KODOWANIE I KOMPRESJA SYGNAŁU MOWY

Katedra Systemów Multimedialnych, Politechnika Gdańska
Autor: Grzegorz Szwoch, kwiecień 2016

Potrzeba kompresji mowy

Cyfrowy sygnał mowy bez kompresji:

- duża przepływność bitowa,
- czas transmisji – powstawanie opóźnień.

Kompresja mowy:

- zmniejszenie przepływności (mniejsze opóźnienia z powodu transmisji), ale:
- wprowadzenie opóźnień związanych z:
  - kodowaniem i dekodowaniem,
  - buforowaniem próbek (przetwarzanie blokowe).

Kodowanie i kompresja

- Kompresja – zmniejszenie liczby danych potrzebnych do zapisania istotnych informacji. Może być bezstratna lub (częściej w przypadku mowy) stratna.
- Kodowanie – zapis potrzebnych informacji za pomocą liczb (bitów) przesyłanych w strumieniu wyjściowym.

Kodek = koder + dekoder

- koder: kompresja, kodowanie
- dekoder: dekodowanie, dekompresja

Pojęcia podstawowe

- Algorytm kodowania – sposób kompresji i kodowania sygnału, np. LPC-10, CELP
- Standard kodowania – dokument opisujący praktyczną implementację algorytmu, opublikowany w celu zapewnienia kompatybilności, np. G.729
- Kodek – oprogramowanie i ew. sprzęt, implementujący algorytm kodowania (może opierać się na standardzie, ale nie musi), np. kodek Speex

Kodeki mowy

Uniwersalne kodeki sygnału nie sprawdzają się w przypadku mowy:

- zbyt duża przepływność,
- jeżeli ograniczymy przepływność: zbyt słaba jakość (zniekształcenia kompresji).

Kodeki mowy są dostosowane do specyficznych właściwości sygnału mowy:

- wąskie pasmo,
- fragmenty dźwięczne i bezdźwięczne,
- formantowość części dźwięcznych,
- entropia sygnału mowy

Zastosowania kodeków mowy

- Telefonia mobilna
- Telefonia internetowa (VoIP)
- Komunikatory internetowe
- Telekonferencje
- Rejestratory mowy
- Archiwizacja nagrań mowy

Typy kodeków mowy

Kodeki sygnałowe (*waveform*)

- operują na próbkach sygnału
- zachowują sygnał dźwiękowy
- wysokie przepływności (mała kompresja)

Kodeki parametryczne (źródłowe)

- kodowanie: analiza sygnału, wyodrębnienie parametrów
- transmisja parametrów sygnału i modelu
- dekodowanie: synteza mowy na podstawie otrzymanych parametrów
- niższe przepływności (większa kompresja)

Kodowanie sygnałowe

Kodeki sygnałowe wykorzystują:

- nieliniowe kodowanie
  - rozkład amplitud sygnału mowy skupia się w zakresie małych wartości,
  - potrzebna większa rozdzielczość dla zakresu małych wartości
- ADPCM: kodowanie różnic między próbkami (usunięcie korelacji, zmniejszenie entropii)
  - kodowany jest błąd predykcji
  - małe wartości – wydajne kodowanie
  - adaptacyjny dobór sposobu kwantyzacji

Nieliniowe kodowanie

- Sygnał mowy ma duży zakres dynamiki
- Przy liniowej kwantyzacji, ciche fragmenty mowy „giną” w szumie
- Stosunek między amplitudą sygnału a odbieraną przez człowieka głośnością jest logarytmiczny (spełnia prawo Webera-Fechnera)
- Nieliniowe kodowanie sygnału mowy pozwala na jego bardziej skuteczne kodowanie

Standardy nieliniowego kodowania

Kompresja (w sensie kompresji dynamiki)
- nieliniowe przekształcenie liniowych wartości amplitudy na wartości zakodowane.

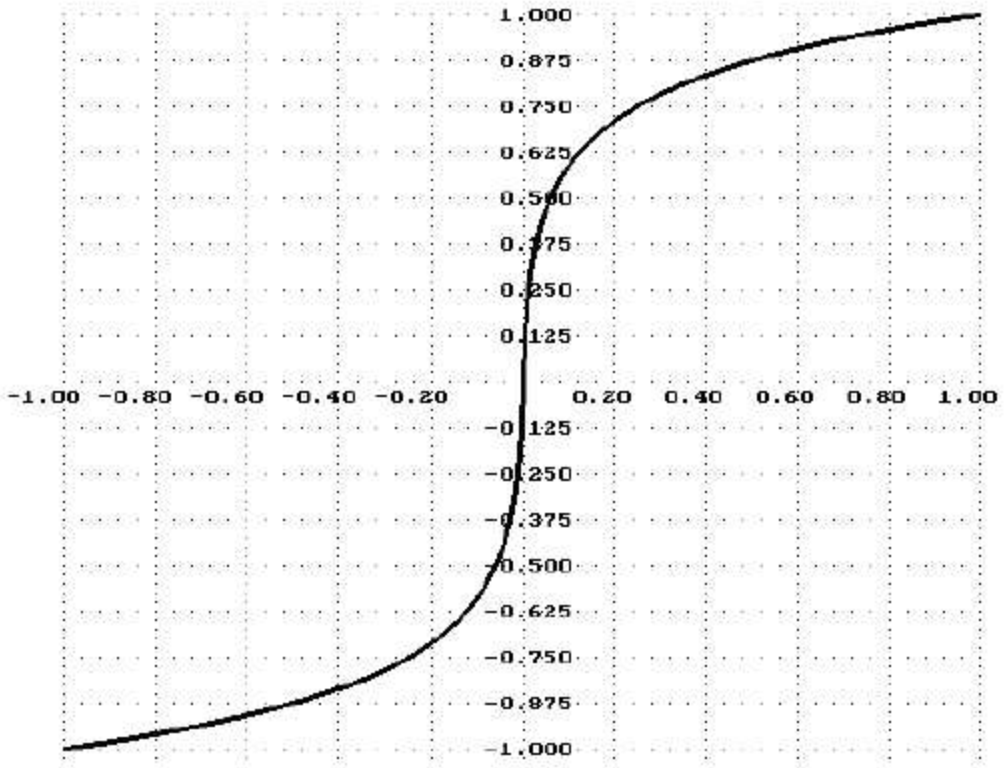
Dwa standardy kodowania:

- μ-law (μ czytamy [miu]) – starszy standard, stosowany w USA i Japonii
- A-law: nowszy standard, stosowany w Europie.

Przy połączeniach międzynarodowych, jeżeli jedna strona używa A-law, druga też musi to zrobić.

Standard μ-law

$$F(x) = \operatorname{sgn}(x) \cdot \frac{\ln(1+\mu \cdot |x|)}{\ln(1+\mu)} \qquad dla \quad -1 \le x \le 1$$



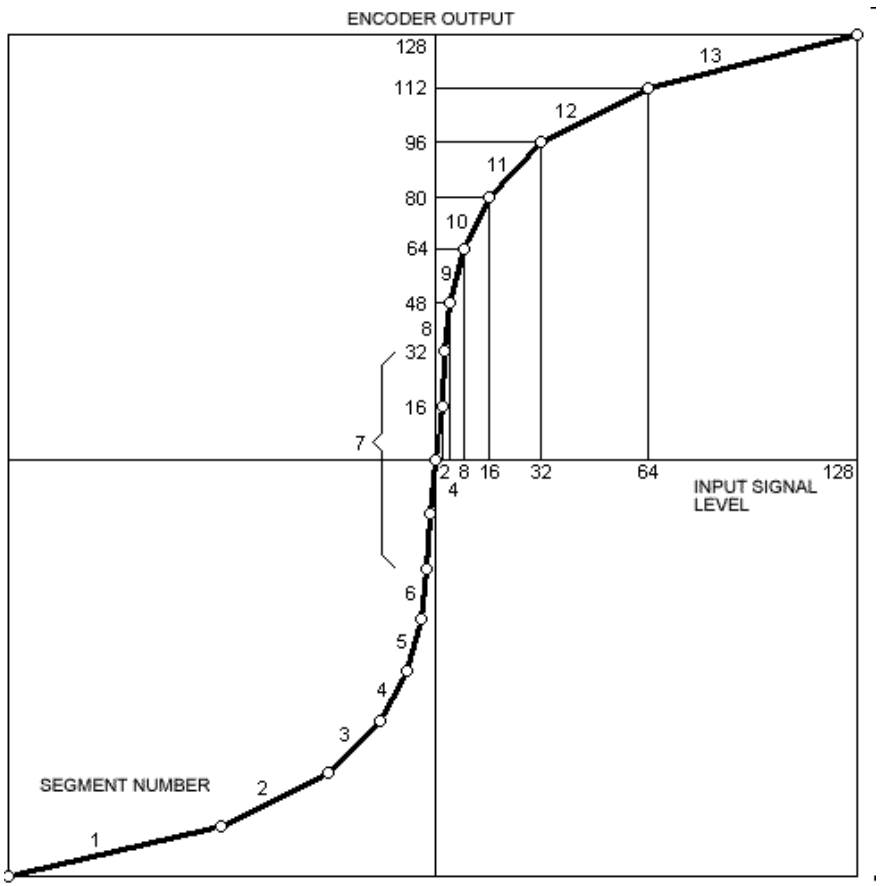
μ = 255

Standard A-law

$$F(x) = \operatorname{sgn}(x) \cdot \frac{A \cdot |x|}{1 + \ln(A)} \qquad dla \quad \frac{1}{A} \le x \le 1 \quad oraz \quad -1 \le x \le -\frac{1}{A}$$

