

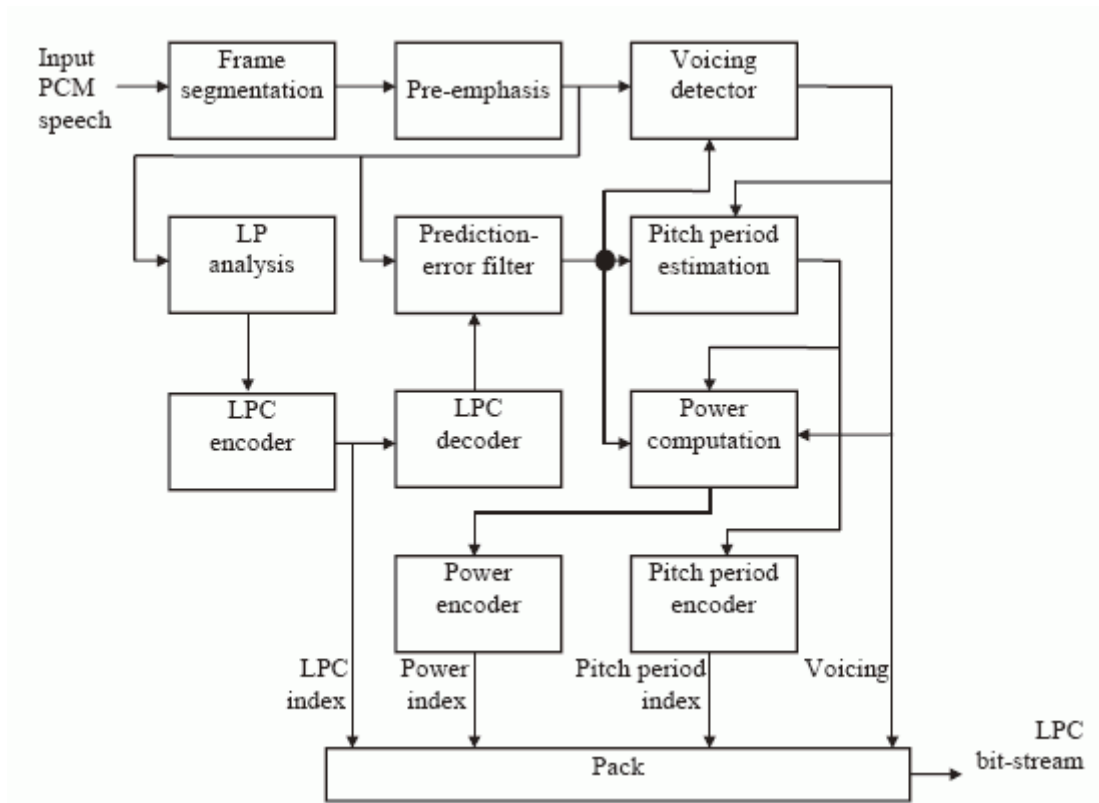
1 ALGORYTMY KODOWANIA ŹRÓDŁOWEGO MOWY

1.1 Liniowe kodowanie predykcyjne (LPC) – kodek FS1015

Większość źródłowych kodeków mowy opartych jest na liniowym kodowaniu predykcyjnym (LPC – *Linear Predictive Coding*) [1]. Algorytmy kodeków źródłowych wykorzystują model wytwarzania sygnału mowy przez człowieka. Sygnał mowy powstaje przez przefiltrowanie pobudzenia przez trakt głosowy. W dostatecznie krótkich przedziałach czasu sygnał mowy może być traktowany jako sygnał stacjonarny, zatem możliwe jest wyznaczenie parametrów filtru cyfrowego, odpowiadającego transmitancji traktu głosowego. Wtedy zamiast przesyłać próbki sygnału mowy, wystarczy przesłać współczynniki filtru, wraz z dodatkowymi informacjami niezbędnymi do odtworzenia sygnału w dekodерze. Operację taką należy wykonać dla każdej z ramek czasowych sygnału. Algorytm LPC wykorzystywany jest do takiego doboru współczynników filtru, aby uzyskać transmitancję filtru najlepiej dopasowaną do transmitancji traktu głosowego. Aby zminimalizować błąd predykcji, najczęściej stosowany jest algorytm Levinsona-Durbina.

Pierwszym powszechnie stosowanym kodekiem mowy, który wykorzystywał metodę LPC, był kodek FS1015, opracowany w roku 1982 do celów militarnych [1]. Oferował on małą przepływność bitową rzędu 2,4 kbit/s przy bardzo znaczącym pogorszeniu jakości sygnału mowy, ale przy zachowaniu zadawalającej zrozumiałości mowy. Kodek ten stał się podstawą dla udoskonalonych algorytmów późniejszych kodeków, które zostaną omówione w dalszej części raportu.

Schemat kodera LPC FS1015 przedstawiono na rys. 1. Sygnałem wejściowym jest sygnał spróbkowany z częstotliwością 8 kHz i rozdzielczością 12 bitów. Sygnał jest dzielony na ramki czasowe o długości 22,5 ms. Następnie sygnał jest poddawany preemfazie, polegającej na wzmocnieniu wysokich częstotliwości. Obwiednia widma sygnału mowy opada w kierunku wysokich częstotliwości. Zastosowanie preemfazy zapobiega błędom, które mogłyby powstać na etapie predykcji liniowej. Preemfaza jest stosowana w niemal wszystkich kodekach źródłowych.



Rys. 1. Schemat blokowy klasycznego kodera LPC (FS1015)

Kodek LPC w odmienny sposób traktuje sygnał mowy dźwięczny (*voiced*) i bezdźwięczny (*unvoiced*). Każda ramka jest analizowana przez detektor dźwięczności (*voicing detector*). Bit na wyjściu detektora sygnalizuje czy analizowana ramka zawiera sygnał dźwięczny czy bezdźwięczny. Analiza ramki opiera się na pomiarze energii (jest większa dla sygnału dźwięcznego), częstości przejść przez zero (większa dla sygnału bezdźwięcznego) i innych parametrów.

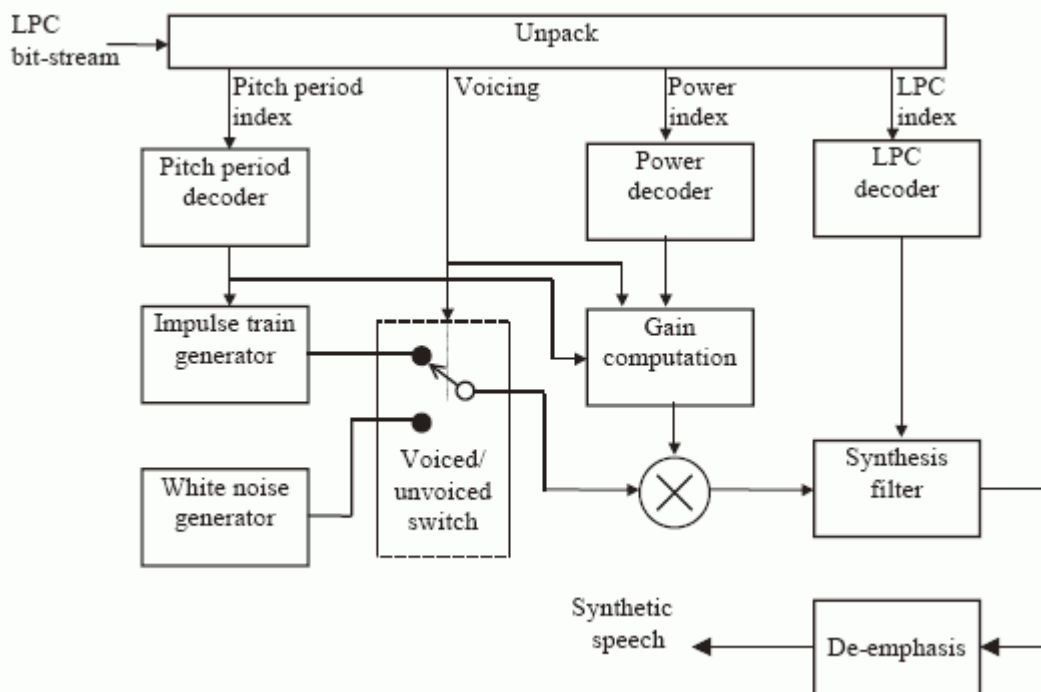
Analiza predykcyjna umożliwia wyznaczenie parametrów filtra predykcyjnego dla analizowanej ramki sygnału. Filtr predykcyjny jest dziesiątego rzędu (stąd częsta nazwa metody: LPC-10). W klasycznym kodeku FS1015 do wyznaczania współczynników predykcji używa się kowariancji, w późniejszych kodekach stosuje się autokorelację. Wyznaczone współczynniki predykcji poddaje się kwantyzacji skalarnej. Etap ten umożliwia uzyskanie informacji na temat obwiedni widma analizowanej ramki czasowej.

