ZESZYTY NAUKOWE POLITECHNIKI ŚLĄSKIEJ 1996

Seria: INFORMATYKA z. 31 Nr kol. 1348

Piotr FABIAN

KOMPRESJA SYGNAŁU MOWY

Streszczenie. W artykule przedstawiono niektóre zagadnienia związane ze zmniejszaniem objętości informacyjnej sygnału mowy. Opisano metody reprezentacji sygnału dźwiękowego w systemach cyfrowych, standardowy format WAVE. Przedstawiono wybrane metody kompresji: wykorzystującą transformatę kosinusową, kodowanie widma i predykcję liniową, problemy występujące przy implementacji tych metod oraz ich porównanie pod względem osiąganego współczynnika kompresji, szybkości działania i jakości odtwarzanego sygnału.

SPEECH SIGNAL COMPRESSION

Summary. The paper presents some problems related to reducing size of digital data streams describing speech signals. Methods of acoustic signal representations in computer systems (including WAVE format) are reviewed. Three methods of compression are presented: discrete cosine transform coding, spectrum coding and linear predictive coding. Implementation problems, compression ratio, speed and quality of output signal are compared.

KOMPRESSION DES SPRACHSIGNALS

Zusammenfassung. In diesem Artikel wurden Probleme die im Bereich der Sprachkompression wichtig sind beschrieben. Methoden der Representation von Audiosignalen (auch der WAVE Format) wurden vorgestellt. Drei Kompressionmethoden wurden vorgestellt: DCT Kompression, Spektrum Kompression und LPC Kompression. Probleme der Implementation, Kompressions Faktors, Geschwindigkeit und Qualität des Ausgangsignals wurden verglichen.

1. Wprowadzenie

Komunikacja z komputerem odbywa się od kilkudziesięciu lat głównie za pośrednictwem klawiatury. Jest to wygodny sposób dla konstruktorów sprzętu, ale nie dla użytkownika. W odpowiedzi na ich wymagania powstały różne urządzenia uzupełniające klawiaturę, takie jak pióra świetlne, myszki, digitizery i inne. Również prezentacja wyników pracy komputera rozwinęła się od prostych wydruków na drukarce wierszowej do rozbudowanej grafiki wyświetlanej na monitorze. Rozwój techniki pozwala na coraz to nowe zastosowania komputerów. Od kilku już lat możliwa jest prezentacja w postaci ruchomych obrazów połączonych z dźwiękiem (multimedia), znacznie lepiej przemawiających do wyobraźni odbiorcy niż ciągi liczb czy teksty. Duże objętości pamięci stałych (CD ROM) pozwalają zawrzeć w atrakcyjnej formie informacje z encyklopedii, słowników i dużych baz danych, a istniejące sieci komputerowe skracają czas dostępu do interesującej informacji do kilku minut. Komunikacja z takimi systemami powinna nadążać za rozwojem formy prezentacji i nie ograniczać się do klawiatury. Naturalnym dla człowieka sposobem porozumiewania się jest mowa. Wydaje się więc sensowne uzupełnienie urządzeń wejściowych komputera właśnie o wejście i wyjście akustyczne obsługiwane przez odpowiednie oprogramowanie.

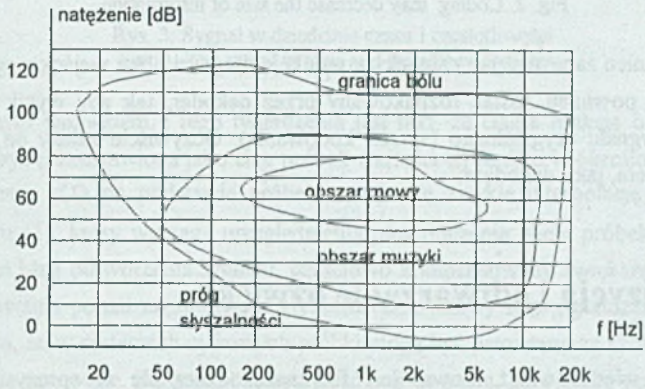
Równie ważną dziedziną wykorzystującą sygnał mowy jest telekomunikacja. Obecnie prawie wszystkie budowane łącza międzycentralowe są cyfrowe; istnieje też standard ISDN definiujący protokół transmisji cyfrowej do i od końcowego abonenta. Przekazanie mowy przez takie łącze wymaga przetworzenia jej na postać cyfrową i rekonstrukcji po stronie odbiorczej.

Jednym z czynników opóźniających szersze zastosowanie mowy w komunikacji maszyny (komputera) z człowiekiem jest **znaczna objętość informacyjna** sygnału mowy zapisywanego bez przetwarzania w pamięci, a w przypadku rozpoznawania - **stopień komplikacji** tego procesu. Przetwarzanie dźwięku (podobnie jak grafiki) wymaga nie tylko **dużych pamięci**, ale także **znacznych mocy obliczeniowych**. Ponieważ mowa jest sposobem komunikacji rozwijanym przez człowieka od około 250 tysięcy lat, człowiek bez większych problemów mówi i rozpoznaje słowa wypowiadane przez inną osobę. Dzieje się tak dzięki dużym obszarom mózgu zaangażowanym w proces artykulacji i rozpoznawania mowy. Jak bardzo skomplikowany jest to proces, można się dopiero przekonać próbując zbudować maszynę mówiącą i rozpoznającą mowę.

W niniejszym artykule przedstawiono wybrane metody umożliwiające zmiejszenie objętości informacyjnej sygnału mowy.

2. Sygnał mowy

Zakres odbieranych przez człowieka częstotliwości i natężeń dźwięku jest ograniczony budową narządu słuchu. Obrazuje go przedstawiony na rys. 1 obszar słyszalności człowieka.

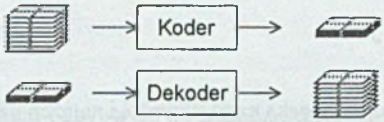


Rys. 1. Obszar słyszalności człowieka
Fig. 1. Human hearing area

Obszar mowy zajmuje stosunkowo niewielką część obszaru słyszalności. Taką co najmniej dynamikę i zakres częstotliwości należy zapewnić w systemie rejestracji (transmisji) sygnału dźwiękowego, aby mowa mogła zostać zrozumiana. W telekomunikacji przyjmuje się nawet węższy zakres częstotliwości: od 300Hz do 3400Hz i dynamikę ok. 50dB.

3. Kompresja i dekompresja

Przetwarzanie sygnału dźwiękowego, które rozważamy, polega głównie na jego kompresji (z utratą informacji) i zapisie w formie zajmującej mniejszą niż oryginał objętość. Ze względu na wymagania dotyczące szybkości komputera przetwarzanie nie odbywa się „w czasie rzeczywistym": wejściowymi i wyjściowymi danymi dla programów są pliki (rys. 2).



Rys. 2. Kodowanie może zmniejszyć objętość informacyjną sygnału
Fig. 2. Coding may decrease the size of information

Koder powinien zapewnić jak największą redukcję długości pliku wejściowego. Uzyskany plik wyjściowy powinien zostać rozpakowany przez dekoder, tak aby wynik jak najmniej różnił się od sygnału oryginalnego (sprzed kodowania). Oczywiście zależy on zarówno od konstrukcji kodera, jak i dekodera.