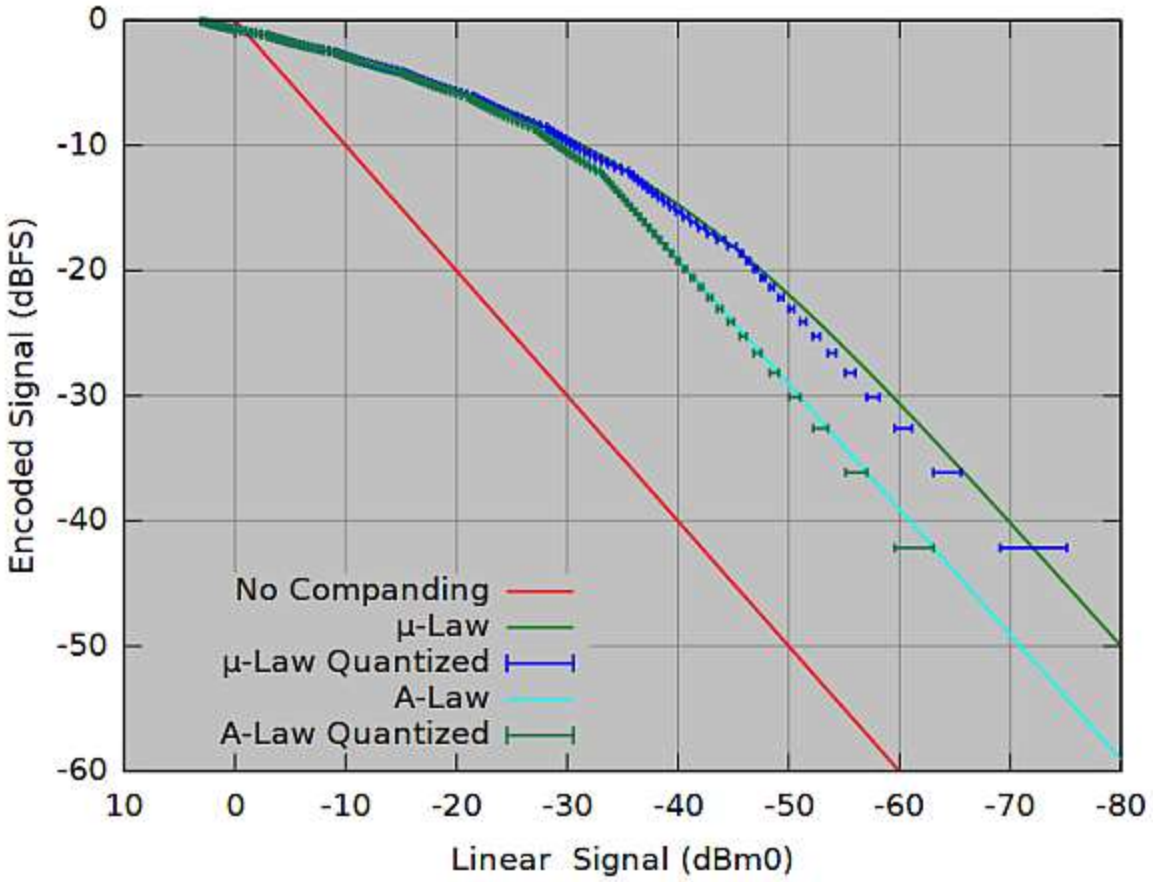
$$F(x) = \operatorname{sgn}(x) \cdot \frac{1 + \ln(A \cdot |x|)}{1 + \ln(A)} \qquad dla \quad -\frac{1}{A} \le x \le \frac{1}{A}$$

A = 87,5



Porównanie standardów

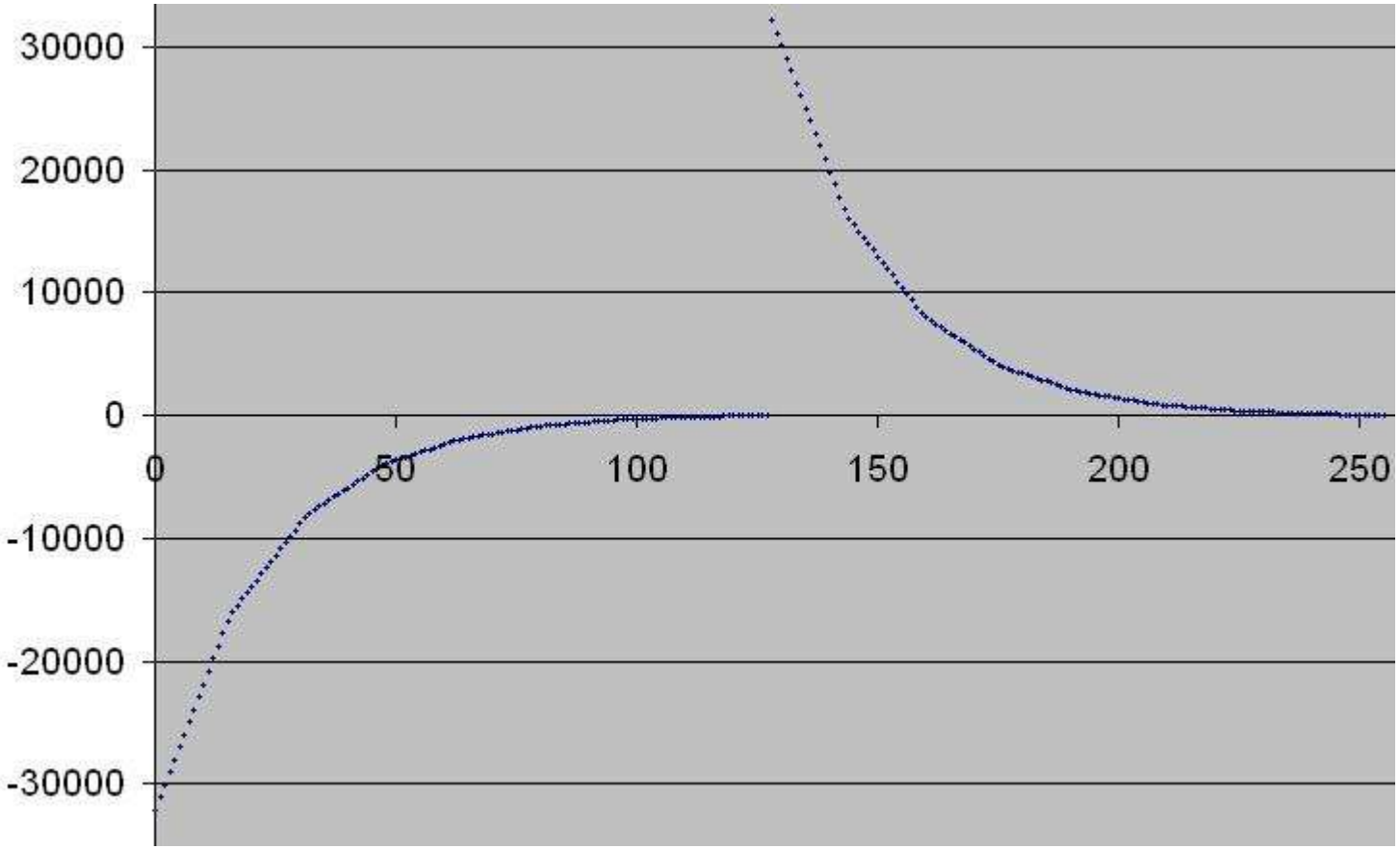
Wykres poziomu sygnału wejście-wyjście



Źródło: Wikipedia

Ilustracja kodowania nieliniowego

Wartości liniowe (oś pionowa) i wartości zakodowane w μ-law (oś pozioma)



Źródło: Wikipedia

Zysk z nieliniowego kodowania

Korzyści:

- zwiększenie dynamiki sygnału o ok. 24 dB, dzięki zwiększeniu precyzji kodowania dla małych amplitud,
- dzięki temu jakość sygnału zakodowanego w μ-law lub A-law na 8 bitach odpowiada jakości sygnału zakodowanego liniowo na 12 bitach
- a zatem przy tej samej dostępnej przepływności mamy lepszą jakość mowy niż przy kodowaniu liniowym

Kodowanie różnicowe (DPCM)

W sygnale mowy dominują składowe o niskich częstotliwościach – małe zmiany amplitudy sąsiednich próbek.