Jeżeli ramka czasowa została zaklasyfikowana jako dźwięczna, wyznaczana jest wysokość dźwięku (*pitch period estimation*). W tym celu, na podstawie skwantyzowanych współczynników predykcji, tworzony jest filtr błędu predykcji (*prediction-error filter*), który zostaje następnie użyty do przefiltrowania sygnału wejściowego po preemfazie. Sygnał po filtracji zostaje następnie wykorzystany do wyznaczenia okresu sygnału. Użycie sygnału błędu predykcji zamiast sygnału oryginalnego ma tą zaletę, że w sygnale błędu predykcji nie są obecne składowe widmowe związane z transmitancją traktu głosowego, a jedynie z samym pobudzeniem, zatem łatwiej jest wyznaczyć

częstotliwość podstawową. Dokonuje się tego obliczając funkcję różnic amplitudy (*magnitude difference function* – MDF) i znajdując jej minimum. Ponadto wyznaczana jest moc sygnału błędu predykcji.

Strumień bitów na wyjściu kodera FS1015 zawiera 54 bity dla każdej ramki analizy. Informacje o rodzaju sygnału (dźwięczny/bezdźwięczny) i o częstotliwości podstawowej zajmują 7 bitów, informacje o mocy sygnału w ramce – 5 bitów. Zakodowane współczynniki LPC zajmują 41 bitów dla ramek dźwięcznych i 20 bitów dla ramek bezdźwięcznych. Jeden bit jest przeznaczony dla synchronizacji. Pozostałych 21 bitów dla ramek bezdźwięcznych wykorzystuje się do kodowania protekcyjnego. W rezultacie uzyskuje się strumień danych o przepływności 2,4 kbit/s.

Schemat dekodera FS1015 przedstawiono na rys. 2. W zależności od stanu bitu dźwięczności wybierany jest jeden z dwóch rodzajów pobudzenia. Dla ramek dźwięcznych sygnałem pobudzającym dla dekodera jest ciąg impulsów o jednostkowej amplitudzie. Częstotliwość impulsów jest dobierana na podstawie odebranych informacji o okresie sygnału. Dla ramek bezdźwięcznych sygnałem pobudzającym jest szum o jednostkowej wariancji. Amplituda sygnału pobudzającego jest ustalana na podstawie informacji o mocy sygnału. Zdekodowane współczynniki predykcji służą do skonstruowania filtru syntetyzującego (*synthesis filter*), który filtruje sygnał pobudzenia. Przefiltrowany sygnał syntetyczny przed przesłaniem na wyjście jest jeszcze poddawany deemfazie.



Rys. 2. Schemat blokowy dekodera LPC (FS1015)

Klasyczny kodek LPC (w implementacji FS1015) posiada szereg ograniczeń [1].

- Kodek klasyfikuje ramki sygnału „binarnie” jako dźwięczne lub bezdźwięczne. W wielu przypadkach nie można jednoznacznie zaklasyfikować ramek sygnału do jednej z tych kategorii (np. w stanach transjentowych).
- Użycie ciągu impulsów jednostkowych jako pobudzenia dla ramek dźwięcznych nie odpowiada praktycznym przypadkom, w których pobudzenie jest najczęściej połączeniem składowych quasi-periodycznych i szumu.
- Tracone są wszelkie informacje o fazie. Widmo sygnału syntetycznego naśladuje widmo sygnału oryginalnego, jednak przebiegi czasowe są całkowicie różne. Chociaż ucho ludzkie jest niewrażliwe na zmiany fazy, zachowanie częściowej informacji o fazie sygnału korzystnie wpływa na jakość sygnału.
- Stosowanie pobudzenia w postaci ciągu impulsów dla ramek dźwięcznych jest sprzeczne z zasadą wyznaczania parametrów filtru za pomocą LPC – zakłada się tam, że pobudzeniem będzie szum biały. Powoduje to zniekształcenia sygnału syntetycznego.

Ograniczenia te stały się przyczyną opracowania udoskonalonych algorytmów kodowania źródłowego. Algorytmu LPC w opisaney w tym punkcie formie nie stosuje się już do kodowania sygnału mowy.

1.2 Kodek CELP

Algorytm CELP (*Code Excited Linear Prediction*) został opracowany w roku 1985 jako rozwinięcie idei kodeka LPC, w celu usunięcia niedoskonałości starszego kodeka i uzyskania poprawy jakości sygnału syntetycznego, kosztem pewnego zwiększenia złożoności obliczeniowej. Przykładem praktycznej implementacji algorytmu jest kodek FS1016 (rok 1992). W kodeku CELP nie ma klasyfikowania ramek sygnału jako dźwięcznych lub bezdźwięcznych oraz osobnego traktowania tych dwóch typów sygnału. Ponadto, zamiast dwóch typów pobudzenia (szum lub ciąg impulsów) stosuje się szereg różnych pobudzeń, zapisanych w tzw. książce kodowej, która może być stała (*fixed*) lub adaptacyjna (*adaptive*). Dla każdej ramki z książki kodowej wybierany jest ten rodzaj pobudzenia, który daje najmniejszy błąd kodowania [1]. Przy odpowiednim doborze wektorów w książce kodowej, obejmujących zakres typowych pobudzeń do wytwarzania sygnału mowy, możliwe jest zminimalizowanie błędów kodowania.

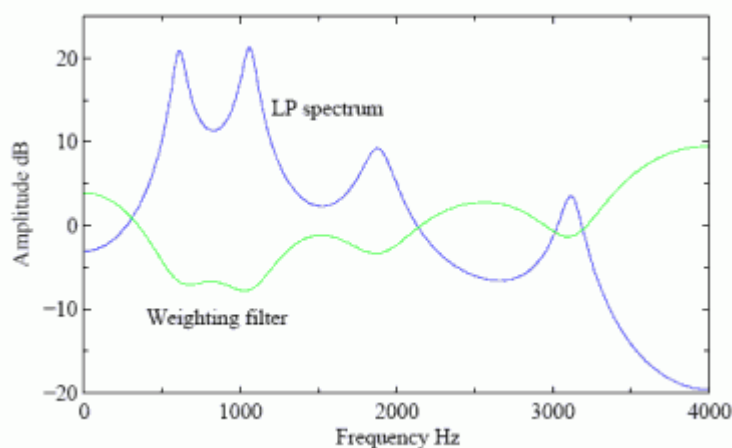
Działanie kodeka CELP oparte jest na zasadzie „analizy przez syntezę” (*analysis-by-synthesis*). Wybór wektora pobudzenia z książki kodowej odbywa się w pętli zamkniętej. Każdy z wektorów zapisanych w książce kodowej jest przetwarzany przez filtr syntetyzujący, którego parametry uzyskiwane są w procesie analizy w pętli otwartej. Uzyskany w ten sposób sygnał

syntetyczny jest porównywany z sygnałem oryginalnym i obliczana jest wielkość błędu. W ten sposób wybierany jest wektor kodowy, który daje najmniejszy błąd kodowania.

Uproszczony schemat kodera CELP przedstawiono na rys. 3. Wejściowy sygnał jest dzielony na ramki o długości od 30 ms (240 próbek). Każda ramka jest dzielona na cztery podramki (*subframes*) o jednakowej długości. Na podstawie ramek sygnału (zwykle poddanego preemfazie) obliczane są współczynniki predykcji (LPC) krótkookresowej (10. rzędu), dzięki czemu uzyskuje się informacje o kształcie widma sygnału. Za pomocą filtru uzyskuje się następnie sygnał błędu predykcji, na podstawie którego wyznaczane są współczynniki długookresowej predykcji, dające informację o okresie sygnału. Predykcja długookresowa dokonywana jest w podramkach, ponieważ dane do wyznaczania okresu sygnału muszą być uaktualniane częściej niż wynika to z długości ramki.