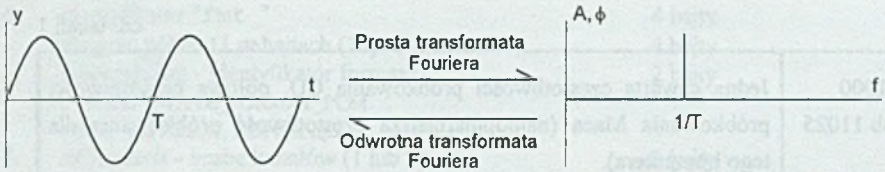
4. Akwizycja i odtwarzanie dźwięku

Dźwięk, a więc również mowa, jest falą rozchodzącą się w sprężystym ośrodku (powietrzu). Drgania cząsteczek ośrodka powodują powstanie dodatkowego ciśnienia p ponad ciśnienie równowagi ośrodka. Zmiany p w dobry sposób opisują falę dźwiękową w jednym punkcie ośrodka. Znajomość zmian ciśnienia w funkcji czasu $p(t)$ jest wystarczająca do odtworzenia fali dźwiękowej. Funkcja $p(t)$ jako opisująca wielkość fizyczną przyjmuje wartości rzeczywiste ze skończonego przedziału i jest ciągła.

Z twierdzenia Kotielnikowa-Shannona wynika, że sygnał $x(t)$ o widmie $G(f)$ może być przedstawiony w równoważny sposób jako ciąg próbek $y_i = x(t_i)$ sygnału wejściowego w wybranych chwilach t_i, jeśli $G(f) = 0$ dla $f > f_g$ oraz próbki pobierane są w jednakowych odstępach czasowych nie większych od $\frac{1}{2f_g}$ ($2f_g$ to częstotliwość Nyquista). Częstotliwość f_g nazywana jest górną częstotliwością graniczną sygnału $x(t)$. Równoważność obu reprezentacji (rys. 3) oznacza możliwość (przynajmniej teoretyczną) odtworzenia na podstawie próbek y_i funkcji $x(t)$ dla dowolnej chwili t (T_p jest okresem próbkowania, a $\omega_g = 2\pi f_g$ jest pulsacją odpowiadającą częstotliwości granicznej):

$$x(t) = \sum_{n=-\infty}^{\infty} x(nT_p) \frac{\sin(\omega_g(t - nT_p))}{\omega_g(t - nT_p)} \tag{1}$$

$$G(f) = \int_{-\infty}^{\infty} x(t) \exp(-i2\pi f t) dt \tag{2}$$



Rys. 3. Sygnał w dziedzinie czasu i częstotliwości
Fig. 3. The signal in time and frequency domain

Praktycznym wnioskiem z tego twierdzenia jest fakt, że ciągła funkcja opisująca sygnał mowy może być przedstawiona jako ciąg próbek wartości tej funkcji w określonych chwilach. Do odtworzenia $x(t)$ na podstawie próbek stosuje się zwykle interpolację wielomianową zamiast wzoru (1), który wymaga uwzględnienia nieskończenie wielu próbek. Powoduje to jednak pewien błąd odtworzenia sygnału, częściowo kompensowany zwiększeniem częstotliwości próbkowania ponad częstotliwość Nyquista. Dodatkowy błąd ograniczający dynamikę wynika z faktu, że w systemach cyfrowych próbki mogą być pamiętane ze skończoną dokładnością, tj. są kwantowane. Zarówno częstotliwość próbkowania, jak i liczbę poziomów kwantyzacji należy dobrać do charakteru sygnału.

Nie będziemy tu omawiać metod konwersji analogowo-cyfrowej i odwrotnej. Większość dostępnych obecnie na rynku kart muzycznych dla komputerów PC umożliwia zamianę analogowego sygnału na cyfrowy i odwrotną z parametrami przewyższającymi potrzeby akwizycji i odtwarzania mowy nawet w przypadku najprostszych kart.

Niektóre częstotliwości próbkowania są bardziej popularne od innych z różnych powodów. Pewne układy próbkujące mogą pracować tylko z określonymi częstotliwościami, wynikającymi np. z zastosowanego rozwiązania dzielnika częstotliwości [7]. Oto częściej stosowane częstotliwości prókowania.

Tabela 1

Częściej stosowane częstotliwości próbkowania

f_s [Hz]	Opis
8000	Dokładnie 8000 próbek na sekundę jest standardem telefonicznym stosowanym z kompresją dynamiki A-LAW i μ-LAW. Niektóre systemy stosują zbliżoną częstotliwość (np. NeXT: 8012.82Hz).

cd. tabeli 1

11000 lub 11025	Jedna czwarta częstotliwości próbkowania CD, połowa częstotliwości próbkowania Maca (najpopularniejsza częstotliwość próbkowania dla tego komputera).
16000	Stosowana np. przez standard kompresji G.722.
18900	Standard CD-ROM/XA.
22000 lub 22050	Połowa częstotliwości próbkowania CD, częstotliwość próbkowania Maca (historycznie: 22254.54 Hz, częst. odchylania poziomego monitora Maca).
32000	Stosowana w radiu cyfrowym, systemie dźwięku telewizyjnego NICAM (Nearly-Instantaneous Companded Audio Multiplex [IBA / BREMA / BBC]), również w zapisie taśmowym DAT ze zmniejszoną częstotliwością próbkowania i w japońskiej HDTV.
37800	Standard CD-ROM/XA dla wyższej jakości.
44056	Częstotliwość stosowana w profesjonalnej telewizji, aby uzyskać całkowitą liczbę próbek przypadającą na jedną ramkę (kadr).
44100	Częstotliwość próbkowania CD i DAT dla kopii z CD (16 bitów).
48000	Normalna częstotliwość próbkowania DAT (magnetofony cyfrowe).

5. Format plików dźwiękowych WAVE

Pliki dźwiękowe formatu WAVE (o nazwie zwykle z rozszerzeniem .wav) zawierające spróbkowany dźwięk bez dodatkowych informacji (np. wyróżnionych chwil czasowych) są stosowane m.in. w systemie MS Windows i mają następującą budowę [7]:

Nagłówek:

1. identyfikator "RIFF" — 4 bajty
2. *rLen*, długość pliku bez pól 1. i 2. — 4 bajty
3. identyfikator "WAVE" — 4 bajty

4.	identyfikator "`fmt`"	4 bajty
5.	długość pól 6..11 w bajtach (16)	4 bajty
6.	*wFormatTag* - identyfikator formatu 0x0001 - WAVE_FORMAT_PCM inne - dla ALAW, ADPCM itp.	2 bajty
7.	*wChannels* - liczba kanałów (1 lub 2)	2 bajty
8.	*dwSamplesPerSec* - częstotliwość próbkowania w Hz	4 bajty
9.	*dwAvgBytesPerSec* = #7 * #8 * #11 / 8	4 bajty
10.	*wBlockAlign* = #7 * #11 / 8	2 bajty
11.	*wBitsPerSample* - liczba bitów na próbkę	2 bajty
12.	identyfikator "`data`"	4 bajty
13.	*dLen* - liczba próbek w bajtach	4 bajty

Dane:
dla próbek 1..8 bitowych: 1 bajt, bez znaku (z przesuniętym zerem),
dla próbek 9..16 bitowych: 2 bajty - liczba całkowita w zapisie uzupełnień do dwóch,
dla (*wBitsPerSample* & 7) != 0: wyzerowane mniej znaczące bity,
dla dwóch kanałów: najpierw kanał lewy, potem prawy.
Kolejność bajtów (również w nagłówku): od najmniej do najbardziej znaczącego.