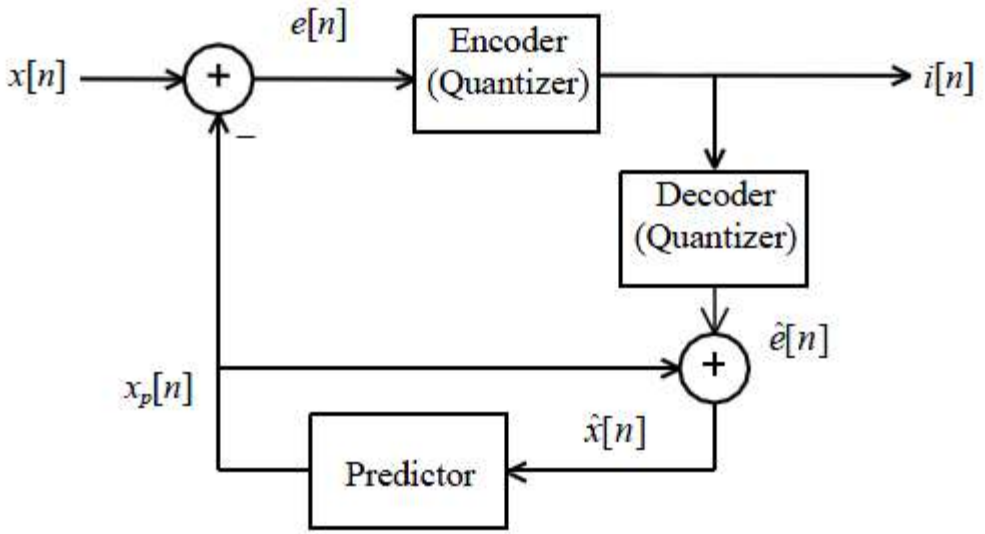
Kodowanie różnicowe DPCM – dwie metody:

- kodowanie (zwykle nieliniowe) różnicy między bieżącą a poprzednią wartością sygnału;
- kodowanie różnicy między rzeczywistą wartością próbki a wartością estymowaną (przewidywaną) za pomocą pewnego modelu

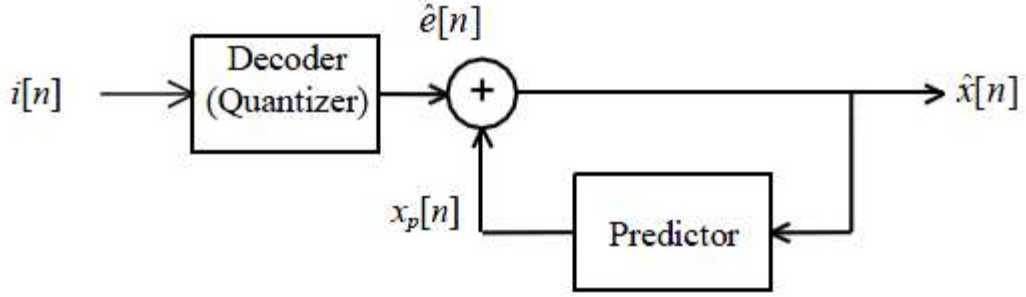
Zalety: mniej bitów potrzebnych do zakodowania, zysk dynamiki 4 – 11 dB.

DPCM

Kodowanie



Dekodowanie



ADPCM (adaptacyjne kodowanie różnicowe)

ADPCM – *adaptive differential pulse modulation coding*

- Najczęściej używana metoda w kodekach sygnałowych. Rozwinięcie DPCM.
- Błąd predykcji jest normalizowany (skalowany)
- Kodowanie dostosowuje się do zmiennych właściwości sygnału
- Dalsze zmniejszenie przepływności bitowej i poprawa dynamiki (SNR)

Kodeki sygnałowe

- G.711 – PCM, nieliniowe kodowanie,
  8 kHz, 8 bit $\Rightarrow$ przepł. 64 kbit/s
  (słowa 14-bitowe kodowane na 8 bitach)
- G726 – ADPCM z nieliniowym kodowaniem,
  tryby: 16, 24, 32, 40 kbit/s
- G722 – sub-band ADPCM, osobne
  przetwarzanie w pasmach częstotliwości,
  przepł. 48, 56, 64 kbit/s

Kodowanie parametryczne

Wytwarzanie sygnału mowy:

- pobudzenie
  - ton krtaniowy (elementy dźwięczne)
  - szum (elementy bezdźwięczne)
- trakt głosowy
  - filtracja pobudzenia

Kodowanie parametryczne – kodujemy:

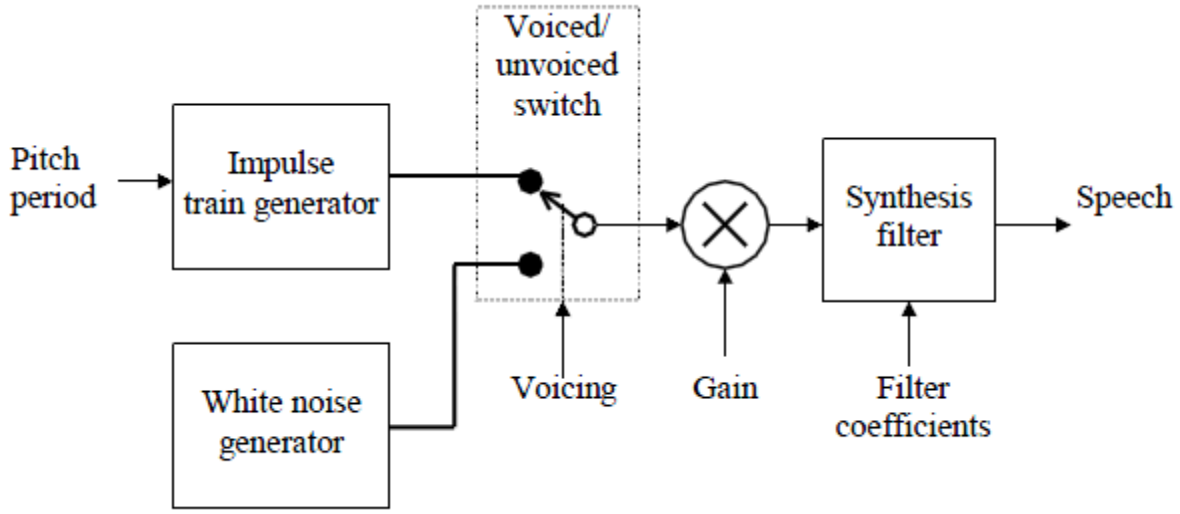
- typ pobudzenia
- parametry modelu traktu głosowego

Predykcja liniowa (LPC)

- Niemal wszystkie kodeki parametryczne wykorzystują liniowe kodowanie predykcyjne LPC (*linear predictive coding*)
- filtr LPC modeluje trakt głosowy dla bloku próbek sygnału mowy
- parametry filtru są kodowane i przesyłane dla każdego bloku (ramki) próbek
- resynteza za pomocą filtru LPC:

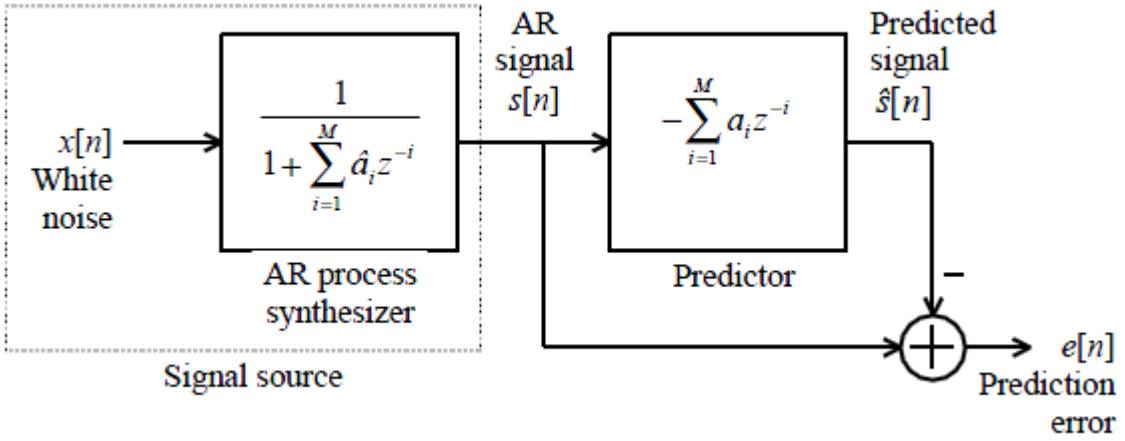
$$H(z) = \frac{1}{1 + \sum_{i=1}^{M} a_i z^{-i}}$$