Kolejnym etapem kodowania jest wyznaczenie sygnału pobudzającego. Długość każdego z wektorów pobudzających w książce kodowej jest równa długości podramki analizy, zatem wyznaczanie wektora pobudzającego odbywa się raz dla każdej z podramek analizy. Każdy z wektorów z książki kodowej jest wstępnie przetwarzany (filtracja, ustalenie amplitudy), a następnie filtrowany za pomocą dwóch filtrów syntetyzujących: filtru wysokości dźwięku (*pitch synthesis filter*) oraz filtru formantowego (*formant synthesis filter*). Parametry pierwszego z filtrów są wyznaczane na podstawie długookresowej analizy LPC. Transmitancja tego filtru umożliwia ustalenie częstotliwości formantów widma. Drugi filtr, o parametrach wyznaczanych na podstawie krótkookresowej analizy LPC, kształtuje obwiednię widma (amplitudy formantów). Przefiltrowanie sygnału pobudzenia przez oba filtry pozwala uzyskać sygnał o odpowiednim widmie (rys. 4).

wejściowego. W tej części algorytmu kodowania źródłowego pojawiają się bowiem elementy algorytmów kodowania perceptualnego [1]. Celem operacji przeprowadzanych w pętli zamkniętej jest wyznaczenie wektora kodowego, który umożliwi uzyskanie najmniejszej wartości błędu (różnicy między sygnałem wejściowym a rzeczywistym). Jako miarę tej różnicy można przyjąć wartość błędu średniokwadratowego, jednak takie podejście nie musi dać najlepszej subiektywnej jakości sygnału mowy. Lepsze rezultaty można uzyskać wykorzystując zjawisko maskowania równoczesnego. Sygnały wejściowy oraz syntetyczny mają formantową strukturę widmową – widmo zawiera lokalne maksima i minima. Sygnał błędu (różnicowy) ma charakter szumowy. Szum ten będzie bardziej uciążliwy w pobliżu minimów widmowych (niski poziom sygnału). Natomiast w pobliżu formantów (maksimów widma) wpływ tego szumu na jakość sygnału będzie mniejszy, ponieważ szum będzie maskowany przez sygnał użyteczny. Zatem sygnał błędu można przefiltrować przy użyciu funkcji wagowej, której kształt jest odwrotnością kształtu widma. Dokonuje tego filtr ważenia perceptualnego (*perceptual weighting filter*), który wzmacnia sygnał błędu w pobliżu minimów widmowych i tłumi go w pobliżu formantów. Przetworzony w ten sposób sygnał błędu pozwala na obliczenie miary błędu kodowania, która bardziej odpowiada percepcji zniekształceń przez ucho ludzkie. Przykładową charakterystykę filtru ważącego przedstawiono na rys. 5. Współczynniki filtru ważącego ustala się na podstawie współczynników filtru formantowego, z uwzględnieniem stałej skalowania γ , której wartość mieści się zwykle w zakresie od 0,8 do 0,9.



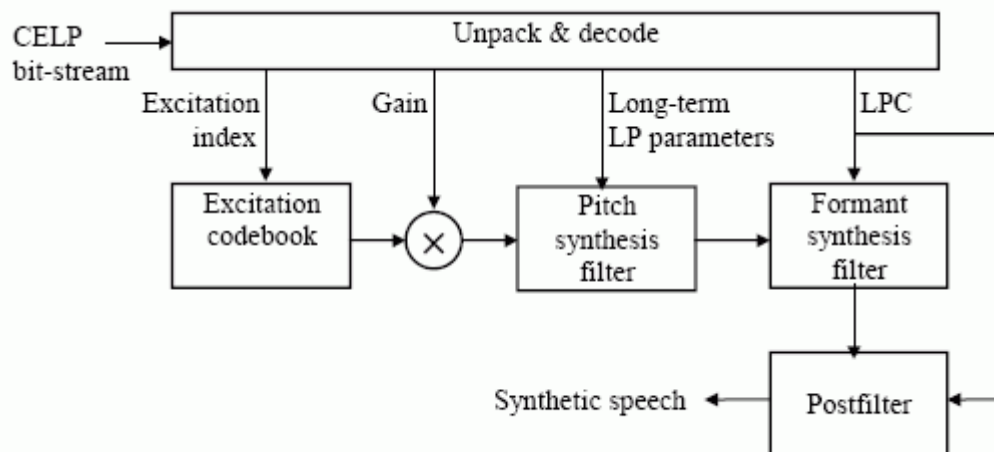
Rys. 5. Przykładowa charakterystyka widmowa sygnału syntetycznego (krzywa niebieska) i odpowiadająca mu krzywa ważąca (charakterystyka filtru ważącego – krzywa zielona)

Należy jeszcze dodać, że w praktycznych implementacjach kodeka, kaskadowo połączone filtry: formantowy oraz ważący, są łączone w jeden filtr, nazywany zmodyfikowanym filtrem syntetyzującym formanty (*modified formant synthesis filter*). Modyfikacja ta ma na celu optymalizację algorytmu (zmniejszenie liczby wymaganych obliczeń).

Operacja wyznaczenia wektora pobudzenia z książki kodowej jest najbardziej złożoną częścią algorytmu kodowania CELP. Wyszukiwanie wektora odbywa się dla każdej podramki sygnału wejściowego i jest poprzedzone przefiltrowaniem sygnału wejściowego przez perceptualny filtr ważący. Dla każdego z wektorów kodowych wyznaczana jest optymalna wielkość wzmocnienia (minimalizująca błąd), a następnie sygnał pobudzający jest skalowany na podstawie wyznaczonego wzmocnienia. Przetworzony wektor pobudzenia jest filtrowany najpierw przez filtr syntezy wysokości dźwięku, a następnie przez zmodyfikowany filtr formantowy. Poprzez odjęcie przetworzonego sygnału pobudzającego od przetworzonego sygnału wejściowego uzyskuje się sygnał błędu, z którego wyznaczana jest wielkość energii (miara błędu). W praktycznych implementacjach kodeków CELP stosuje się szereg optymalizacji, których celem jest rozdzielenie operacji na wykonywane jednokrotnie dla podramki sygnału i na wykonywane dla każdego wektora kodowego (usunięcie redundancji w obliczeniach).

Strumień bitów na wyjściu koder CELP zawiera skantyzowane wartości: współczynników LPC, wzmocnienia, indeksu wektora z książki kodowej oraz parametrów długookresowej predykcji.

Blok dekodera CELP jest znacznie prostszy od koder (rys. 6). Na podstawie zdekodowanego indeksu wybierany jest z książki kodowej odpowiedni wektor pobudzenia (dekoder dysponuje tą samą książką kodową co koder). Wektor pobudzający jest skalowany za pomocą odebranej wartości wzmocnienia, a następnie filtrowany za pomocą dwóch filtrów syntetyzujących, których współczynniki są wyznaczone przez parametry długookresowej i krótkookresowej LPC.



Rys. 6. Schemat blokowy dekodera CELP

Dodatkowym blokiem, jaki pojawia się w dekodery jest filtr końcowy (*postfilter*). Jest to blok opcjonalny, może zostać wyłączony. Jego zadaniem jest poprawa subiektywnej jakości zdekodowanego sygnału poprzez zmniejszenie poziomu szumu percypowanego przez słuchacza. Filtr końcowy jest odpowiednikiem perceptualnego filtru ważącego w koderze. Jego zadaniem jest poprawa jakości sygnału wynikowego poprzez uwypuklenie formantów widmowych.

Współczynniki filtru końcowego wyznaczane są na podstawie parametrów LPC. Stosowanych jest kilka różnych implementacji filtru końcowego. W celu uniknięcia zniekształceń, obok filtru końcowego stosuje się również blok automatycznej regulacji wzmocnienia. Filtr końcowy, podobnie jak filtr wazący sygnał błędu w koderze, również w pewnym stopniu uwzględnia model perceptualny słyszenia.