6. Implementacja metod kompresji i dekompresji

6.1. Kompresja dźwięku - wprowadzenie

Jak wspomniano wcześniej, minimalna częstotliwość próbkowania sygnału zależna jest od częstotliwości najwyższej harmonicznej tego sygnału, a liczba poziomów kwantyzacji zwykle wpływa na zniekształcenia odtwarzanego sygnału (im więcej poziomów kwantyzacji, tym mniejsze błędy). Dla sygnałów dźwiękowych wybór tych parametrów jest zwykle kompromisem między jakością zapisu a wymaganiami sprzętowymi, rosnącymi wraz z częstotliwością próbkowania i liczbą poziomów kwantyzacji. „Na szczęście" człowiek dysponuje jedynie ograniczonymi możliwościami percepcji dźwięku i możliwe jest ograniczenie pasma i dynamiki zapisywanego sygnału [13].

W telekomunikacji przyjęto, że zrozumiałą mowę można przesłać kanałem o paśmie przenoszonych częstotliwości od 300Hz do 3400Hz i dynamice ok. 50dB (przy pełnym wysterowaniu). Takie parametry można uzyskać próbkując sygnał z częstotliwością 8000 Hz i rozdzielczością 8 bitów. Nazwa Pulse Code Modulation (PCM) oznacza zapis sygnału w postaci próbek - liczb binarnych o stałej liczbie bitów.

Strumień informacji opisującej dźwięk w bezpośredni sposób zależy od częstotliwości próbkowania i liczby poziomów kwantyzacji:

$$b = f_s \cdot \log_2 k \tag{3}$$

gdzie b jest strumieniem informacji w bitach na sekundę, f_s - częstotliwością próbkowania, k - liczbą poziomów kwantyzacji. Dla zapisu stereofonicznego b ulega podwojeniu.

Tabela 2

Przykładowe wartości parametrów próbkowania i uzyskiwanej jakości

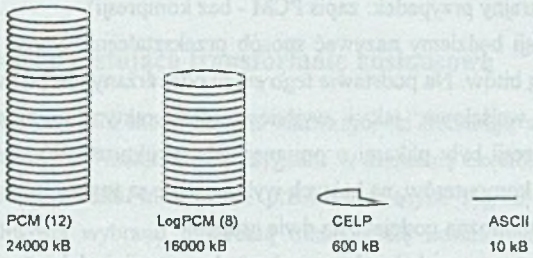
Jakość	...porównywalna z	f_s	k	$\log_2 k$	b
Bardzo wysoka	Compact Disc	44.1 kHz	65536	16	2*705 600
Wysoka	transmisja radiowa FM	30 kHz	16384	14	2*420 000
Dobra	transmisja radiowa AM	11 kHz	4096	12	132 000
Dobra (dla mowy)	transmisja telefoniczna	8 kHz	4096	12	96 000

Do transmisji mowy wystarczająca jest jakość „telefoniczna", nawet z mniejszą liczbą poziomów kwantyzacji. W telekomunikacji przyjęto rozdzielczość dwunastobitową, aby pokryć szerszy zakres dynamiki (umożliwić korzystanie z telefonu osobom ciszej i głośniej mówiącym przy porównywalnej jakości); wartości kwantowane równomiernie z rozdzielczością 12 bitów poddawane są tzw. kompresji dynamiki dając w wyniku próbki ośmiobitowe.

Przekazanie sygnału mowy wymaga więc przesłania ok. 8000 słów ośmiobitowych na każdą sekundę mowy. Zapis dziesięciominutowego odcinka mowy to prawie 5 MB informacji. Tymczasem w ciągu dziesięciu minut można wypowiedzieć około 1500 słów - 8000 znaków. Stosując nieskomplikowane metody zapisu można ten tekst zapisać w ok. 3000 bajtów. Iloraz 5 MB i 3000 bajtów (ponad 1600) wykazuje, jak wielka jest nadmiarowość sygnału mowy. Pomaga ona w zrozumieniu mowy przez człowieka w niesprzyjających warunkach, ale sprawia kłopoty w transmisji i przechowywaniu tego sygnału.

Próby kompresji dźwięku podejmowane były najwcześniej w telekomunikacji. Początkowo były to metody analogowe, np. kompresja dynamiki, później zastosowane również w technikach cyfrowych. Kompresją mowy, umożliwiającą przesłanie jej kanałem cyfrowym o małej szybkości transmisji (jakim jest np. modem podłączony do linii telefonicznej), interesuje się też sektor wojskowy ze względu na nieporównywalnie większe bezpieczeństwo szyfrowania informacji cyfrowej metodami cyfrowymi niż szyfrowania sygnału analogowego metodami analogowymi.

Opracowano różne metody różniące się stopniem komplikacji, uzyskiwanym stopniem kompresji i jakością sygnału. Oto porównanie kilku z nich na przykładzie ilości informacji potrzebnej do zapisu 10000 znaków (rys. 4).



Rys.4. Ilustracja zajętości pamięci dla różnych metod zapisu dźwięku (mowy). Jeden krążek oznacza 1000 kB
Fig. 4. Comparation of memory requirements for different methods of speech coding. One disc stands for 1000 KB

Skróty pod rysunkiem oznaczają:

PCM (Pulse Code Modulation) - zapis próbek bez przetwarzania z rozdzielczością 12 bitów,
LogPCM - zapis próbek z kompresją dynamiki (precyzja zależna od wartości próbki),
CELP (Code Excited Linear Prediction) - zapis wykorzystujący predykcję liniową,
ASCII - zapis w postaci tekstu.

Chociaż porównanie to wygląda obiecująco, i tutaj konieczny jest kompromis. Otóż wraz ze wzrostem współczynnika kompresji (ilorazu ilości informacji przed kompresją do ilości informacji po kompresji) wzrasta stopień komplikacji metody kompresji i jej czasochłonność (rys. 5) [9]:

Rys. 5. Zależność jakości od współczynnika kompresji i komplikacji kodera
Fig. 5. Dependence of quality on compression ratio and complexity of coding

„Jakość" dotyczy sygnału po kompresji i dekompresji. Szczególnie ekstremalnie małe szybkości transmisji (duże wartości współczynnika kompresji) wymagają zastosowania złożonych metod (skrajny przypadek: kompresja przez pełne rozpoznanie mowy). Przy dopuszczeniu wyższych szybkości transmisji różnice jakości sygnału uzyskanego różnymi metodami maleją (skrajny przypadek: zapis PCM - bez kompresji).

Metodą kompresji będziemy nazywać sposób przekształcenia zapisu dźwięku w postaci PCM na pewien ciąg bitów. Na podstawie tego ciągu odtwarzany jest później zapis PCM.

Zarówno dane wejściowe, jak i wyjściowe dla praktycznie zrealizowanych metod kompresji i dekompresji były plikami o opisanej dalej strukturze. Wynika to z ograniczonej mocy obliczeniowej komputerów, na których wykonywane są te programy.