Model LPC wytwarzania mowy

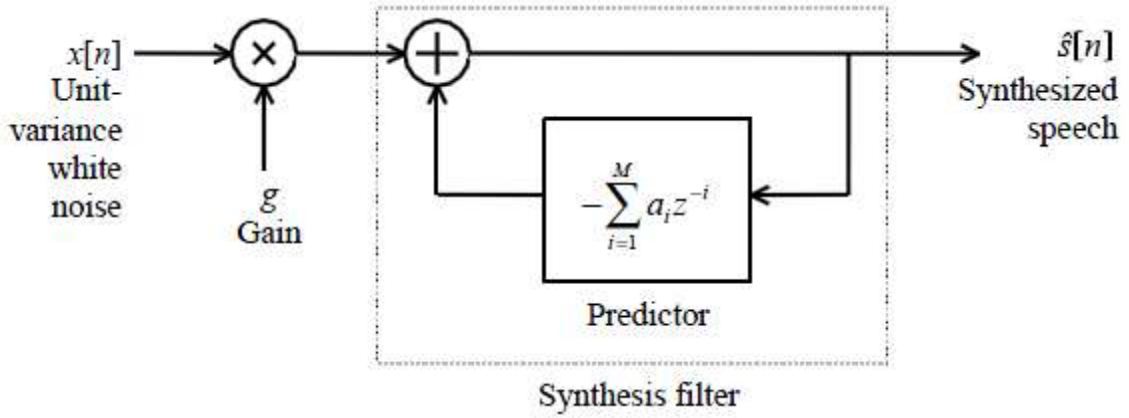


Predykcja liniowa (LPC)

Predykcja



Synteza

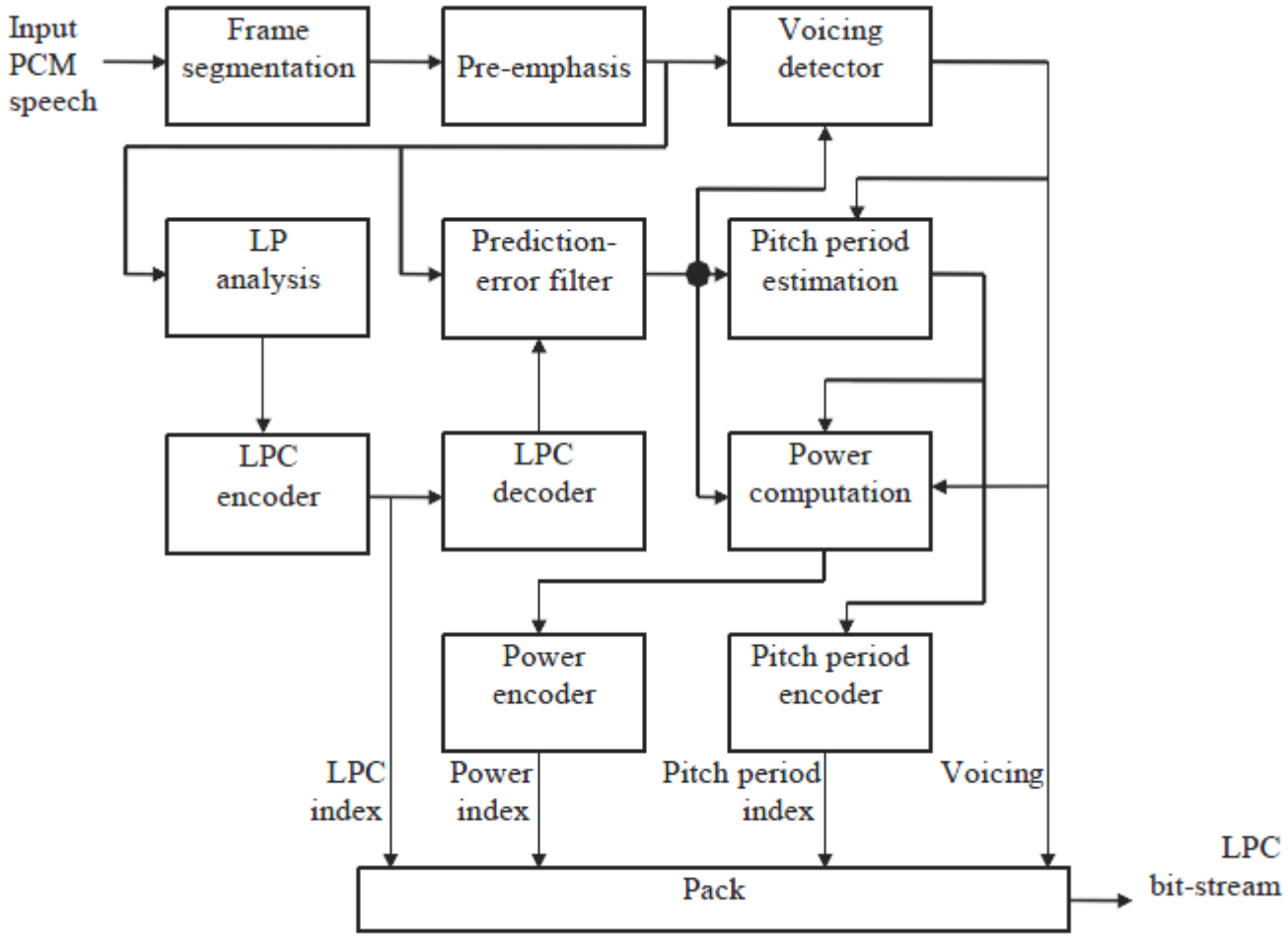


Kodek FS-1015 (LPC-10)

Kodowanie:

- sygnał 8 kHz, 12 bit/próbkę
- podział na ramki po 180 próbek (22,5 ms)
- preemfaza – podbicie wysokich częstotliwości
- klasyfikacja ramek
  - dźwięczne: LPC 10. rzędu (LPC-10)
  - bezdźwięczne: LPC 4. rzędu
- wyznaczenie współczynników predykcji
- kwantyzacja współczynników
- wyznaczenie wysokości dźwięku
- kodowanie, wysłanie

Koder FS-1015



Koder FS-1015

Zakodowany strumień – dla jednej ramki:

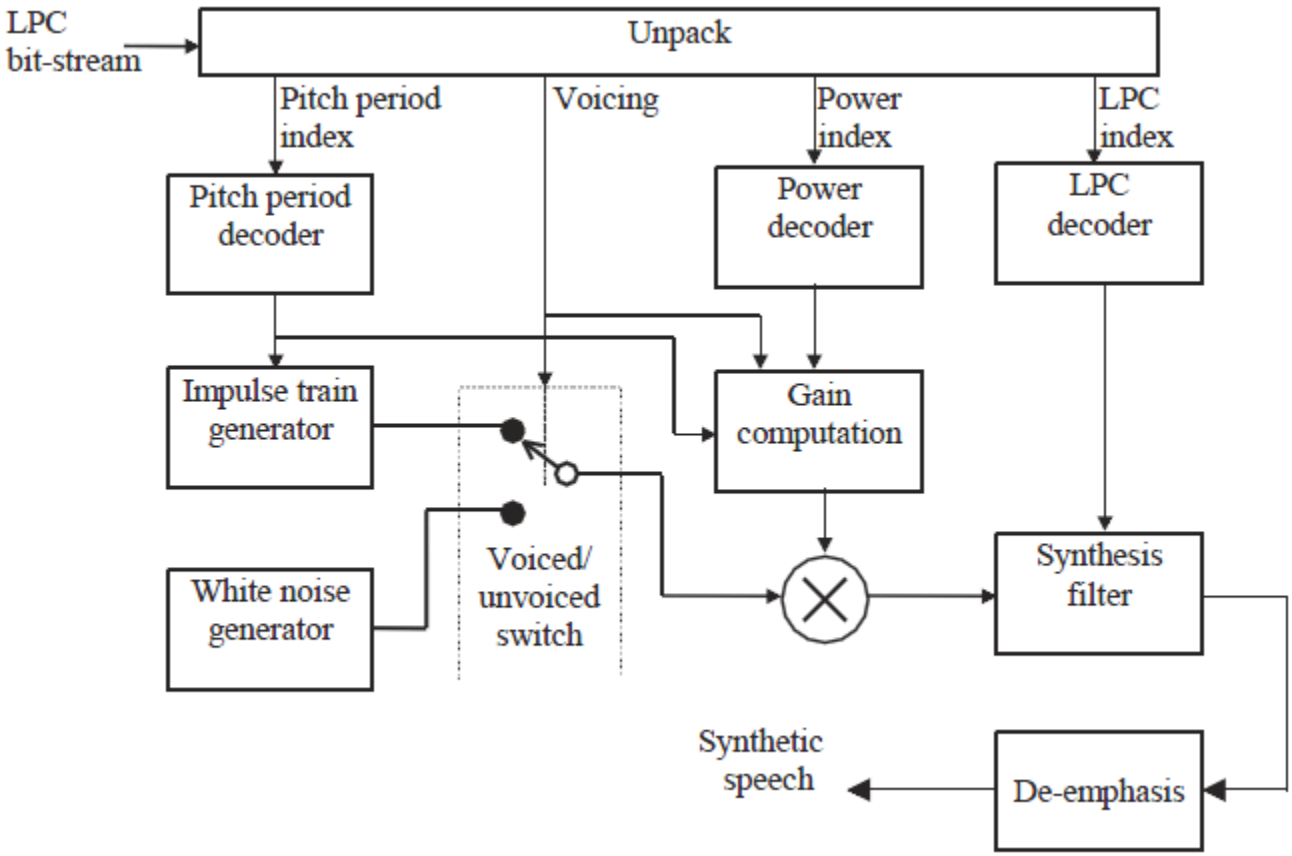
- 7 bitów: rodzaj (dźwięczny/bezdźwięczny) oraz wysokość dźwięku
- 5 bitów: wzmocnienie sygnału
- 41 bitów (dźwięczne) lub 20 bitów (bezdźw.): skwantowane współczynniki LPC
- 1 bit: synchronizacja
- 21 bitów (tylko bezdźwięczne): kodowanie protekcyjne.