W roku 1992 przyjęto w USA algorytm CELP jako standard kodowania mowy, oznaczony symbolem FS1016 [1]. W stosunku do tradycyjnego algorytmu CELP wprowadzono usprawnienia, których celem była poprawa jakości mowy. Pomimo tego, że standard FS1016 nie jest już stosowany, modyfikacje te znalazły zastosowanie również w innych kodekach mowy. W oryginalnym kodeku CELP, przy minimalizacji błędu kodowania nie uwzględniano wyznaczania okresu sygnału za pomocą długookresowej predykcji liniowej. Proces wyszukiwania wektora kodowego umożliwiającego uzyskanie najmniejszego błędu powinien również uwzględniać predykcję długookresową. Nie jest to możliwe przy używaniu ustalonej z góry książki kodowej. Dlatego wprowadzono w kodeku FS1016 drugi rodzaj książki kodowej, nazwanej adaptacyjną (*adaptive codebook*). Służy ona do wyznaczania okresu sygnału. Adaptacyjna książka kodowa zmienia się z każdą podramką analizowanego sygnału (jest uaktualniana). Wektory w adaptacyjnej książce kodowej nakładają się na siebie. Dla każdej podramki wyszukiwany jest wektor kodowy dający najmniejszą wielkość błędu. Wektor ten służy do wyznaczania okresu sygnału, upraszcza on również obliczenia. W skrócie, wartość okresu sygnału odpowiada indeksowi wektora kodowego zawierającego sygnały pobudzające dla poprzednich podramek, który daje najmniejszy błąd dla aktualnej podramki. Wartość okresu wyznaczana jest najpierw jako liczba całkowita (*integer pitch period*), a następnie, metodą interpolacji, jako bardziej dokładna liczba ułamkowa (*fractional pitch period*). Daje to większą rozdzielczość czasową analizy długookresowej. Okres sygnału jest kodowany jako indeks okresu (dla podramek 1 i 3) lub jako przesunięcie względem poprzedniej wartości (podramki 2 i 4).

Druga (stała) książka kodowa jest w koderze FS1016 nazywana stochastyczną (*stochastic codebook*) i służy do wyznaczenia wektora pobudzenia. Wprowadzono jedną modyfikację: wektory kodowe nakładają się na siebie (posiadają części wspólne). Daje to zmniejszenie objętości książki kodowej oraz zmniejszenie złożoności obliczeniowej algorytmu. Stochastyczna książka kodowa zawiera 1082 sygnały uzyskane na drodze przetworzenia szumu gaussowskiego o jednostkowej wariancji. Stochastyczna książka kodowa wprowadza składowe szumowe do sygnału – jest to główne źródło zniekształceń sygnału syntetycznego w kodeku FS1016.

Strumień bitów na wyjściu koderza FS1016 zawiera skwantyzowane dane dotyczące: indeksu parametrów LPC, indeksu i amplitudy pobudzenia z adaptacyjnej książki kodowej (informacje o

okresie sygnału) oraz indeksu i amplitudy wektora ze stochastycznej książki kodowej. Łączna liczba bitów dla jednej ramki sygnału wynosi 144, co daje przepływność 4,8 kbit/s (przy standardowej długości ramki 30 ms).

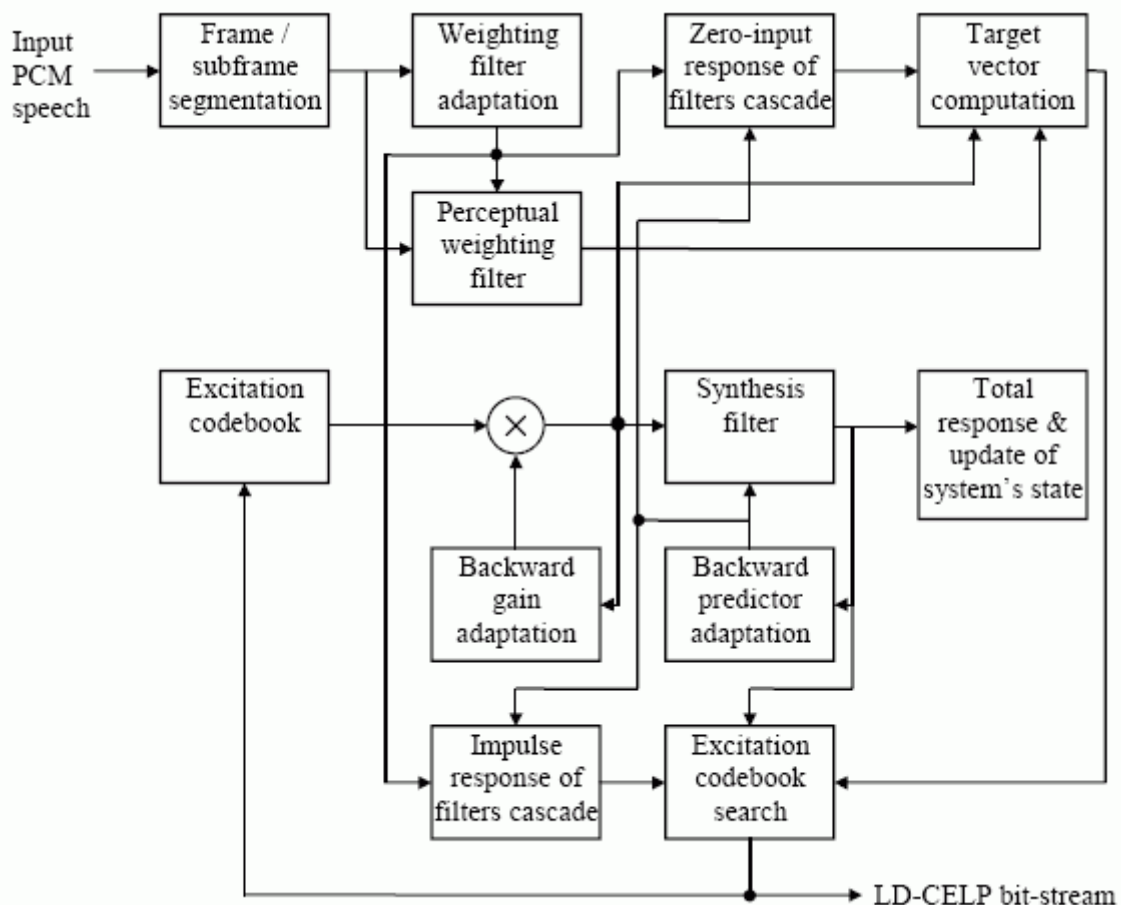
1.3 Kodek Low Delay CELP (LD-CELP) – G.728

Większość kodeków źródłowych mowy skupiała się na uzyskaniu możliwie małej przepływności bitowej, przy zachowaniu zrozumiałości mowy. Odbywa się to kosztem zwiększenia opóźnień przesyłania danych, co jest związane z koniecznością buforowania danych (przetwarzania ramek sygnału). Duże wartości opóźnień mogą powodować jednak pogorszenie jakości odbieranego sygnału, zwłaszcza w takich zastosowaniach, jak telefonia VoIP. Aby nie trzeba było dokonywać kompromisu pomiędzy wielkością opóźnienia a przepływnością bitową, opracowano kodek *Low Delay CELP* (LD-CELP), charakteryzujący się małymi opóźnieniami przy kodowaniu sygnału, przy zachowaniu względnie niskiej przepływności bitowej (16 kbit/s). Odbywa się to kosztem zwiększenia złożoności obliczeniowej algorytmu [1]. Kodek wykorzystujący algorytm LD-CELP został objęty standardem ITU-T w roku 1992 i jest znany pod oznaczeniem G.728 [3].

Najważniejsze rozwiązania wprowadzone w kodeku LD-CELP w celu zmniejszenia opóźnień (skrócenia czasu obliczeń) wymieniono poniżej.

- Zmniejszenie długości ramki analizy do 20 próbek. Daje to zmniejszenie opóźnień związanych z buforowaniem próbek (opóźnienie wprowadzane przez koder jest rzędu 1,25-1,875 ms w porównaniu z 20-30 ms w kodeku CELP). Kodowanie może rozpocząć się już po otrzymaniu pięciu próbek sygnału.
- Rekursywna estymacja autokorelacji przy wyznaczaniu współczynników predykcji (wykorzystanie okna Chena).
- Predykcja zewnętrzna (*external prediction*) – parametry LPC są wyznaczone na podstawie poprzednich próbek sygnału i są stosowane do bieżącej ramki analizy.
- Adaptacyjna predykcja wstecz (*backward adaptive prediction*) – parametry LPC są obliczane na podstawie sygnału syntetycznego, a nie sygnału wejściowego. Dzięki temu nie ma konieczności kwantyzacji i przesyłania współczynników predykcji w strumieniu bitowym (dzięki temu zmniejsza się przepływność bitowa).
- Wysoki rząd predykcji, równy 50. Nie ma predykcji długookresowej.
- Adaptacja wzmocnienia sygnału pobudzającego wstecz (*backward excitation gain adaptation*). Wartość wzmocnienia jest wyznaczana na podstawie poprzednich sygnałów pobudzających i nie musi być kwantyzowana i przesyłana.