Metody kompresji można podzielić na dwie grupy:

- **Metody bezstratne** - w których sygnał po kompresji i dekompresji jest identyczny z oryginalnym. Takie metody stosują na przykład popularne programy kompresji danych: PKZIP, ARJ i podobne. Niżej przedstawiona tabela 3 zawiera porównanie efektywności metod bezstratnych dla różnych rodzajów danych - dźwięk nie poddaje się łatwo takiej kompresji. W związku z istnieniem tych efektywnych programów metody bezstratne kompresji *nie przetworzonego* sygnału nie będą przedmiotem naszego zainteresowania.
- **Metody stratne** - w których sygnał wynikowy zachowuje tylko istotne cechy sygnału oryginalnego, a odpowiadające sobie próbki tych sygnałów nie muszą być identyczne. Głównym problemem jest tu stwierdzenie, które cechy sygnału są nieistotne i mogą w związku tym zostać odrzucone jako nie przenoszące informacji o sygnale.

Tabela 3

Porównanie współczynników kompresji bezstratnej dla różnych rodzajów danych

	Dł. przed kompresją	Dł. po kompresji		Skrócenie w %	
		PKZIP	ARJ	PKZIP	ARJ
ENIGMA.PAS - tekst	16849	4829	4849	71.34	71.22
ENIGMA.EXE - program	16000	8786	8837	45.10	44.77
UFO1.WAV - dźwięk	18170	14900	14958	18.00	17.68

Przedstawione zostaną trzy zaimplementowane metody kompresji:

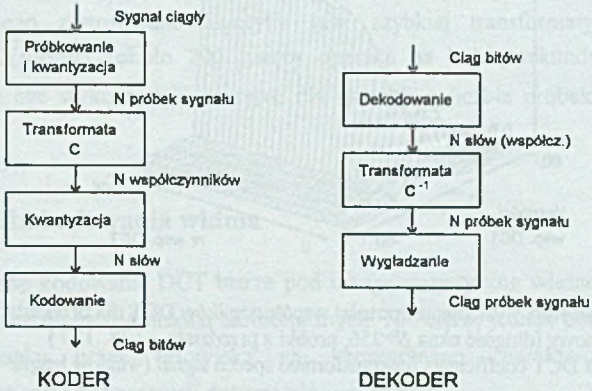
- Metoda wykorzystująca dyskretną transformatę kosinusową (DCT),
- Metoda kodowania harmonicznych,

- Metoda predykcji liniowej.

Aby możliwa była jakakolwiek kompresja (jest to ogólna uwaga, dotycząca nie tylko dźwięku), pakowane dane muszą wykazywać pewną regularność. Zadaniem kodera jest wychwycenie tej regularności i jej wykorzystanie do efektywnego zapisu danych.

6.2. Metoda wykorzystująca transformatę kosinusową

Obserwując widmo sygnału mowy można zauważyć, że dominują w nim harmoniczne o niskich częstotliwościach. Przekształcenie sygnału w dziedzinę częstotliwości pozwoli na jego kompresję. Spośród kilku transformat przekształcających sygnał z dziedziny czasu w dziedzinę częstotliwości wybrano dyskretną transformatę kosinusową. Ogólny schemat kompresji i dekompresji przedstawiony jest na rys 6.



Rys. 6. Schemat kodera/dekodera wykorzystującego transformatę
Fig. 6. Outline of transform coder and decoder

Oto definicja jednowymiarowej dyskretnej transformaty kosinusowej **C** (Discrete Cosine Transform, DCT) (Ahmed, Natarajan i Rao, 1974); **x** jest wektorem próbek sygnału, θ wektorem współczynników DCT:

$$\theta = \mathbf{C}\mathbf{x}; \quad \mathbf{x} = \mathbf{C}^{-1}\theta. \tag{4}$$

Transformata prosta:

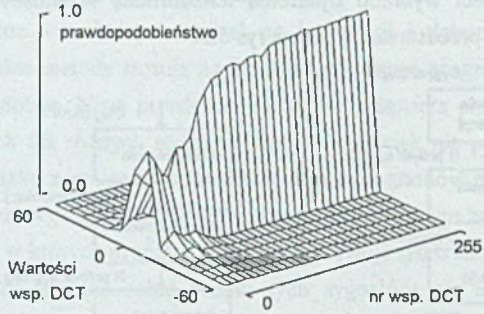
$$\theta(k) = \sqrt{\frac{2}{N}}\,\alpha(k)\sum_{n=0}^{N-1} x(n)\cos\frac{(2n+1)k\pi}{2N}; \quad k = 0, 1, \ldots, N-1. \tag{5}$$

Transformata odwrotna:

$$x(n)=\sqrt{\frac{2}{N}}\sum_{k=0}^{N-1}\alpha(k)\,\theta(k)\cos\frac{(2n+1)k\pi}{2N};\quad n=0,1,\ldots,N-1, \tag{6}$$

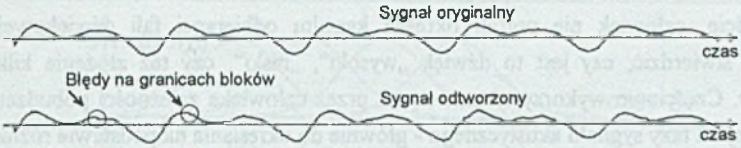
$$\alpha(0)=1/\sqrt{2},\quad \alpha(k)=1 \text{ dla } k>0.$$

KODER dokonuje podziału ciągu próbek sygnału wejściowego na fragmenty - ciągi N próbek. Każdy fragment przekształcany jest przy użyciu dyskretnej transformaty kosinusowej (DCT), dając w wyniku N rzeczywistych współczynników DCT $\theta(k)$. Następnie wyznaczana jest różnica bieżącego i poprzedniego wektora współczynników. Rozkład prawdopodobieństw wystąpienia poszczególnych wartości tych współczynników przedstawia rys. 7.



Rys. 7. Prawdopobobieństwa wystąpienia wartości współczynników DCT dla przekształconego sygnału mowy (długość okna N=256, próbki z przedziału $-128..127$)
Fig. 7. Distribution of DCT coefficients for transformed speech signal (window length = 256 samples)

Składowe różnicy wektorów poddawane są kwantyzacji i kodowane metodą Huffmana uzupełnioną o kodowanie RLE (kodowanie ciągów zer - kolejne współczynniki często są zerowe, efektywne jest kodowanie liczby zer zamiast niezależnego kodowania współczynników). Stosowane są trzy różne słowniki dla kolejnych grup współczynników. DEKODER wykonuje odwrotne operacje. Utrata informacji następuje tu przy kwantyzacji wyników DCT i objawia się niewielkimi zmianami zawartości harmonicznych w odtwarzanych odcinkach sygnału. Uciążliwe są błędy odtworzonego sygnału na granicach bloków (rys. 8).



Rys. 8. Błędy na granicach bloków
Fig. 8. Errors at blocks boundaries

Te błędy są łagodzone przez zastosowanie częściowo nakładających się bloków, na które dzielony jest sygnał wejściowy. Dyskusję błędów zamieszczono w pracy [8].

Bezpośrednia realizacja DCT (według wzoru (5)) jest bardzo czasochłonna - ma złożoność czasową $O(n^2)$ ze względu na wymiar wektora próbek (dla N=256: ponad 256^2=65636 mnożeń i tyle samo dodawań; dla f_s=11 kHz ponad 5.5 miliona operacji na każdą sekundę sygnału). Dlatego zastosowano algorytm tzw. szybkiej transformaty kosinusowej [5] o złożoności $O(n\log n)$ (około 200 tysięcy operacji na każdą sekundę sygnału). Szybki algorytm może być wykorzystywany tylko dla okienek o liczbie próbek będącej całkowitą potęgą dwójki.