Razem: 54 bity na jedną ramkę (180 próbek).

Bez kodowania: 180*12 = 2160 bitów

Przepływność kodeka: 2,4 kbit/s

Dekoder FS-1015



Wady kodeka LPC-10

- Ramki nie są zwykle czysto dźwięczne lub bezdźwięczne
- Ciąg impulsów nie symuluje odpowiednio dokładnie pobudzenia dla ramek dźwięcznych
- Utrata informacji o fazie
- Efekt działania kodeka:
  - słaba zrozumiałość mowy
  - bardzo słaba jakość mowy
  - duży poziom szumu
  - nienaturalna (syntetyczna) mowa

CELP

CELP – *Code-Excited Linear Prediction*

Rozwinięcie idei kodowania LPC

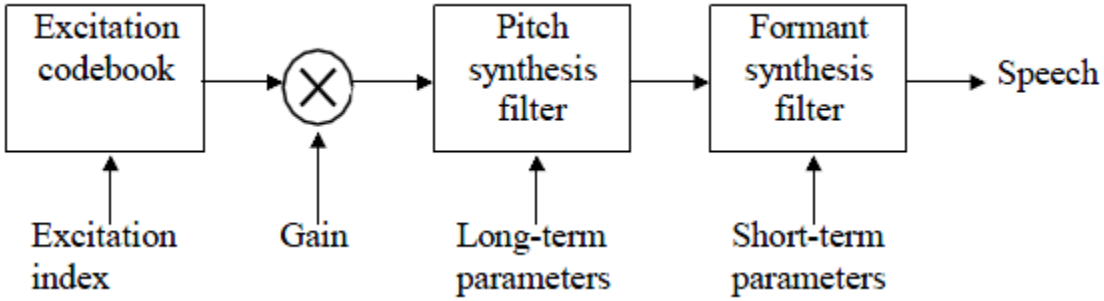
- Zamiast dwóch typów pobudzeń: zbiór („książka kodowa”) reprezentatywnych pobudzeń.
- Wybierane jest pobudzenie dające najmniejszy błąd predykcji.
- Kod wybranego pobudzenia jest przesyłany.
- Dekoder odtwarza sygnał.

CELP jest podstawą praktycznie wszystkich używanych obecnie kodeków parametrycznych.

„Czysty” CELP: kodek FS-1016; 4,8 kbit/s

CELP

Model wytwarzania mowy wg CELP

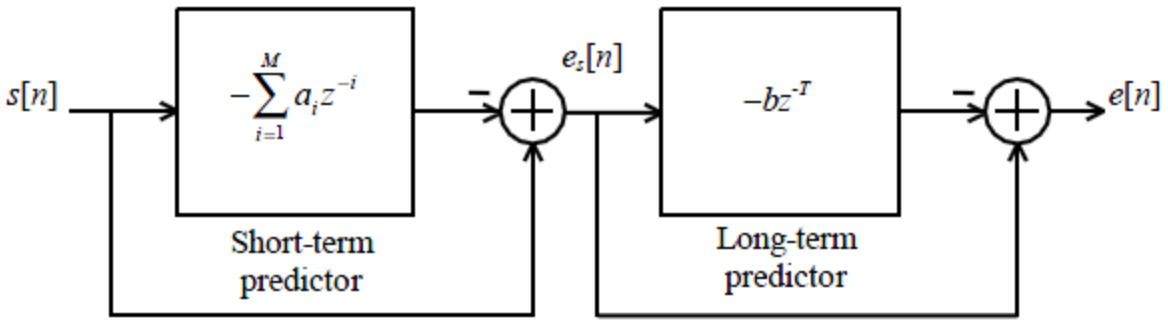


LPC short-term i long-term

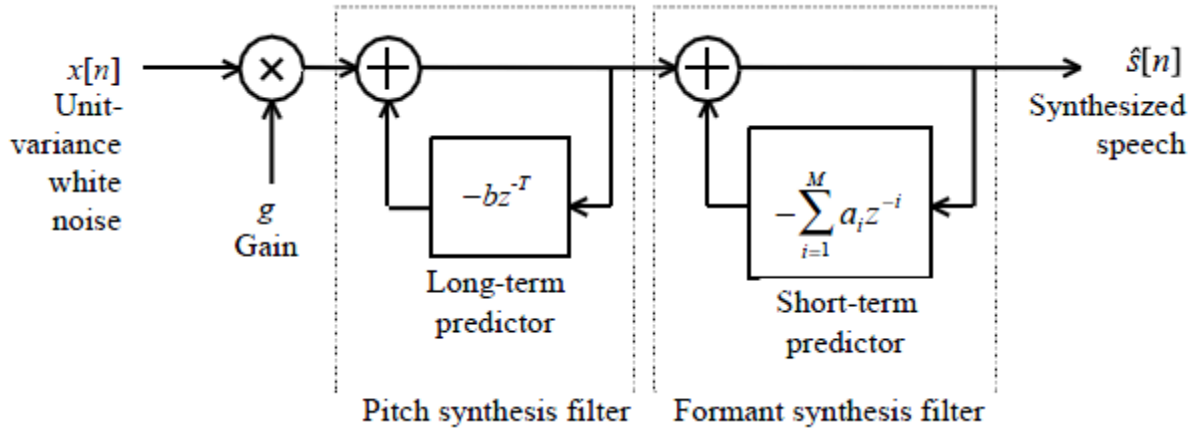
- Predykcja w krótkich ramkach próbek (*short-term LPC*)
  - pozwala na estymację kształtu widma (formantów),
  - ale nie można wyznaczyć wysokości sygnału mowy
- Dodatkowa predykcja długoterminowa (*long-term LPC*)
  - obejmuje co najmniej jeden okres
  - pozwala wyznaczyć wysokość

LPC short-term i long-term

Predykcja

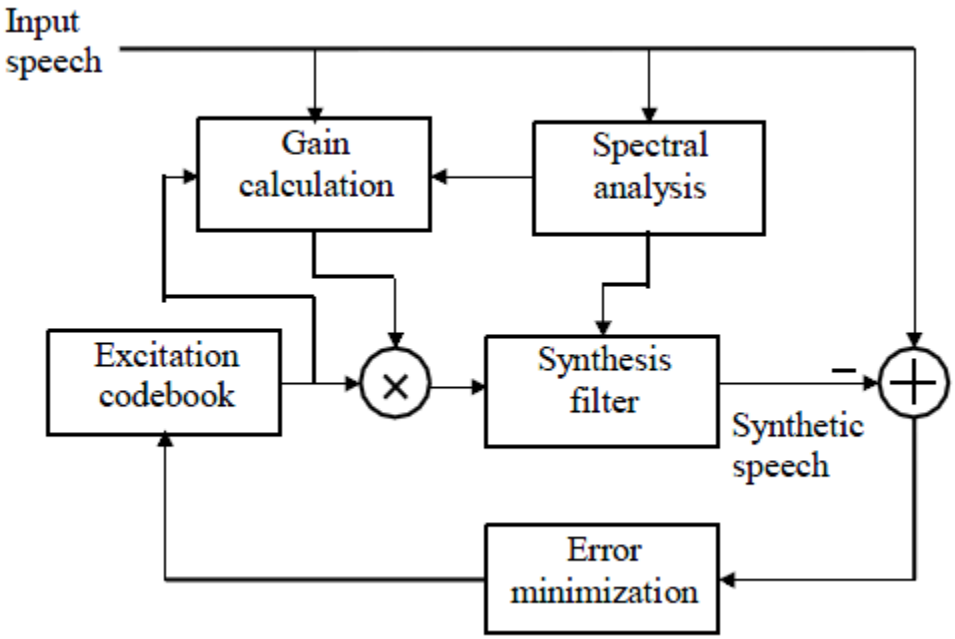


Synteza



Wybór pobudzenia

Wybór pobudzenia z książki kodowej (*codebook*) odbywa się w pętli zamkniętej, na zasadzie „analizy przez syntezę”.