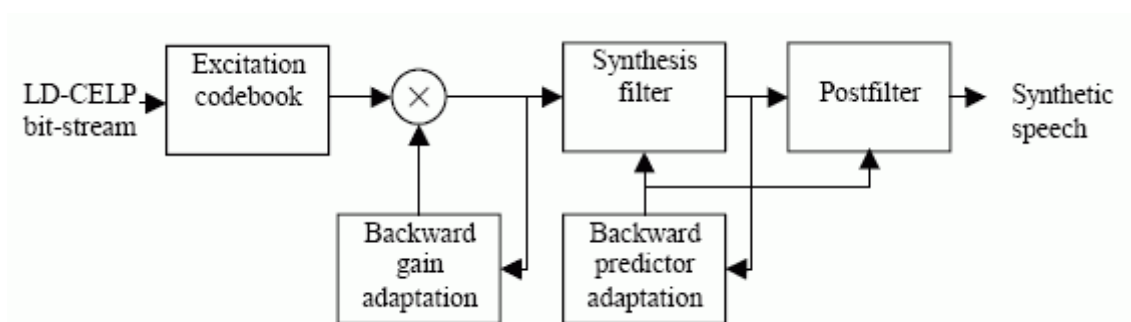
Schemat kodera LD-CELP G.728 przedstawiono na rys. 7. Dane są przetwarzane w ramach o długości 20 próbek (4 podramki). Predykcja liniowa jest dokonywana niezależnie w trzech różnych blokach algorytmu. Perceptualny filtr ważący wykorzystuje wyniki analizy LPC 10. rzędu, uzyskane na podstawie poprzednich próbek sygnału wejściowego. Współczynniki filtru są uaktualniane dla każdej ramki. Współczynniki filtru syntezy są wyznaczone na podstawie analizy LPC 50. rzędu próbek sygnału syntetycznego z poprzednich ramek analizy. Wzmocnienie sygnału pobudzającego (*excitation gain*) wyznaczone jest dla każdej podramki przy użyciu predykcji 10. rzędu w skali logarytmicznej, a współczynniki filtru są uaktualniane dla każdej ramki na podstawie poprzednich ramek. Wyszukiwanie wektora pobudzającego w książce kodowej jest przeprowadzane dla każdej podramki. Przeskalowane sygnały pobudzeń są filtrowane przez filtr syntetyzujący, a następnie wyszukiwany jest wektor dający najmniejszą wartość błędu. Wyjściowy strumień bitów zawiera indeksy pobudzeń z książki kodowej. Dla każdej podramki transmitowanych jest 10 bitów, co daje przepływność bitową 16 kbit/s, w praktyce jednak dodawane są jeszcze bity synchronizacji.



Rys. 7. Schemat blokowy kodera LD-CELP G.728

Schemat dekodera LD-CELP przedstawiono na rys. 8. Na podstawie zdekodowanego strumienia wejściowego wybierane są z książki kodowej odpowiednie sygnały pobudzające. Sygnały te są przetwarzane przez filtr syntetyzujący. Współczynniki filtru syntezy oraz wartości

wzmocnienia dla sygnałów pobudzających są obliczane na podstawie poprzednich próbek zrekonstruowanego sygnału syntetycznego. Nie jest tu stosowane perceptualne ważenie sygnału, stosuje się natomiast filtr końcowy (*postfilter*) w celu poprawy jakości sygnału. Filtr końcowy jest kaskadowym połączeniem dwóch filtrów: krótkookresowego (współczynniki filtru wyznaczone są na podstawie analizy LPC 10. rzędu, w praktyce używa się współczynników obliczanych dla filtru syntetyzującego) oraz długookresowego (filtr grzebieniowy, którego maksima są położone na wielokrotnościach częstotliwości podstawowej). Bloki perceptualnego filtru ważącego w koderze oraz filtru końcowego w dekodерze stanowią elementy kodowania opartego na modelu psychoakustycznym w źródłowym kodeku LD-CELP.



Rys. 8. Schemat blokowy dekodera LD-CELP G.728

Warto jeszcze wspomnieć, że z powodu odmiennej niż w przypadku CELP struktury algorytmu (brak bloku predykcji długookresowej do wyznaczania okresu sygnału), kodek LD-CELP nadaje się w pewnym stopniu również do kodowania sygnałów innych niż mowa, a w każdym razie nie powoduje tak wyraźnego zniekształcenia tych sygnałów (należy jednak pamiętać o ograniczeniu pasma częstotliwości).

1.4 Kodek CS-ACELP G.729

Algorytm ACELP (*Algebraic Code Excited Linear Prediction*) powstał jako rozwinięcie idei algorytmu CELP. Modyfikacje oryginalnego algorytmu skierowane zostały w stronę zmniejszenia złożoności obliczeniowej algorytmu (a przez to zmniejszenia opóźnień) oraz poprawy jakości sygnału syntetycznego [1]. Podstawy algorytmu ACELP sformułowano w roku 1987, a praktyczną implementację kodeka objęto standardami ITU-T w roku 1995 jako dwa odrębne algorytmy: G.729 oraz G.723.1. W tym podrozdziale zostanie omówiony pierwszy z wymienionych kodeków.

Kodek G.729 nosi oznaczenie CS-ACELP (*Conjugate Structure ACELP*) [4]. Wprowadza on do dziedziny kodowania źródłowego mowy dwa nowatorskie rozwiązania. Pierwszym jest algebraiczna książka kodowa (*algebraic codebook*). Wektory pobudzające z książki kodowej są

uzyskiwane na drodze przekształceń matematycznych – dodawania i przesuwania. Nie ma konieczności przechowywania książki kodowej, ponieważ może ona być w każdej chwili odtworzona. Drugą innowacją jest zastosowanie zespolonej kwantyzacji wektorowej, w której kwantyzacji poddawane są nie pojedyncze wartości uzyskane z analizy sygnału wejściowego, ale cały wektor parametrów. Umożliwia to dokonanie optymalnej kwantyzacji i poprawę jakości sygnału syntetycznego. Wszystkie nowoczesne kodeki źródłowe mowy stosują kwantyzację wektorową zamiast skalarnej.

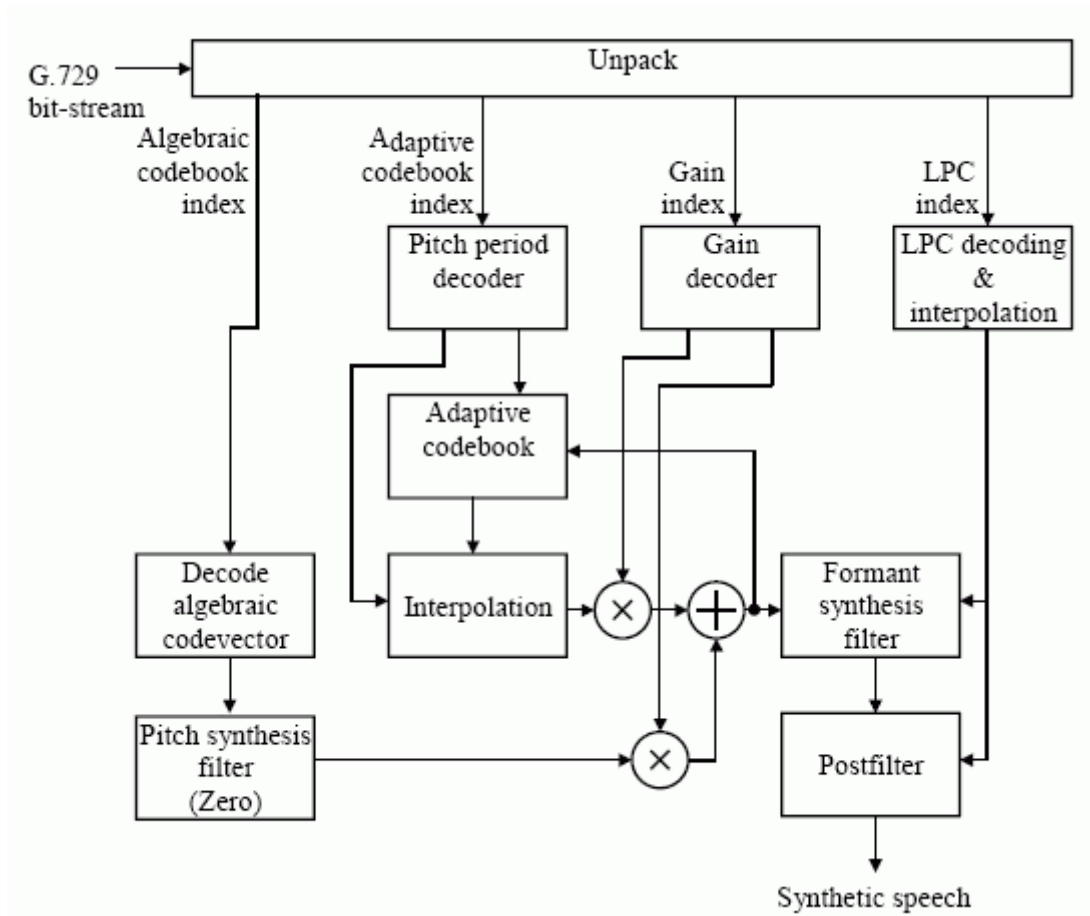
Algorytm kodeka G.729 wykorzystuje dwie książki kodowe: algebraiczną i adaptacyjną. Algebraiczna książka kodowa zbudowana jest według zasady *interleaved single-pulse permutation*. Książka kodowa składa się z czterech wierszy po 40 próbek każdy. Dla każdej z 40 próbek, w jednym z wierszy znajduje się impuls, który może mieć amplitudę +1 lub -1. Wektory kodowe są konstruowane w ten sposób, że każdy wektor zawiera cztery impulsy (wartość +1 lub -1), a na pozostałych pozycjach znajdują się zera. Indeks wektora kodowego jest zapisywany za pomocą 17 bitów.