6.3. Metoda kodowania widma

Przedstawione kodowanie DCT bierze pod uwagę statystyczne własności sygnału dźwiękowego - większą zawartość niskich harmonicznych. Nie bierze jednak pod uwagę charakteru percepcji dźwięku przez człowieka ani szczególnego charakteru sygnału mowy (a przedstawiane metody kompresji dotyczą głównie mowy).

Właściwymi receptorami reagującymi na drgania powietrza przetwarzane w uchu człowieka przez złożony układ mechaniczny są tzw. komórki rzęsaste z rzęskami (stereocilia), rozmieszczone na błonie podstawnej w tzw. przewodzie ślimakowym [18]. Błona ta wzdłuż kanału ślimaka zmienia swą szerokość, masę i sztywność, w wyniku czego drgania wywołują nierównomierne oscylacje poszczególnych punktów błony. Zależnie od częstotliwości drgań większemu pobudzeniu ulegają różne części błony, a przy złożonych drganiach powstaje na błonie podstawnej „mapa" pobudzeń - fale o różnych długościach wprawiają w maksymalne drgania różne jej rejony. Wspomniane wcześniej komórki rzęsaste dotykają tzw. błony nakrywkowej i poddawane są naprężeniom mechanicznym podczas drgań błony podstawnej. Naprężenia te zamieniane są na impulsy nerwowe przesyłane nerwem słuchowym do pnia mózgu. Główną częścią układu słuchowego człowieka jest więc **analizator częstotliwościowy** (o dobroci $Q \approx 200$ przy czasie analizy ok. 10 ms).

Rzeczywiście, człowiek nie potrafi określić kształtu odbieranej fali dźwiękowej, potrafi natomiast stwierdzić, czy jest to dźwięk „wysoki", „niski", czy też złożenie kilku takich dźwięków. Częściowo wykorzystywane są też przez człowieka zależności pobudzenia błony podstawnej od fazy sygnału akustycznego - głównie do określania na podstawie różnicy faz fal docierających do lewego i prawego ucha kierunku źródła dźwięku (dla niskich częstotliwości ten sposób zawodzi). Wydaje się, że efektywne może być wykorzystanie tych właściwości ludzkiego ucha zamiast (albo oprócz) właściwości samego sygnału. Metoda kodowania widma jest próbą ich wykorzystania.

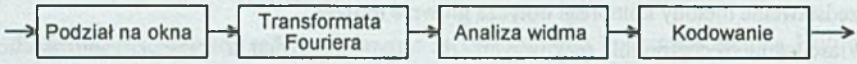
Jako informację o dźwięku wykorzystano zakodowany kształt widma częstotliwościowego. Korzystne dla celów kodowania własności słuchu to m.in.:

- mała wrażliwość na fazę sygnału,
- efekt „zagłuszania" słabych harmonicznych przez sąsiadujące z nimi harmoniczne o dużych wartościach,
- logarytmiczna wrażliwość na wartość amplitudy i częstotliwości harmonicznych.

Również sam sygnał, który ma być kodowany (mowa), ma korzystne własności:

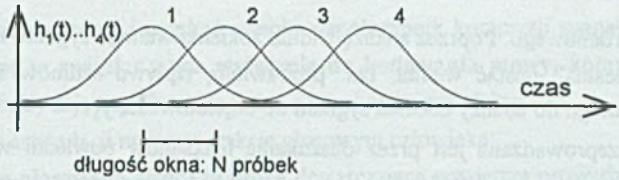
- większą zawartość niskich niż wysokich harmonicznych,
- obecność w widmie dla głosek dźwięcznych wyłącznie harmonicznych o częstotliwościach równych całkowitym wielokrotnościom częstotliwości tonu krtaniowego.

Ponieważ rozpoznanie mowy odbywa się na podstawie widma sygnału, a nie jego przebiegu w dziedzinie czasu, kodowaniu może podlegać tylko postać widma. Kodowanie przeprowadzane jest według schematu przedstawionego na rys. 9.



Rys. 9. Ogólny schemat kodowania widma
Fig. 9. General outline of spectrum coding

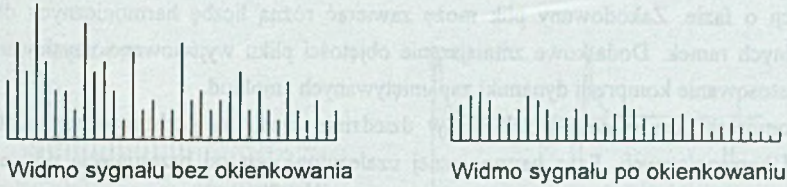
Dźwięk próbkowany jest z częstotliwością 8..11kHz (lub sprowadzany do tej częstotliwości). Okienkowanie nie może tu być prostym odrzuceniem próbek spoza okna (okno prostokątne), ponieważ taka operacja zniekształca widmo sygnału. Przyjęto okienkowanie tzw. oknem Gaussa o korzystnych własnościach (nie wprowadza tzw. listków bocznych do widma) (rys. 10). Wartości próbek mnożone są przez $h_i(t)$, gdzie i jest numerem bloku.



Rys. 10. Okienkowanie sygnału
Fig. 10. Signal windowing

Wejściowy ciąg próbek podlega podziałowi na okna o długości 512 próbek. Taki fragment sygnału poddawany jest okienkowaniu oknem Gaussa, a następnie transformacie Fouriera przy wykorzystaniu algorytmu FFT. Parametry okna Gaussa dobrano tak, aby istotne wartości próbek pochodziły z około 128 - próbkowego fragmentu z całego ciągu 512 próbek.

Uzyskane widmo przy standardowej częstotliwości próbkowania 8000 Hz ma rozdzielczość 15.625 Hz - uzyskuje się harmoniczne o częstotliwościach 0 Hz, 15.625 Hz, 31.25 Hz itd. aż do 4000 Hz. Zastosowanie okienkowania zmniejsza tę rozdzielczość, co objawia się „rozmyciem" prążków (rys. 11).



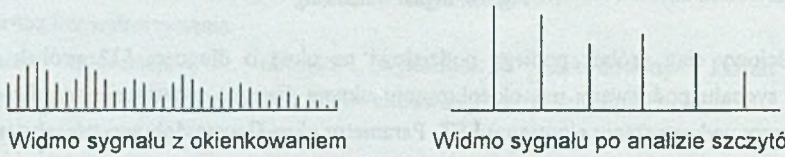
Rys. 11. Wpływ okienkowania na rozdzielczość analizy
Fig. 11. Influence of windowing on analysis resolution

Dobór parametrów okna zależy od charakteru sygnału. Metoda kompresji polega na zakodowaniu własności widmowych sygnału, a więc zakłada niezmienność widma w pewnym odcinku czasowym. Dla sygnału mowy zmiany widma powodowane są ruchami narządów artykulacyjnych, a ich szybkość ograniczona jest bezwładnością mięśni. Przyjmuje się [4], że zmiany takie występują z częstotliwością od ok. 40 Hz do 100 Hz. Okienko dla 130 próbek odpowiada częstotliwościom 62 Hz .. 85 Hz dla częstotliwości próbkowania 8000 Hz .. 11025 Hz.

