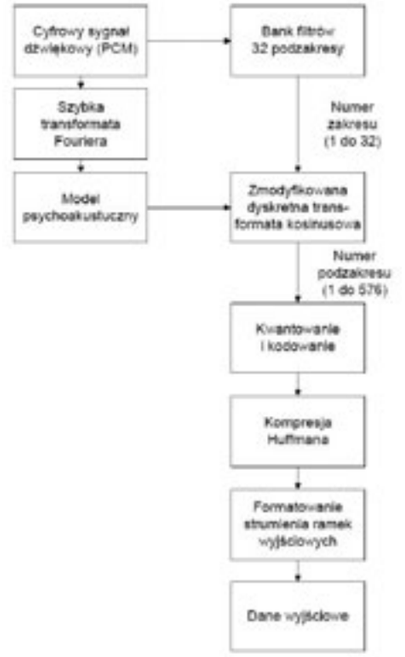
Proces kwantyzacji w kompresji MP3 jest realizowany na zasadzie dwóch pętli, jedna zagnieżdżona w drugiej (rys. 11). Zawiera on także część procesu formowania dźwięku.

Pierwsza z pętli, wewnętrzna, to pętla kontroli współczynnika kompresji. Przeprowadzany jest w niej proces kwantyzacji dla poszczególnych pasm częstotliwościowych, następnie symulowane jest kodowanie skwantowanych współczynników. Jeżeli po kodowaniu okaże się, że jest przekroczony limit przepływności, czyli plik po kompresji byłby zbyt duży, to wskaźnik przyrostu jest dopasowywany do danych i cała pętla jest powtarzana od nowa.

Druga pętla, zewnętrzna, pętla kontroli zniekształceń rozpoczyna się od ustawienia indywidualnych współczynników kwantyzacji na 1, po czym obliczany jest błąd kwantyzacji. Jeśli błąd ten przekracza oszacowany przez model psychoakustyczny próg percepcji, to jest odpowiednio zmieniany współczynnik kwantyzacji i obliczenie błędu odbywa się ponownie. Gdy nie jest możliwe uzyskanie żądanej przepływności i spełnienie wymagań modelu psychoakustycznego, to dźwięk jest kodowany mimo niespełnienia wymagań.

Po procesie kwantyzacji następuje proces kompresji algorytmem Huffmana. W celu dopasowania procesu kompresji do fragmentu danych źródłowych wybierana jest najbardziej pasująca tablica kodów Huffmana z całego zestawu. W celu otrzymania lepszego dopasowania, różne tablice kodów Huffmana są wybierane dla różnych części widma. Jest to proces usuwania nadmiarowych danych bez utraty informacji. Bazuje on na słowie kodowym – kluczu o zmiennej długości, w której klucze krótkie przypisane są do często występujących wzorców, a długie do rzadko występujących. Algorytm rozpoczyna działanie od utworzenia histogramu

(tablicy częstości występowania danych w pliku). W drugim kroku tworzy listę drzew binarnych, które w węzłach przechowują symbol i częstość jego wystąpienia. Następnie w pętli, dopóki jest jeszcze więcej niż jedno drzewo na liście, usuwane są dwa drzewa, które mają w korzeniu zapisane najmniejsze zsumowane częstości, i wstawiane jest nowe drzewo, którego korzeń zawiera sumę częstości usuniętych drzew.



Rysunek 11.
Kodowanie MP3

Końcowym etapem procesu kompresji jest formatowanie ramek wyjściowych i zapis do strumienia wyjściowego. Niektóre pliki MP3 dodatkowo zawierają sumy kontrolne. Suma kontrolna to 16-bitowa liczba, która jest zapisywana w każdej ramce oddzielnie i służy do weryfikacji poprawności strumienia MP3.

4.2 STRUMIEŃ BITOWY

Gęstość strumienia bitowego (ang. *bitrate*) określa współczynnik kompresji sygnału algorytmem MP3. Wyznacza on liczbę bitów przypadającą na sekundę finalnego zapisu. Ustawienie odpowiedniej wartości strumienia bitowego jest kompromisem między jakością a rozmiarem pliku wynikowego (rys. 12).