Koder w standardzie G.729 jest znacznie bardziej złożony niż w starszych kodekach (rys. 9). Wzrost złożoności obliczeniowej kodeka w celu poprawy jakości sygnału jest akceptowalny, biorąc pod uwagę wzrost mocy obliczeniowej procesorów sygnałowych. Sygnał wejściowy jest przetwarzany przez filtr górnoprzepustowy i dzielony na ramki o długości 10 ms (80 próbek). Każda ramka jest dzielona na dwie podramki o długości 5 ms (40 próbek). W wyniku analizy LPC uzyskuje się dwa zestawy współczynników LPC (zapisywanych w postaci LSF – *Line Spectral Frequency*) – skwantyzowane i oryginalne. Oba zestawy parametrów są interpolowane pomiędzy podramką poprzednią a bieżącą. Oryginalne współczynniki LPC są wykorzystywane przez perceptualny filtr ważący do przetworzenia sygnału wejściowego. Współczynniki filtru ważącego są uaktualniane dla każdej podramki. Transmitancja filtru jest adaptacyjnie dostosowywana do sygnału (wykorzystuje się miarę płaskości widma).

Wyznaczanie okresu sygnału odbywa się w pętli otwartej dla każdej ramki. Na podstawie dwóch podramek sygnału przetworzonego przez filtr perceptualny wyznaczana jest wartość autokorelacji dla trzech zakresów wartości opóźnienia. W każdym zakresie wyznaczane jest maksimum autokorelacji, po czym dokonywana jest normalizacja i wyznaczana jest największa wartość autokorelacji, wyznaczająca okres sygnału. Dla pierwszej podramki kodowana jest ułamkowa wartość okresu sygnału, dla drugiej – przesunięcie względem pierwszej podramki.

Przeszukiwanie adaptacyjnej książki kodowej odbywa się na podstawie trzech sygnałów: sekwencji pobudzającej (wyznaczonej na podstawie sygnałów pobudzających z poprzednich ramek oraz sygnału błędu predykcji z bieżącej ramki), sekwencji docelowej (sygnał wejściowy

interpolowane wartości współczynników LPC. Sygnał wyjściowy jest przetwarzany przez filtr końcowy.



Rys. 10. Schemat blokowy dekodera CS-ACELP G.729

Zamieszczony w tym rozdziale opis kodeka G.729 jest bardzo skrótowy. Pełny opis algorytmu kodeka można znaleźć w normach ITU-T. Jednak z powyższego opisu można wywnioskować, że struktura algorytmów współczesnych kodeków źródłowych mowy jest bardzo złożona w porównaniu do stosunkowo prostej struktury algorytmu LPC w kodeku FS1015. Kodeki mowy takie jak G.729 realizują zaawansowane operacje przetwarzania sygnału, wymagające dużej mocy obliczeniowej.

1.5 Kodek ACELP G.723.1

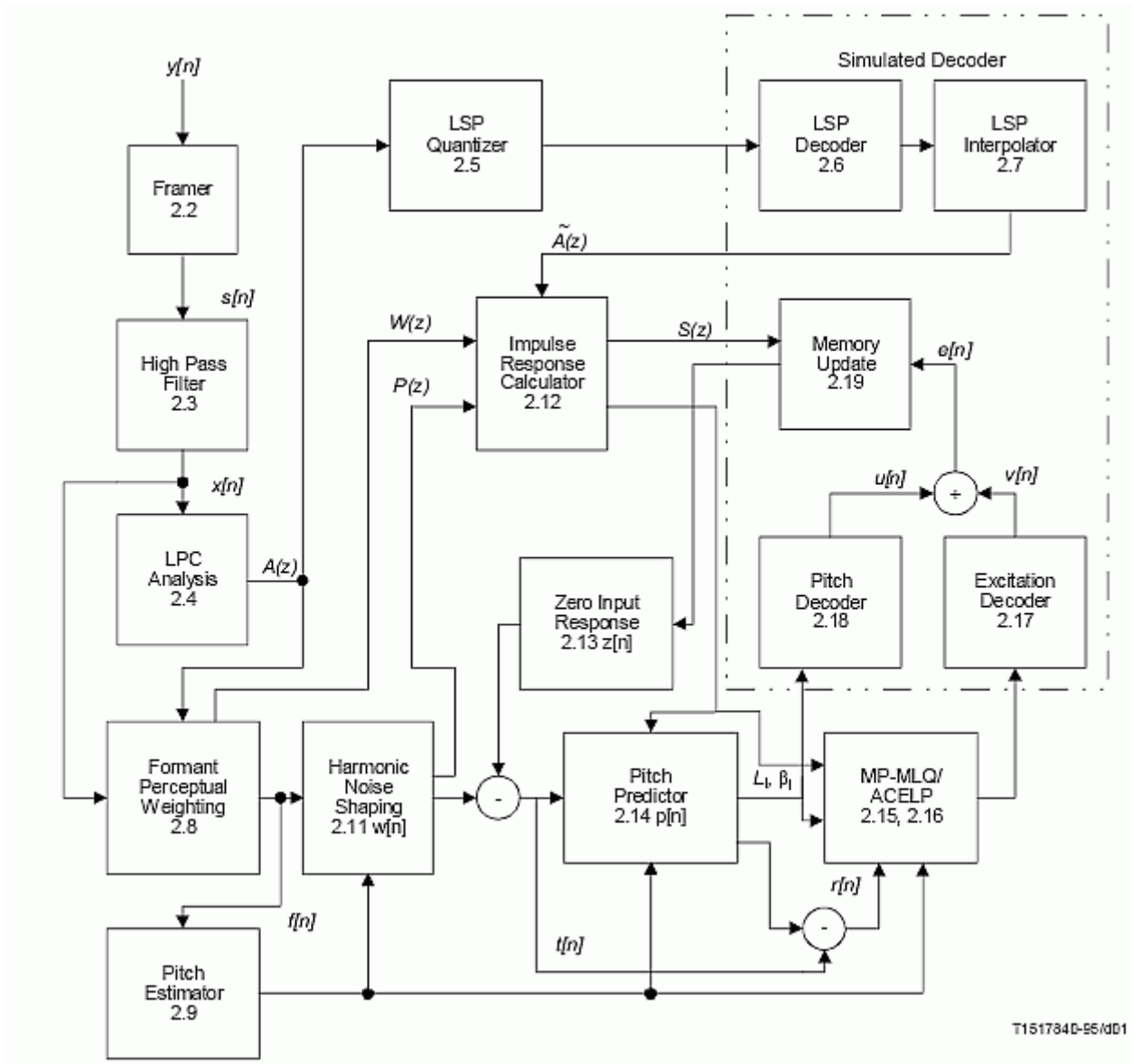
Kodek G.723.1 przeznaczony jest głównie do kodowania sygnału mowy w sytuacjach, w których jest on przesyłany jednocześnie z obrazem (wideotelefon, telekonferencje, itp). Pracuje on w dwóch trybach: pierwszy o przepływności 5,3 kbit/s wykorzystuje algorytm ACELP, drugi – o przepływności 6,3 kbit/s – algorytm MPC-MLQ (*Multipulse LPC with Maximum Likelihood Quantization*) [1,5,6]. Oba algorytmy różnią się między sobą m.in. strukturą algebraicznej książki

kodowej (która jest również inna niż w kodeku G.729). Nietypowe oznaczenie kodeka wynika z tego, że zastąpił on starszy standard G.723 o podobnym przeznaczeniu.