Ze względu na sposób artykulacji mowy przez człowieka widmo zwykle zawiera składowe o częstotliwościach $n \cdot f_k$, gdzie f_k jest częstotliwością tonu krtaniowego (zwykle od 100 Hz dla głosów męskich do 250 Hz dla głosów żeńskich). Niezerowe wartości harmonicznych pomiędzy częstotliwościami $n \cdot f_k$ uzyskiwane przy analizie rzeczywistego sygnału dla głosek dźwięcznych wynikają z dodatkowych zakłóceń (szumy itp.) oraz faktu, że widmo nie jest niezmienne w rozpatrywanym odcinku czasu ani odcinek nie zawiera całkowitej liczby o-

kresów tonu krtaniowego. Poprzez analizę widma zokienkowanego sygnału można próbować odtworzyć „idealną" postać widma, tzn. pozbawioną wpływu szumów i niedopasowania długości wybranego do analizy odcinka sygnału do częstotliwości f_k.

Analiza przeprowadzana jest przez odszukanie maksimów obwiedni widma zokienkowanego sygnału i uwzględnienie energii pomijanych harmonicznych przez odpowiednie zwiększenie energii harmonicznych, które nie są pomijane (rys. 12).



Rys. 12. Analiza kształtu widma
Fig. 12. Spectrum shape analysis

Do celów rozpoznawania mowy (przez człowieka) wystarczy przesłanie od kilku do kilkunastu harmonicznych (w implementacji - domyślnie 10). Nie jest konieczne przesyłanie informacji o fazie. Zakodowany plik może zawierać różną liczbę harmonicznych dla poszczególnych ramek. Dodatkowe zmniejszenie objętości pliku wyjściowego uzyskiwane jest przez zastosowanie kompresji dynamiki zapamiętywanych amplitud.

Dekompresja jest syntezą przebiegu w dziedzinie czasu na podstawie zapamiętanych amplitud harmonicznych. Faza harmonicznej uzależniona jest od przesunięcia odtwarzanej ramki względem początku pliku (dla każdej harmonicznej zachowana jest ciągłość w dziedzinie czasu). Wartości amplitud harmonicznych są interpolowane na podstawie informacji z dwóch kolejnych ramek.

Najbardziej czasochłonną operacją jest tu wyznaczanie transformaty DFT podczas kodowania i IDFT przy dekodowaniu. Przekształceniu poddawany jest 512-elementowy ciąg próbek i czas tego przekształcenia jest w przybliżeniu stały. Dla okienka o długości N próbek koniecznych jest $\frac{N}{2}\log_2 N$ mnożeń zespolonych i $N\log_2 N$ dodawań zespolonych. Dla częstotliwości próbkowania 8kHz i N=512 przetworzenie jednej sekundy sygnału wymaga ponad 1.5 miliona operacji.

6.4. Kodowanie predykcyjne

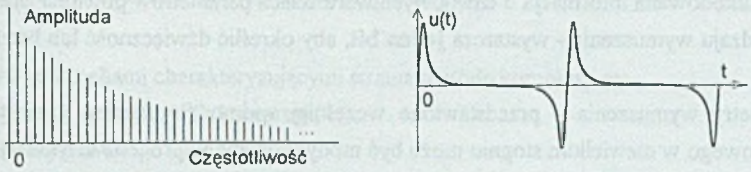
Przedstawione dwie metody kodowania uwzględniały charakter kodowanego sygnału, mało dokładnie jednak naśladowały sposób wytwarzania i percepcji mowy przez człowieka. Metoda predykcji liniowej (LPC, linear predictive coding) pozwala zbliżyć się do naturalnych

sposobów wytwarzania mowy i uzyskać wysoki współczynnik kompresji sygnału mowy. Oczywiście nadaje się w związku z tym **wyłącznie do kodowania mowy**; kompresja innych sygnałów (np. muzyki) daje zupełnie niezgodne z oczekiwaniami efekty.

Oto krótki opis artykulacji mowy w trakcie głosowym człowieka:

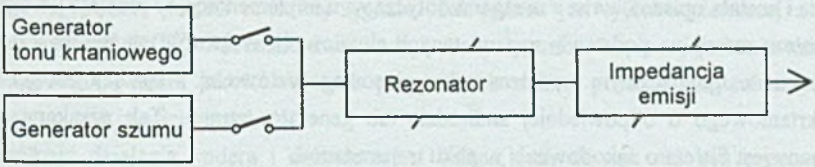
W skład **traktu głosowego** wchodzą **płuca** dostarczające powietrza prowadzonego przez **oskrzela** i **tchawicę** do **krtani**, w której drgające **struny głosowe** są źródłem **tonu krtaniowego** dla dźwięcznych fragmentów sygnału mowy. Ton ten jest następnie odpowiednio kształtowany we wnękach rezonansowych tworzonych przez **język**, **podniebienie**, **zęby** i **wargi**. W przypadku głosek nosowych istotną rolę odgrywa zamknięta jama ustna. Dla głosek bezdźwięcznych źródłem dźwięku jest nie ton krtaniowy, a szum.

Częstotliwość tonu krtaniowego zależy od płci, wieku i od cech osobniczych. Dla normalnej mowy (nie śpiewu) waha się w granicach 80..130 Hz dla mężczyzn i 150..250 Hz dla kobiet. Ton krtaniowy powstaje na skutek przetłaczania powietrza przez szparę głośni. Tak uzyskany dźwięk ma złożone widmo (rys. 13), w którym można wyróżnić nawet do 30 harmonicznych. Amplitudy harmonicznych maleją z nachyleniem ok. 12 dB na oktawę.



Rys. 13. Widmo i przebieg czasowy tonu krtaniowego (częst. podstawowa: 100..250Hz)
Fig. 13. The spectrum and waveform of laryngeal sound (base frequency: 100..250 Hz)

Ton krtaniowy (lub szum) kształtowany jest w dalszych częściach toru głosowego, który może być traktowany jako filtr o zmiennych w czasie parametrach (rys. 14).



Rys. 14. Model traktu głosowego człowieka
Fig. 14. A model of human vocal tract

W ogólnym przypadku filtr może dowolnie kształtować widmo sygnału. Dla celów generacji mowy można uwzględnić tylko wspomniane wnęki rezonansowe, uwypuklające w sygnale pewne częstotliwości, dlatego też na schemacie występuje Rezonator. Mówić można z różnymi poziomami głośności; w schemacie uwzględnia ten fakt zmienna w czasie Impedancja emisji. Parametr „głośność" może również opisywać amplitudę generatora tonu krtaniowego (lub szumu).

Sygnał mowy może więc być w pełni zdefiniowany parametrami sterującymi opisanym modelem traktu głosowego; są nimi:

- rodzaj wymuszenia (dźwięczne - ton krtaniowy lub bezdźwięczne - szum),
- parametry wymuszenia - częstotliwość, kształt fali i amplituda dla tonu krtaniowego, moc dla szumu,
- parametry filtru (rezonatora).