Rysunek 12.
Ilustracja pojęcia strumienia bitowego [źródło: 8]

Kompresja MP3 może przebiegać:
- ze stałą gęstością strumienia bitowego (ang. *constant bitrate*) – tryb CBR,
- ze zmienną gęstością strumienia bitowego (ang. *variable bitrate*) – tryb VBR.

tryb CBR – każda sekunda dźwięku jest skompresowana za pomocą tej samej liczby bitów, co powoduje jednak, że różne fragmenty utworu mają niejednakową jakość (spokojny fragment wykonany na instrument solo brzmi lepiej niż mocne uderzenie całej orkiestry wspomaganej chórem),
tryb VBR – koduje sygnał uwzględniając jego dynamikę, dzięki czemu przydziela więcej bitów fragmentom sygnału, który zawiera dużo ważnych informacji oraz mniej bitów dla części sygnału, które są mniej złożone. Każda sekunda dźwięku skompresowana jest za pomocą odpowiednio dobranej liczby bitów, dzięki czemu cały utwór ma stałą jakość. W tym przypadku spokojny fragment wykonany na instrument solo (dający się mocniej skompresować) brzmi tak samo dobrze, co mocne uderzenie całej orkiestry wspomaganej chórem (wymagające mniejszego stopnia kompresji). Kompresja w trybie VBR wymaga podania przedziału tolerancji, w jakim może się zmieniać gęstość strumienia bitowego.

Ponieważ zadana gęstość strumienia bitowego obowiązuje dla każdej ramki, w przypadku bardzo złożonych fragmentów może okazać się niewystarczająca i program kodujący nie będzie w stanie zapewnić żądanej jakości zapisu w ramach przydzielonej liczby bitów. Aby zapobiec temu zjawisku standard MP3 zapewnia możliwość skorzystania z dodatkowej rezerwy umożliwiającej zapisanie nadmiarowych danych, tzw. **rezerwy bitowej**. Rezerwa ta powstaje w miejscu pustych fragmentów ramek, w których po zakodowaniu sygnału zostało trochę miejsca.

4.3 ŁĄCZENIE KANAŁÓW ZAPISU STEREOFONICZNEGO

Jak wiemy, sygnał stereo składa się z dwóch odseparowanych od siebie kanałów. Przez znaczną część czasu kanały te jednak przenoszą jeśli nie identyczne to bardzo zbliżone do siebie informacje. Jeśli tak jest, to wtedy koder MP3 wykorzystuje tzw. **algorytm joint stereo**, który powtarzające się dźwięki w obu kanałach zapisuje jako jeden.

Dodatkową możliwością podczas kodowania sygnału z funkcją joint stereo jest **stereofonia różnicowa**. Polega ona na zapisaniu dwóch ścieżek – kanału środkowego będącego sumą sygnałów R i L oraz kanału bocznego, będącego ich różnicą, służącego później do rekonstrukcji sygnału oryginalnego podczas odtwarzania pliku. Warto dodać, że algorytm joint stereo jest bardzo efektywny – powoduje redukcję do 50% liczbę potrzebnych danych.

Ogólnie algorytm MP3 umożliwia skompresowanie dźwięku do postaci:

dual channel – kanały lewy i prawy są traktowane jako dwa niezależne kanały mono, każdy z nich otrzymuje dokładnie połowę dostępnej przepływności; w praktyce jest nieekonomiczny, nie jest więc używany;

stereo – kanały lewy i prawy są traktowane jako stereo, przepływność dzielona jest pomiędzy kanały dynamicznie (np. jeżeli w lewym kanale akurat jest cisza, to prawy dostaje większą część dostępnej przepływności – daje to lepszą jakość dźwięku w prawym kanale) – używany do kompresji w wysokich przepływnościach (192 kbps i więcej);

joint stereo (stereofonia różnicowa) – kanały lewy i prawy są rozbijane na kanały mid/side (*mid* – środek, czyli to, co jest identyczne w obu kanałach i *side* – otoczenie, czyli to, czym różnią się oba kanały) – używany do kompresji w średnich przepływnościach (128–192 kbps);

intensity stereo – kanały lewy i prawy są zamieniane na jeden kanał mono, do którego jest dodawana informacja o uśrednionym kierunku, z którego dźwięk dochodzi (dzięki czemu podczas odsłuchu dźwięk nie dochodzi ze środka tylko z jakiegoś kierunku) – używany do kompresji w niskich przepływnościach (128 kbps i mniej);

mono – kanały lewy i prawy są zamieniane na jeden kanał mono, który jest potem kompresowany, dźwięk odtwarzany jest jako mono – używany do bardzo niskich przepływności (32 kbps i mniej), głównie do kompresji głosu.