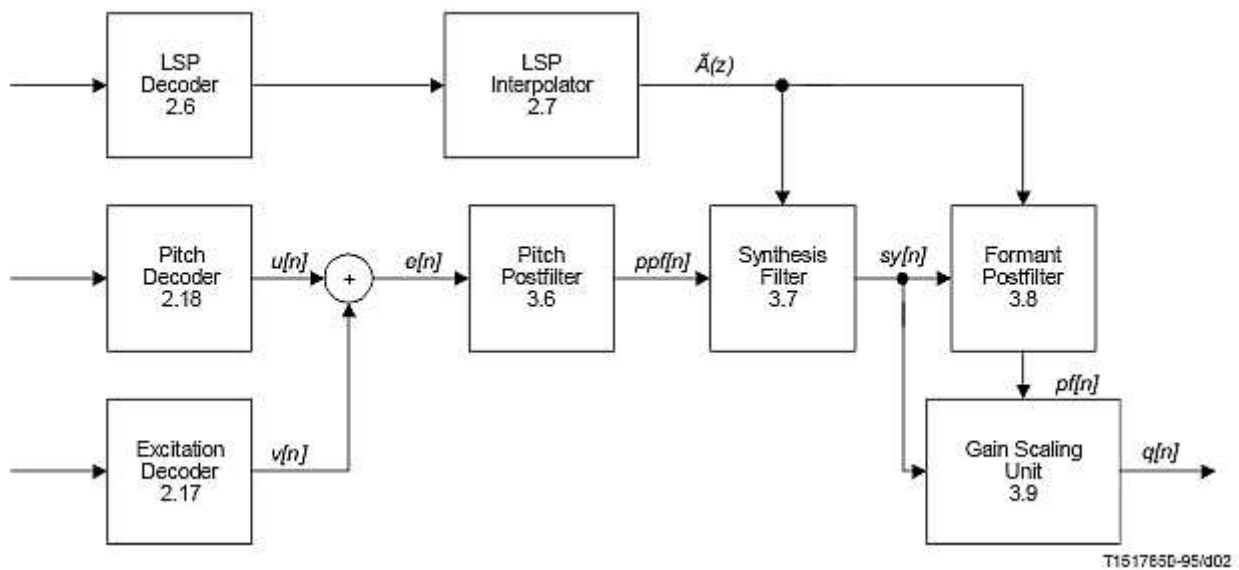
Zasada działania kodera G.723.1, przedstawiona na rys. 11, jest zbliżona do stosowanej w kodeku G.729. Sygnał jest dzielony na ramki o długości 30 ms (240 próbek), każda ramka dzielona jest na cztery podramki o równej długości. Dla każdej podramki wyznaczane są parametry LPC (analiza 10. rzędu) na podstawie sygnału wejściowego. Parametry LPC wyznaczone dla ostatniej z podramek są poddawane kwantyzacji przy użyciu algorytmu *predictive split vector quantization* (PSVQ). Parametry LPC nie poddane kwantyzacji służą do skonstruowania perceptualnego filtru ważącego, który przetwarza całą ramkę sygnału wejściowego. W kolejnym kroku, na podstawie przetworzonego sygnału, dla każdych dwóch podramek (120 próbek) wyznaczana jest wartość estymaty okresu sygnału (w pętli otwartej). Na podstawie wstępnie wyznaczonego okresu sygnału konstruowany jest filtr kształtujący szum harmoniczny (*harmonic noise shaping filter*). Następnie wyznaczana jest odpowiedź impulsowa kaskadowego połączenia trzech filtrów: syntetyzującego, perceptualnego filtru ważącego oraz kształtującego szum. Estymata okresu sygnału oraz odpowiedź impulsowa filtrów są używane do wyznaczenia okresu sygnału w pętli zamkniętej, jako odchyłki od uzyskanej wcześniej estymaty. W ostatnim kroku wyznaczana jest nieokresowa składowa pobudzenia na podstawie jednego z algorytmów, zależnie od wybranego trybu (MP-MLQ lub ACELP).

W dekodерze G.723.1, na podstawie zdekodowanych wartości indeksu wektora kodowego algebraicznego i adaptacyjnego, odtwarzane są sygnały pobudzające, które po skalowaniu, sumowaniu i przetworzeniu przez końcowy filtr wysokości dźwięku (*pitch postfilter*) podawane są na wejście filtru syntetyzującego. Współczynniki tego filtru wyznaczane są na podstawie zdekodowanych i interpolowanych wartości współczynników LPC. Wynikowy sygnał jest przetwarzany przez końcowy filtr formantowy (*formant postfilter*). Oba filtry końcowe kształtują sygnał w taki sposób, aby poprawić jego jakość.

Warto jeszcze wspomnieć, że standard G.723.1 przewiduje również funkcję odzyskiwania utraconych pakietów (*packet loss concealment*). Nie służy do tego pojedynczy, wydzielony blok, ale algorytm dekodera został zaprojektowany w taki sposób, że można do pewnego stopnia wypełnić „zagubione” odcinki dekodowanego sygnału [6].



Rys. 11. Schemat blokowy kodera G.723.1



Rys. 12. Schemat blokowy dekodera G.723.1

1.6 Kodek iLBC

Kodek iLBC oparty jest na algorytmie CELP. Jest on wart uwagi m.in. z tego powodu, że jest on dostępny na licencji *freeware*, czyli może być wykorzystywany znacznie swobodniej niż kodeki serii G objęte wysokimi opłatami licencyjnymi. Kody źródłowe kodeka są powszechnie dostępne w sieci Internet. Kodek iLBC jest wykorzystywany w wielu komunikatorach działających w sieci IP, m.in. w popularnym programie *Skype*.

Kodek iLBC może pracować w dwóch trybach: z przepływnością bitową 13,3 kbit/s (długość ramki 30 ms) lub z przepływnością 15,2 kbit/s (długość ramki 20 ms) [7]. Główną różnicą pomiędzy kodekiem iLBC a innymi kodekami opartymi na CELP jest to, że kodek ten został dostosowany do sytuacji, w której następuje utrata części pakietów [8]. Sytuacja taka często występuje np. w sieciach VoIP. W kodeku iLBC stosuje się algorytm długookresowego kodowania predykcyjnego niezależnego od ramki (*frame-independent long-term predictive coding*). W klasycznych kodekach CELP adaptacyjna książka kodowa jest wypełniana sygnałami pobudzającymi przed rozpoczęciem kodowania. Podejście takie może być źródłem zniekształceń sygnału, np. gdy część pakietów zostanie utracona lub zniekształcona, książki kodowe w koderze i dekoderze mogą się różnić. W kodeku iLBC stosuje się adaptacyjną książkę kodową do próbek poprzednich oraz następnych (adaptacja w przód i wstecz). Z ramki sygnału wyodrębniany jest wektor startowy, na podstawie maksimum energii sygnału rezydualnego. Wektor ten jest stanem początkowym do długookresowego kodowania predykcyjnego. Położenie i postać wektora startowego są kodowane dla każdej ramki. Adaptacyjna książka kodowa jest w pierwszym kroku wypełniana zdekodowanym wektorem startowym. Następnie ta książka kodowa jest wykorzystywana do długookresowej predykcji w przód (próbki sygnału od wektora startowego do końca ramki). Podczas kodowania, adaptacyjna książka kodowa jest na bieżąco uzupełniania dekodowanymi fragmentami sygnału. Następnie do książki kodowej wprowadzane są: zdekodowany wektor startowy oraz pierwszy zakodowany segment sygnału. Adaptacyjna książka kodowa jest teraz wykorzystywana do predykcji długookresowej wstecz, tzn. dla próbek od wektora startowego do początku ramki. Dzięki temu, na zawartość adaptacyjnej książki kodowej w dekoderze nie mają wpływu utracone lub zniekształcone pakiety. Jakość zdekodowanego sygnału w kodeku iLBC zależy przede wszystkim od dokładności wyznaczenia wektora startowego [8].