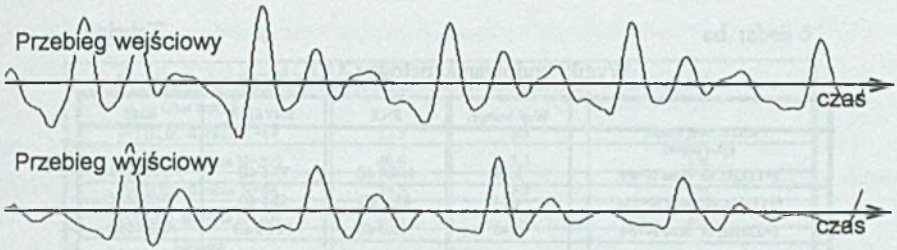
Wymienione parametry zmieniają się w czasie. Dlatego należy kodować poszczególne składowe wektora parametrów co najmniej z taką szybkością, z jaką mogą się zmieniać. Zmiany te, w przeciwieństwie do drgań strun głosowych, wynikają ze świadomych ruchów narządów mowy i nie są przez to szybkozmienne; zmiany odpowiadają częstotliwościom 50 .. 100 Hz. Zakodowana informacja o chwilowych wartościach parametrów powinna zawierać:

- dla rodzaju wymuszenia - wystarcza **jeden bit**, aby określić dźwięczność lub bezdźwięczność;
- parametry wymuszenia - przedstawione wcześniej widmo (i przebieg czasowy) tonu krtaniowego w niewielkim stopniu może być modyfikowane w procesie artykulacji mowy; dlatego przyjęto, że jedynymi istotnymi parametrami są **częstotliwość** i **amplituda** tonu krtaniowego; kształt tej fali w procesie syntezy jest stały;
- parametry filtru - filtr traktowany jako rezonator charakteryzuje się pewnymi częstotliwościami rezonansowymi, zwanymi **formantami**.

Metoda ekstrakcji wymienionych cech z przebiegu czasowego sygnału mowy jest dosyć złożona i została opisana, wraz z uwagami dotyczącymi implementacji, w pracy [8].

Synteza mowy na podstawie zapamiętanych chwilowych wektorów parametrów odbywa się zgodnie z uproszczonym modelem traktu głosowego człowieka: włączany jest generator tonu krtaniowego o odpowiedniej amlitudzie lub generator szumu. Tak uzyskany sygnał filtrowany jest filtrem o zakodowanej w pliku transmitancji.

Wyjściowy przebieg ma podobną jak wejściowy charakterystykę częstotliwościową, ale w dziedzinie czasu przebiegi wejściowy i wyjściowy mogą być przesunięte względem siebie, a w przypadku głosek szumowych - mieć zupełnie różny kształt (rys. 15).



Rys. 15. Wyniki kodowania predykcyjnego
Fig. 15. Results of predictive coding

Analiza (kompresja) sygnału jest bardzo złożona obliczeniowo - wymaga łącznie około 1.5 miliona operacji zmiennoprzecinkowych na każdą sekundę sygnału. Dekompresja - ok. 200 tysięcy operacji / sekundę.

7. Porównanie metod kompresji

Głównymi cechami charakteryzującymi stratną metodę kompresji są:

- uzyskiwany współczynnik kompresji,
- jakość uzyskiwanego sygnału,
- szybkość działania kodera i dekodera.

Współczynnik kompresji to stosunek długości pliku nie kodowanego do długości pliku zakodowanego (lub strumienia informacji nie kodowanej do strumienia informacji kodowanej).

Wskaźniki jakości takie jak stosunek sygnału do szumu (SNR) czy segmentowy stosunek sygnału do szumu (SNRSEG) opisują dobrze jakość sygnału wtedy, gdy układ koder/dekoder realizuje kodowanie postaci czasowej sygnału. Tak jednak jest tylko w pierwszej z trzech opisanych metod (DCT). Pozostałe dwie kodują cechy sygnału i błąd w dziedzinie czasu nie jest odpowiednią miarą jakości. W tych przypadkach można posłużyć się subiektywnymi metodami oceny jakości sygnału.

Szybkość działania kodera i dekodera jest ważna, jeśli dźwięk ma być przetwarzany w czasie rzeczywistym (dla mowy - co najmniej 8000 próbek na sekundę).

Oto wyniki uzyskane przy zastosowaniu metod kompresji CODCT (kodowanie DCT), COFFG1 (kodowanie widma) i PLPC (kodowanie predykcyjne) dla sygnałów o różnym charakterze.

Tabela 4

Wyniki kodowania metodą CODCT

	Wsp. kompr.	SNR	SNRSEG	RMS
Głos męski				
f=11kHz, dł. okna N=64	3.84	104.4 dB	79.6 dB	0.000128
f=11kHz, dł. okna N=256	4.44	85.7 dB	62.5 dB	0.000305
f=22kHz, dł. okna N=64	5.40	110.3 dB	104.6 dB	0.000099
f=22kHz, dł. okna N=256	9.05	92.3 dB	90.1 dB	0.000222
Głos żeński				
f=11kHz, dł. okna N=64	3.27	112.2 dB	91.3 dB	0.000137
f=11kHz, dł. okna N=256	4.85	90.1 dB	70.6 dB	0.000392
f=22kHz, dł. okna N=64	4.93	117.8 dB	103.2 dB	0.000105
f=22kHz, dł. okna N=256	7.57	99.0 dB	87.2 dB	0.000254
Muzyka				
f=11kHz, dł. okna N=64	1.42	103.4 dB	92.5 dB	0.000186
f=11kHz, dł. okna N=256	2.16	76.6 dB	72.4 dB	0.000667
f=22kHz, dł. okna N=64	2.07	104.4 dB	92.3 dB	0.000178
f=22kHz, dł. okna N=256	3.17	80.9 dB	74.9 dB	0.000544

Tabela 5

Wyniki kodowania metodą COFFG1

	Wsp. kompr.	Jakość (1..5)	Uwagi
Głos męski			
f=11kHz	9.5	3.5	mowa zrozumiała
f=22kHz	15.4	3.5	j.w.
Głos żeński			
f=11kHz	11.3	3.5	mowa zrozumiała
f=22kHz	12.1	3.5	j.w.
Muzyka			
f=11kHz	5.6	2.5	jakość
f=22kHz	5.6	2.5	niezadowalająca

Tabela 6

Wyniki kodowania metodą PLPC

	Wsp. kompr.	Jakość (1..5)	Uwagi
Głos męski			
f=11kHz, dł. okna N=64	12.1	4	mowa zrozumiała
f=11kHz, dł. okna N=200	39.1	3.5	j.w.
f=22kHz, dł. okna N=64	24.2	4	j.w.

cd. tabeli 6

f=22kHz, dł. okna N=200	78.2	3.5	j.w.
Głos żeński f=11kHz, dł. okna N=64	12.3	3.5	mowa zrozumiała
f=11kHz, dł. okna N=200	40.6	3.5	j.w.
f=22kHz, dł. okna N=64	24.5	3.5	j.w.
f=22kHz, dł. okna N=200	81.1	3.5	j.w.
Muzyka f=11kHz, dł. okna N=64	11.8	1	Sygnał nie przypomina oryginału
f=11kHz, dł. okna N=200	37.6	1	
f=22kHz, dł. okna N=64	23.9	1	
f=22kHz, dł. okna N=200	76.2	1	

Koder PLPC wykorzystuje stałą wewnętrzną częstotliwość próbkowania i dlatego długość wynikowego pliku zależy od czasu trwania dźwięku z pliku wejściowego, a nie od liczby próbek - dla dwukrotnie wyższej częstotliwości próbkowania kompresja jest dwukrotnie wyższa. Mowa jest zrozumiała - w przeciwieństwie do muzyki, której bardzo trudno domyślić się w odtworzonym sygnale.