Ciekawostką jest to, że specyfikacja formatu MP3, zawarta w dokumencie ISO/IEC 11172-3, nie określa dokładnie sposobu samego kodowania, a jedynie prezentuje ogólny zarys techniki i podaje wymagany poziom zgodności zapisu z normą. Innymi słowy, ustala ona kryteria, jakie musi spełniać struktura pliku, by można było go sklasyfikować jako zgodny ze standardem MP3. Podejście takie ma na celu promowanie różnorodności implementacji programów kodujących i dekodujących dźwięk w standardzie MP3 realizowanych przez różnych producentów. Specyfikacja ISO pełni jedynie rolę bazowego zestawu reguł, określających sposób funkcjonowania standardu tak, aby za pomocą dowolnego kodera można było wygenerować plik odtwarzany przez dowolny dekoder.

4.4 ZALETY I WADY STANDARDU MP3

Niewątpliwie standard kodowania dźwięku MP3 ma wiele zalet. Do najważniejszych należą:
- duży stopień kompresji – stosując kompresję MP3 uzyskujemy plik wynikowy o rozmiarze ok. 10 razy mniejszym od oryginału;
- możliwość sterowania stopniem kompresji i tym samym dostosowania jakości dźwięku do indywidualnych potrzeb;
- metoda ta umożliwia uzyskanie sygnałów o stosunkowo dobrej jakości;
- dekompresja wymaga znacznie mniejszej mocy obliczeniowej niż kompresja;
- twórcy standardu bezpłatnie udostępnili kod źródłowy programów kodujących i dekodujących, dzięki czemu standard ten stał się niezwykle popularny.

Warto jednak pamiętać, że MP3 to metoda kompresji stratnej, a tym samym uniemożliwia zrekonstruowanie sygnału oryginalnego. Ocena jakości dźwięku odtworzonego z pliku MP3 jest bardzo indywidualnym doznaniem. Ponieważ algorytm opiera się na matematycznym modelu percepcji słuchowej przeciętnego człowieka, to siłą rzeczy zawsze będzie grupa ludzi, która usłyszy brakujące, wycięte dźwięki. Oczywiście bardzo duże znaczenie będą miały tu parametry dobrane przez twórcę pliku. Osoba nadzorująca proces kompresji MP3 nie ma co prawda bezpośredniego wpływu na współczynnik kompresji lub też na poziom stratności, może jednak ustalać liczbę bitów przypadających na sekundę docelowego zapisu tzw. przepływność. A to przekłada się bezpośrednio na jakość.

LITERATURA

1. Barański J., *MP3 – internetowy standard zapisu dźwięku,* „Magazyn Elektroniki Użytecznej” maj 2000
2. Beach A., *Kompresja dźwięku i obrazu wideo*, Helion, Gliwice 2009
3. Butryn W., *Dźwięk cyfrowy*, WKiŁ, Warszawa 2002
4. Butryn W., *Dźwięk cyfrowy. Systemy wielokanałowe*, WKiŁ, Warszawa 2004
5. Czyżewski A., *Dźwięk cyfrowy. Wybrane zagadnienia teoretyczne, technologia, zastosowania*, Exit, Warszawa 2001
6. Kołodziej P., *Komputerowe studio muzyczne i nie tylko. Przewodnik*, Helion, Gliwice 2007
7. Nasiłowski D., *Jakościowe aspekty kompresji obrazu i dźwięku. Poglądowo o DivX*, Mikom, Warszawa 2004
8. Rak R., Skarbek W. (red.), *Wstęp do inżynierii multimediów*, Politechnika Warszawska, Warszawa 2004