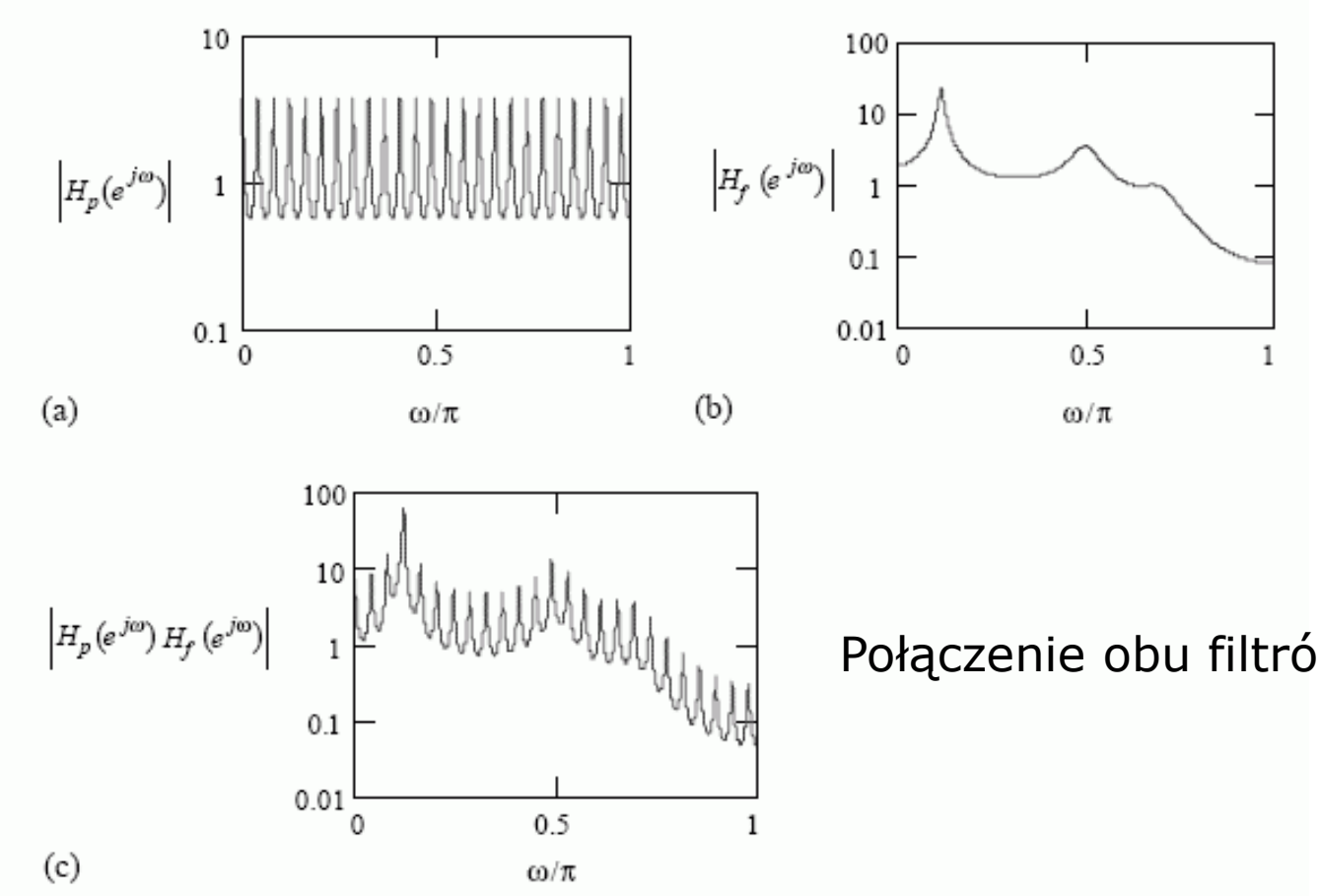
CELP – wybór pobudzenia

- Filtracja ramki próbek filtrem ważącym perceptualnie (zjawiska maskowania)
- Dla każdego słowa kodowego w książce:
  - wyznaczenie optymalnego wzmocnienia
  - synteza sygnału za pomocą filtrów LPC (wysokości dźwięku i formantowego)
  - wyznaczenie błędu predykcji
- Wybór pobudzenia, dla którego błąd predykcji jest najmniejszy.

CELP – wybór pobudzenia

Filtr wysokości
(*pitch filter*)

Filtr formantowy
(*formant filter*)