Szybkość kodowania i dekodowania zależy oczywiście od zastosowanego komputera. Można przyjąć, że przy użyciu procesora i486DX 50MHz w ciągu sekundy wykonywanych jest ok. 250 tysięcy operacji elementarnych (tzn. operacji na próbkach łącznie z przygotowaniem danych i zapamiętaniem wyników). W czasie rzeczywistym mogą więc działać algorytmy CODCT (kompresja i dekompresja) oraz PLPC (dekompresja). Pozostałe wymagają kilkakrotnie szybszego procesora lub muszą być przetwarzane „wsadowo".

8. Podsumowanie

Kompresja sygnału akustycznego przeprowadzana była trzema metodami:

CODCT - kompresja wykorzystująca transformatę kosinusową dobrze nadaje się do każdego rodzaju dźwięku. Odtwarza przebieg oryginalnego sygnału w dziedzinie czasu. Współczynnik kompresji zwiększa się wraz ze wzrostem długości okna *N*. Niestety równocześnie rośnie błąd odtworzenia sygnału. Uzyskiwane współczynniki kompresji wynoszą od 3 do 12 (dla okna o długości 512 próbek), przy czym przy maksymalnej kompresji mimo błędów w dziedzinie czasu mowa jest doskonale zrozumiała.

COFFG1 - jako metoda kodująca cechy nie może być oceniana standardowymi wskaźnikami jakości. Uzyskiwane współczynniki kompresji zależnie od charakteru sygnału wahają się od 4 do 12. Błędy objawiają się w postaci nieco zmienionego głosu, mowa jest zrozumiała.

PLPC - jest również metodą kodującą cechy. Pozwala na uzyskanie maksymalnej kompresji spośród przedstawionych metod, sięgającej nawet 35. Błędy mogą dotyczyć nieprawidłowego rozpoznania głoski dźwięcznej i bezdźwięcznej.

Wydaje się, że do kodowania mowy z wysoką jakością najodpowiedniejsza jest metoda CODCT. Niezależnie od charakteru sygnału (mówcy) uzyskiwany jest niewielki błąd, zależny od przyjętej długości okna. Metoda może być zaimplementowana jako działająca w czasie rzeczywistym. Dodatkową zaletą jest możliwość kodowania nie tylko mowy, ale dowolnych dźwięków.

Zastosowanie metody widmowej i predykcyjnej ogranicza się do kompresji sygnału mowy ze względu na naśladowanie sposobu generacji dźwięku przez człowieka. Rekompensatą jest uzyskiwany większy współczynnik kompresji niż w CODCT.

Najskuteczniejsza metoda (PLPC) daje kompresję do ok. 300 bajtów na sekundę mowy. Dalsze zmniejszanie strumienia informacji opisującej mowę przy zachowaniu jej zrozumiałości możliwe jest przez wykorzystanie bardziej czasochłonnych algorytmów kodowania współczynników (np. kwantyzacji wektorowej), a granicą kompresji jest zastosowanie pełnego rozpoznawania mowy i kodowania słownikowego, co powinno dać efekty w postaci strumienia informacji rzędu kilku bajtów na sekundę mowy (pięć lat sygnału na dysku CD-ROM).

LITERATURA

[1] Basztura C.: Źródła, sygnały i obrazy akustyczne. Przetwarzanie, analiza, rozpoznawanie. WKiŁ, Warszawa 1988.

[2] Basztura C.: Rozmawiać z komputerem. Wydawnictwo Prac Naukowych „FORMAT", Wrocław 1992.

[3] Beauchamp K. G.: Przetwarzanie sygnałów metodami analogowymi i cyfrowymi. WNT, Warszawa 1978.

[4] Borodziewicz W., Jaszczak K.: Cyfrowe przetwarzanie sygnałów. WNT, Warszawa 1987.

[5] Chen W. S., Smith C. H., Fralick S. C.: A fast computational algorithm for the discrete cosine transform. IEEE Trans. Commun., 1977, vol. COM-25, s. 1004-1009.

[6] Dr. Dobb's Multimedia Sourcebook, Winter 1994 (dodatek do miesięcznika).

[7] Electronic Communal Temporal Lobe - lista dyskusyjna dotycząca problemów syntezy, analizy, kompresji, przetwarzania i transmisji mowy. LISTSERV: ectl@snowhite.cis.uoguelph.ca.

[8] Fabian P.: Pakiet wybranych programów kompresji mowy i karta przetworników A/C, C/A. Praca magisterska, Politechnika Śląska, Gliwice 1994.

[9] Jayant N. S., Noll P.: DIGITAL CODING OF WAVEFORMS - Principles and Applications to Speech and Video. PRENTICE-HALL, INC. Englewood Cliffs, New Jersey 1984.

[10] Kleijn W. B., Krasinski D. J., Ketchum R. H.: Fast Methods for the CELP Speech Coding Algorithm. IEEE ASSP-38, No. 8, August 1990, s. 1330-1342.

[11] Kulka Z., Libura A., Nadachowski M.: Przetworniki analogowo - cyfrowe i cyfrowo - analogowe. WKiŁ, Warszawa 1987.

[12] Mała Encyklopedia Muzyki. Praca zbiorowa, PWN, Warszawa 1981.

[13] Nelson M.: The Data Compression Book. M&T Books.

[14] Okubo K., Okumura H.: Data Compression Algorithms of LARC and LHarc. Artykuł z comp.binaries.ibm.pc.

[15] Skarbek W.: Metody reprezentacji obrazów cyfrowych. Akademicka Oficyna Wydawnicza PLJ, Warszawa 1993.

[16] Skowronek M.: Wykłady z modelowania cyfrowego układów dynamicznych (notatki - 1993).

[17] Tadeusiewicz R.: Sygnał mowy. WKiŁ, Warszawa 1988.

[18] Tramain T. E.: The Government Standard Linear Predictive Coding Algorithm: LPC-10. Speech Technology, April 1982, s. 40-49.

Recenzent: Dr inż. Ryszard Winiarczyk

Wpłynęło do Redakcji 9 czerwca 1995 r.

Abstract

The paper presents some problems related to reducing size of digital data streams describing speech signals. The speech signal is very memory deserving when stored as raw data without additional coding (e.g. as WAVE files described in chapter 5). Human ear and brain don't need to have exact waveform to recognise voices. Coding may decrease the amount of memory needed to store speech data. Three methods of coding are presented. The Discrete Cosine Transform Coding combines DCT and entropy coding (chapter 6.2). This method is applicable for all kinds of voice, giving best results for speech. The next method, Spectrum Coding (chapter 6.3), stores only the shape of input signal spectrum. Only few harmonics are enough for medium-quality speech. The third method, Linear Predictive Coding (chapter 6.4), is the most complex. It simulates human articulation system (larynx, vocal tract) and can be used for human speech only, but gives very high compression ratio. Problems with base frequency extraction are discussed. Each method is discussed - especially implementation problems. Finally, most important features of these three methods are compared and future applications are mentioned.