Koder iLBC działa następująco (opis dotyczy trybu 13,3 kbit/s). Sygnał jest dzielony na ramki o długości 240 próbek. Z ramki tej wybierane są za pomocą okna czasowego dwie podramki: jedna z początkowego fragmentu ramki, druga z końcowego fragmentu. W obu podramkach wyznaczane są parametry LPC. Oba zestawy parametrów są interpolowane i służą do wyznaczenia parametrów filtru. Sygnał rezydualny otrzymany po filtracji jest dzielony na podramki o długości

40 próbek. Znajdywane są dwie podramki o największej energii (80 próbek). Z tych dwóch podramek wybieranych jest 57 pierwszych lub 57 ostatnich próbek (w zależności od tego, który zestaw próbek charakteryzuje się większą energią), które są kodowane jako wektor startowy. Następnie adaptacyjna książka kodowa jest inicjalizowana zdekodowanym wektorem startowym w celu zakodowania pozostałych 23 próbek z dwóch wyznaczonych podramek. Do kodowania dalszych podramek (za wektorem startowym) wykorzystywane są wszystkie zdekodowanych 80 próbek. Podobnie dokonywane jest kodowanie wcześniejszych podramek (przed wektorem startowym). Adaptacyjna książka kodowa jest zatem wykorzystywana w trzech etapach, w każdym kolejnym etapie uzyskiwana jest dokładniejsza reprezentacja każdej z podramek. Pozostaje jeszcze wyznaczenie i zakodowanie wartości wzmocnienia korekcyjnego.

Dekoder iLBC dokonuje najpierw zdekodowania wektora startowego. Następnie dekodowane są późniejsze podramki (za wektorem startowym), a w kolejnym kroku wcześniejsze podramki (przed wektorem startowym). Tak uzyskany sygnał jest przetwarzany przez filtr końcowy wysokości dźwięku (*pitch postfilter*), a następnie przez filtr syntetyzujący. Jeżeli wykryta zostanie utrata pakietu, uruchamiane są odpowiednie procedury PLC (*packet loss concealment*).

Porównanie dokonane przez autorów kodeka iLBC wykazało, że w sytuacji, w której nie występuje utrata pakietów, subiektywna jakość sygnału mowy w kodeku iLBC jest porównywalna z kodekami G.723.1 i G.729. Wraz ze wzrostem liczby traconych pakietów, jakość sygnału dla kodeków serii G znacząco maleje, natomiast dla kodeka iLBC spadek jakości sygnału jest mniejszy (różnica na korzyść kodeka iLBC wzrasta ze zwiększaniem się liczby traconych pakietów) [7].

1.7 Kodek Speex

Kodek Speex, podobnie jak iLBC, należy do grupy algorytmów darmowych, przy czym jest on dostępny na licencji *open source*, czyli może być wykorzystywany we własnym oprogramowaniu bez ponoszenia kosztów. Kodek ten, choć jest oparty na algorytmie CELP, różni się znacząco od innych kodeków tego typu. Oferuje on trzy tryby pracy o różnej częstotliwości próbkowania: wąskopasmowy (8 kHz), szerokopasmowy (16 kHz) i ultra-szerokopasmowy (32 kHz). Posiada funkcje kodowania sygnału ze zmienną przepływnością (VBR – *variable bitrate*) oraz kodowania sygnałów stereofonicznych. Wyposażony jest w algorytmy wykrywania sygnału mowy (VAD – *Voice Activity Detection*), odzyskiwania pakietów (*packet loss concealment*) i wykrywania przerw w transmisji (DTX – *Discontinus Transmission*). Z uwagi na różnorodność trybów pracy, możliwe do uzyskania wartości przepływności bitowej mieszczą się w zakresie 2,15–24,6 kbit/s dla trybu wąskopasmowego i 4,0–44,2 kbit/s dla trybu szerokopasmowego. Duża funkcjonalność nie

powoduje znaczącego wzrostu opóźnień kodowania (są one porównywalne z kodekiem G.723.1). Głównym przeznaczeniem kodeka Speex są aplikacje VoIP [9].

Autorzy nie opublikowali pełnej specyfikacji kodeka, ale dostępne są kody źródłowe, które można przeanalizować. Z informacji zamieszczonych na stronie internetowej projektu wynika, że kodek Speex oparty jest na algorytmie CELP, z wykorzystaniem adaptacyjnej i stałej książki kodowej. W trybie wąskopasmowym (częstotliwość próbkowania 8 kHz), analiza sygnału dokonywana jest w ramkach o długości 20 ms (160 próbek). Ramki są dzielone na 4 podramki o jednakowej długości. Analiza LPC jest przeprowadzana w podramkach, z wykorzystaniem interpolacji wartości LSF. Parametry LPC wyznaczone dla czwartej podramki są kodowane z wykorzystaniem kwantyzacji wektorowej. Kodek Speex używa perceptualnego filtra ważącego. W dekodерze stosowany jest blok o nazwie *perceptual enhancement*, który prawdopodobnie działa jako filtr końcowy (*postfilter*). Na podstawie tego szcątkowego opisu można stwierdzić, że kodek Speex oparty jest na klasycznym algorytmie CELP, unowocześnionym m.in. przez wprowadzenie kwantyzacji wektorowej (Speex nie wykorzystuje algorytmu ACELP ze względów patentowych).

W trybie szerokopasmowym (częstotliwość próbkowania 16 kHz) kodek Speex działa następująco. Pasma częstotliwości jest dzielone na dwa zakresy (0–4 kHz i 4–8 kHz) za pomocą filtrów kwadraturowych (QMF). Sygnał z dolnego pasma jest kodowany w taki sam sposób, jak sygnał wąskopasmowy. W paśmie wyższym nie jest przeprowadzana detekcja wysokości dźwięku (nie jest ona celowa w tym zakresie częstotliwości), pozostałe różnice dotyczą liczby bitów przypadających na poszczególne parametry.

Kodek Speex wyróżnia się na tle innych kodeków mowy możliwościami, jednak nie przeprowadzono dotąd testów odsłuchowych pozwalających ocenić jakość sygnału mowy uzyskaną przy pomocy tego kodeka. Trzeba jednak pamiętać, że kodek Speex jest przez cały czas rozwijany i udoskonalany.

BIBLIOGRAFIA

1. Chu W.C., *Speech Coding Algorithms. Foundation and Evolution of Standardized Coders*, John Wiley & Sons, Hoboken 2003.
2. Goldberg R., Riek L., *A Practical Handbook of Speech Coders*, CRC Press, Boca Raton 2000.
3. ITU-T Recommendation G.728, *Coding of Speech at 16 kbit/s Using Low-Delay Code Excited Linear Prediction*, Geneva 1992.
4. ITU-T Recommendation G.729, *Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic-Code Excited Linear Prediction (CS-ACELP)*, Geneva 1996.
5. ITU-T Recommendation G.723.1, *Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 kbit/s*, Geneva 1996.

6. Kabal P., *ITU-T G.723.1 Speech Coder: A Matlab Implementation*, Department of Electrical & Computer Engineering, McGill University, 2004.
7. Global IP Sound, *iLBC – Designed for the Future (iLBC Whitepaper)*, www.globalipsound.com.
8. Andersen S.V., Kleijn W.B., Hagen R. *et al.*, *iLBC – a Linear Predictive Coder with Robustness to Packet Loss*, IEEE 2002 Workshop on Speech Coding, Tsukuba 2002.
9. Valin J.M., *The Speex Codec Manual*, www.speex.org.
10. Kulesza M, *Opracowanie architektury kodeka mowy na potrzeby telefonii VoIP*, Raport Wewnętrzny, Katedra Systemów Multimedialnych PG, Gdańsk 2005.
11. Sen D., Holmes W.H., *Perceptual Enhancement of CELP Speech Coders*, ICASSP-94, Adelaide 1994.
12. Kubin G., Bastiaan Kleijn W., *On Speech Coding in a Perceptual Domain*, ICASSP-99, Phoenix 1999.
13. Tang B., Shen A., Alwan A., Pottie G., *A Perceptually Based Embedded Subband Speech Coder*, IEEE Transactions on Speech and Audio Processing, vol. 5, no. 2, 1997.
14. Najafzadeh-Azghandi H., Kabal P., *Perceptual Coding of Narrowband Audio Signals at 8 kbit/s*, Proc. IEEE Workshop Speech Coding, Pocono Manor 1997.
(<http://www.tsp.ece.mcgill.ca/Kabal/papers/P1997.html>)