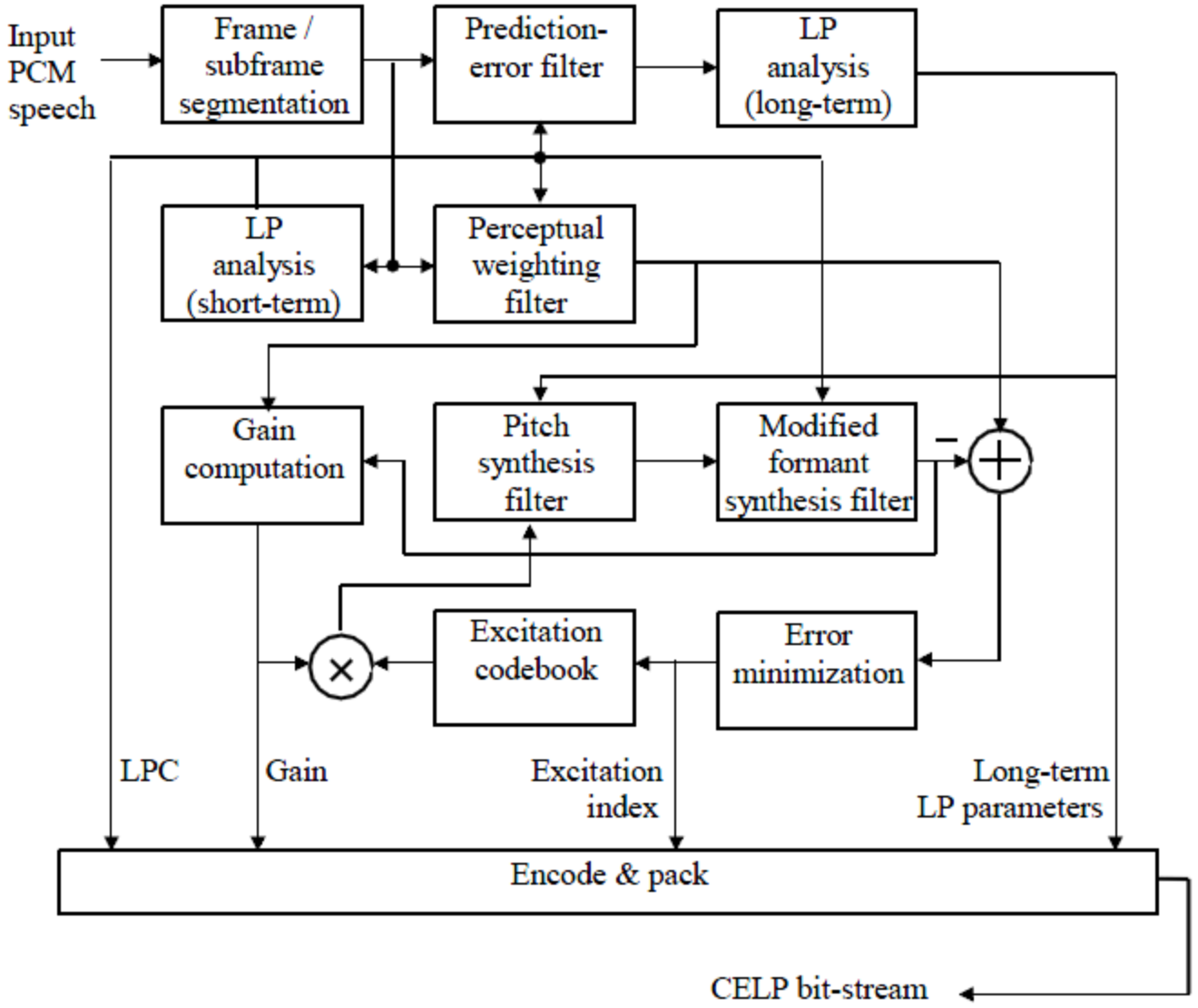


Połączenie obu filtrów

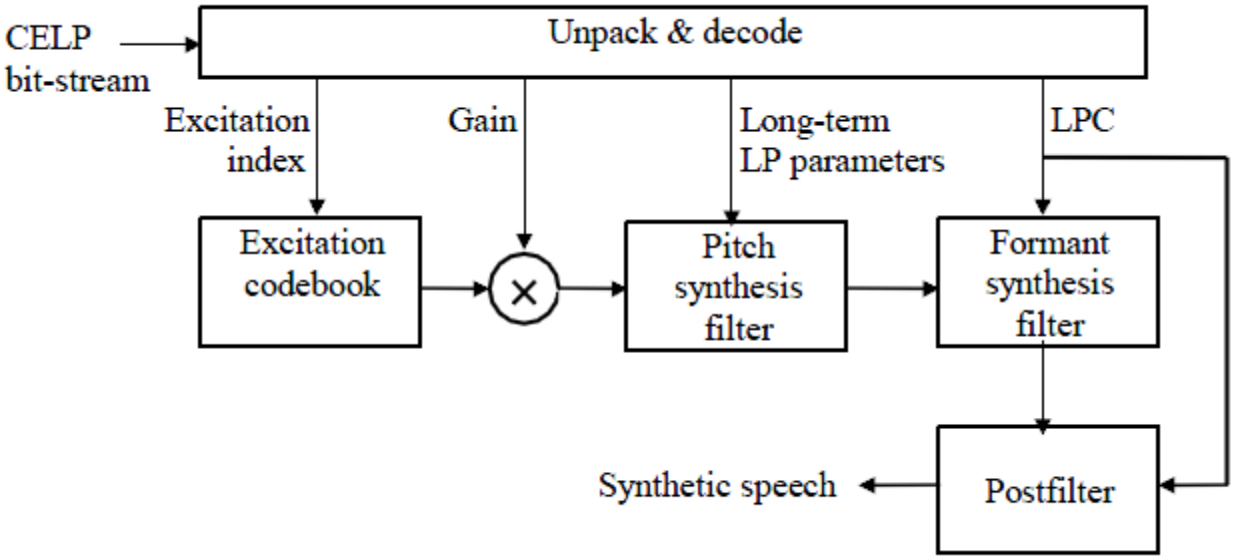
Koder CELP

- Segmentacja (ramki 20-30 ms, każda ramka ma zwykle 4 podramki)
- Analiza LPC *short-term* (w ramkach) oraz *long-term* (w podramkach)
- Wyznaczenie pobudzenia z książki kodowej
- Kodowanie strumienia bitowego
  - indeks pobudzenia
  - wzmocnienie
  - parametry long-term LPC (wysokość)
  - współczynniki LPC (short-term)

Koder CELP


Input
PCM
speech
Frame /
subframe
segmentation
Prediction-
error filter
LP
analysis
(long-term)
LP
analysis
(short-term)
Perceptual
weighting
filter
Gain
computation
Pitch
synthesis
filter
Modified
formant
synthesis filter
Excitation
codebook
Error
minimization
LPC
Gain
Excitation
index
Long-term
LP parameters
Encode & pack
CELP bit-stream

Dekoder CELP



Postfilter – filtracja perceptualna sygnału na podstawie współczynników LPC, uwydatnia formanty, poprawia jakość sygnału

CELP a LPC

- Poprawa jakości i zrozumiałości – lepszy dobór pobudzenia dla filtrów LPC, dostosowany do właściwości sygnału mowy
- Zwiększenie przepływności (dodatkowe informacje)
- Większa złożoność obliczeniowa (głównie dobór pobudzenia), a więc zwiększenie opóźnień.

Stosowane obecnie kodeki parametryczne w różny sposób optymalizują algorytm CELP.

Adaptacyjna książka kodowa

Praktyczne implementacje kodeków CELP wykorzystują dwie książki kodowe:

- adaptacyjna
  - reprezentuje przewidywalną część pobudzenia (na podstawie historii)
  - jest uaktualniana dla każdej ramki
- stała (*fixed*, stochastic)
  - reprezentuje zmienną część pobudzenia

Sygnał pobudzenia: suma wektorów z obu książek kodowych. Uproszczenie obliczeń, szczególnie wyznaczania wysokości dźwięku.

Low-delay CELP (LD-CELP)

Modyfikacja algorytmu CELP:

- zmniejszenie opóźnienia <2 ms
  (małe bloki próbek)
- zastosowanie wstecznej adaptacji predyktora:
  wsp. LPC obliczane z sygnału syntetycznego,
  nie muszą być przesyłane
- wysoki rząd predykcji (50), tylko short-term
- brak predykcji wysokości dźwięku
- znaczne większa złożoność

Kodek G.728, 16 kbit/s

Algebraic CELP (ACELP)

- Zmniejszenie złożoności i opóźnień, poprawa jakości kodowanej mowy
- Algebraiczna książka kodowa: wektory kodowe uzyskiwane przez dodawanie i przesuwanie ciągów impulsów
- Adaptacyjna książka kodowa
- Zespolona kwantyzacja wektorowa obliczonych parametrów

Kodek G.729 (CS-ACELP): 6,4 / 8,0 / 11,8 kbit/s

Odporność na utratę pakietów

- Klasyczne kodeki nie nadają się do zastosowań VoIP. Gdy tracone są pakiety danych, jakość mowy pogarsza się.
- Kodek iLBC - oparty na CELP, dostosowany do transmisji w sieciach IP
- Predykcja w niezależnych blokach próbek – zabezpieczenie przed zniekształceniami przy utracie pakietów
- Adaptacja książki kodowej w przód i wstecz
- Tryby pracy: 13,3 / 15,2 kbit/s

Tryby pracy kodeka

- Pasmo częstotliwości do 4 kHz (cz. próbkowania 8 kHz, tryb *narrowband*) wystarcza do celów zrozumiałości mowy.
- Nie wystarcza do uzyskania zadawalającej jakości mowy, trudno przesyłać muzykę
- Szerokopasmowe tryby pracy kodeków:
  - *wideband*: cz. próbkowania 16 kHz (pasmo do 8 kHz)
  - *ultra wideband*: 32 kHz (pasmo do 16 kHz)
  - *full band*: 48 kHz (do 24 kHz)

Metody zmniejszenia przepływności

- Zmienna przepływność bitowa (*variable bit rate*, VBR)
  - dostosowanie przepływności do parametrów sygnału (wejściowy parametr: *quality*, czyli pożądana jakość sygnału)
- Wykrywanie obecności mowy (*voice activity detection*, VAD)
- Wykrywanie przerw w transmisji (*discontinuous transmission*, DTX)
  - mowa nie jest kodowana i przesyłana gdy nie występuje na wejściu

Kodek Speex

- Speex jest dostosowany do transmisji w sieciach IP (utrata pakietów)
- Wykorzystuje CELP
- Tryby pracy *narrowband*, *wideband*
- Opcjonalne kodowanie VBR z podaną jakością
- Licencja *open source*

Przepływności:

- *narrowband*: 2,15 – 24,6 kbit/s
- *wideband*: 4,0 – 44,2 kbit/s

Kodeki GSM

Kodeki oparte na LPC, wykorzystywane do transmisji w sieciach komórkowych.

- GSM-FR (*Full Rate*, GSM 06.10)
  - najstarszy kodek GSM
  - RPE-LPT (*Regular Pulse Excitation – Long Time Prediction*), LPC 8. rzędu
  - 13 kbit/s (8 kHz), kiepska jakość
- GSM-HR (*Half Rate*, GSM 06.20)
  - mniejsza przepływność: 5,6 kbit/s
  - VSELP (*Vector-Sum Excited LP*)
  - gorsza jakość, mniejsze zużycie energii

Kodeki GSM

- GSM-EFR (*Enhanced Full Rate*, GSM 06.60)
  - wykorzystuje ACELP
  - lepsza jakość niż FR, większe zużycie energii
  - przepływność 12,2 kbit/s
- AMR (*Adaptive Multi-Rate*)
  - wykorzystuje ACELP
  - kodek hybrydowy (+kodowanie sygnałowe)
  - wykorzystanie VAD i DTX
  - poprawa jakości względem FR/EFR
  - 8 przepływności od 4,75 do 12,2 kbit/s

Kodeki GSM

- AMR-WB (AMR *Wideband*, G.722.2)
  - kodek szerokopasmowy do 7 kHz
  - przepływności 6,60 / 8,85 / 12,65 kbit/s
  - do „trudniejszych” zastosowań (hałas, muzyka): do 23,85 kbit/s
  - wymagane szerokopasmowe sieci
- AMR-WB+ (*Extended AMR-WB*)
  - sygnały stereo i wyższe cz. próbkowania
  - zaprojektowane dla „usług telekomunikacyjnych”
  - przepł. od 5,2 do 48,0 kbit/s

Kodek OPUS

OPUS (dawniej SILK) – kodek na licencji open source.

- Kodowanie mowy oraz muzyki
- Odporność na utratę pakietów
- Elementy dźwięczne: predykcja long-term LPC w wybielonym sygnale
- Elementy bezdźwięczne: LPC w sygnale przefiltrowanym górnoprzepustowo
- Skalowalność, niskie opóźnienia
- Cz. próbkowania 8 / 12 / 16 / 24 / 48 kHz
- Przepływności od 8 do 40 kbit/s

OPUS – kodowanie mowy

- Długość ramki: 10 lub 20 ms
- Regulowany stopień złożoności: od 0 do 10.
- Tryby pracy:

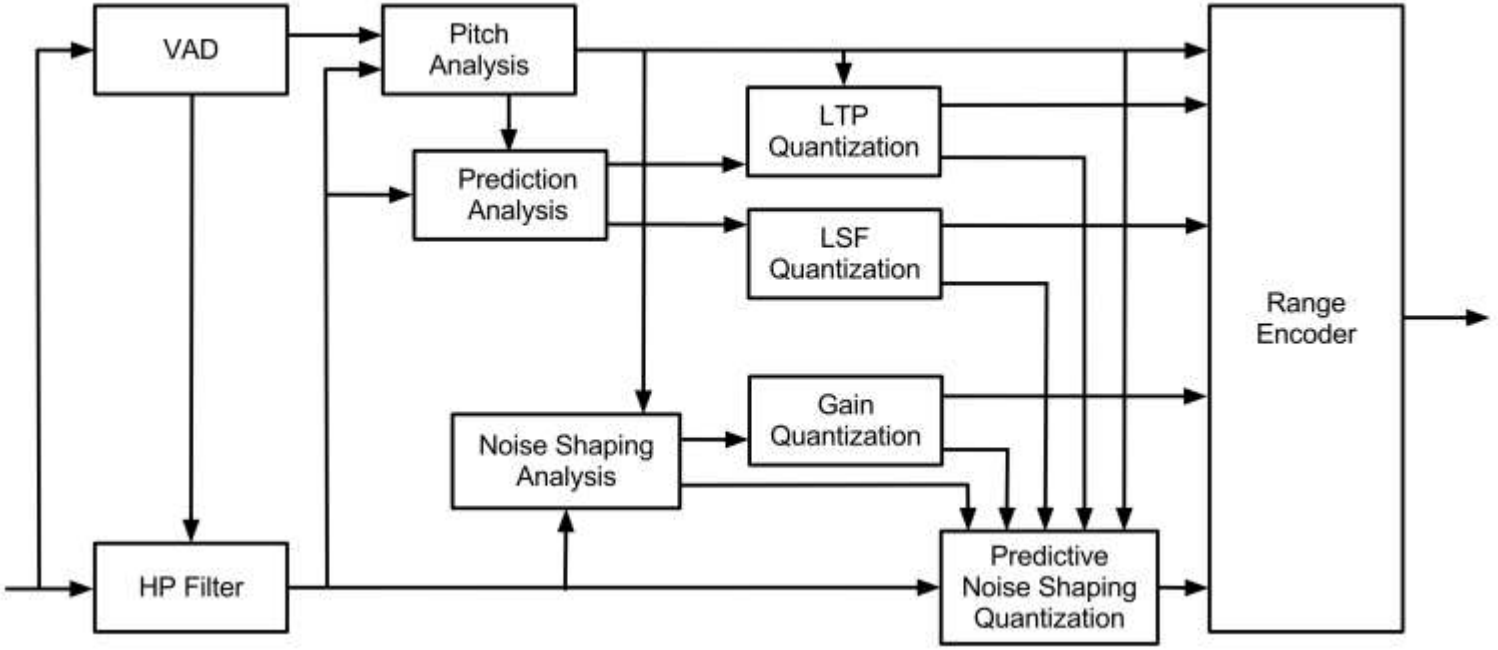
Sample Frequency	Name
48 kHz	Fullband
24 kHz	Super-wideband
16 kHz	Wideband
12 kHz	Mediumband
8 kHz	Narrowband

Input Type	Recommended Bitrate Range	
	Mono	Stereo
FB	28-40 kbps	48-72 kbps
SWB	20-28 kbps	36-48 kbps
WB	16-20 kbps	28-36 kbps
MB	12-16 kbps	20-28 kbps
NB	8-12 kbps	14-20 kbps

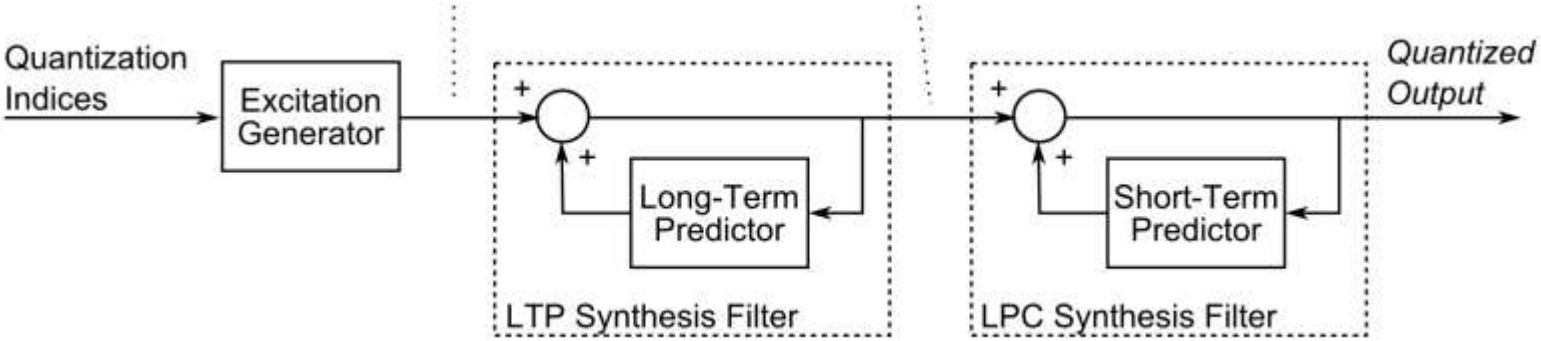
OPUS – kodowanie mowy

- Niskie częstotliwości (do 8 kHz)
  - kodowanie oparte na CELP.
- Wysokie częstotliwości (powyżej 8 kHz)
  - kodowanie w dziedzinie częstotliwości, oparte na współczynnikach MDCT (zmodyfikowana transformacja kosinusowa).
- Wyznaczanie wysokości dźwięku – analiza korelacji w sygnale po filtrze GP oraz po filtrze wybielającym. Wyznaczanie dźwięczności ramki.
- Kształtowanie szumu kwantyzacji (*noise shaping*) – redukcja uciążliwości, stosuje model psychoakustyczny.

OPUS – koder


VAD
Pitch Analysis
LTP Quantization
Prediction Analysis
LSF Quantization
Range Encoder
Gain Quantization
Noise Shaping Analysis
HP Filter
Predictive Noise Shaping Quantization

OPUS - dekoder



Literatura

- Wikipedia – hasła: *Speech coding*, *Speech codecs*
- Chu W.C., *Speech Coding Algorithms. Foundation and Evolution of Standardized Coders*, John Wiley & Sons, Hoboken 2003.
- *SPEEX: A free codec for free speech*. http://www.speex.org/ http://developer.skype.com/silk
- *K. Vos et al.: Voice Coding with OPUS*. https://www.opus-codec.org/presentations/opus_voice_aes135.pdf
- G. Szwoch: *Algorytmy kodowania źródłowego mowy* (raport). Dostępny na stronie Katedry (dział „Materiały pomocnicze”).
- Instrukcja do ćwiczenia laboratoryjnego nr 2: *Badanie algorytmów kodowania mowy*.