# ROZDZIAŁ 6. METODY PREDYKCYJNE

Metody predykcyjne mogą być traktowane jako uzupełnienie entropijnych metod kodowania danych, jak również metod słownikowych. Podobnie jak metody słownikowe, są one techniką zwiększenia skuteczności algorytmów bezstratnej kompresji danych poprzez wstępną redukcję nadmiarowości pierwotnej reprezentacji danych w strumieniu. Predykcja wykorzystuje różne modele lokalnych zależności danych, czy też ich skorelowania bazując na możliwie szerokiej wiedzy dostępnej *a priori* lub też zbieranej we wstępnym etapie analizy strumienia, często też adaptacyjnie modyfikowanej. Mające swoje ograniczenia metody entropijne mogą być następnie, tj. przy znacznie uproszczonej strukturze nadmiarowości w strumieniu, efektywniej użyte do kodowania pośredniej reprezentacji wykorzystując wiarygodne modele probabilistyczne.

W kolejnych punktach tego rozdziału przedstawiona została podstawowa koncepcja predykcji, zarówno liniowej jak i nieliniowej, oraz sposoby konstrukcji skutecznych schematów predykcji, dobrze modelujących zależności danych w strumieniu, zarówno w wersji statycznej jak i adaptacyjnej.

## 6.1. Wprowadzenie w zagadnienie predykcji

Zasadniczą koncepcją metod predykcyjnych jest konstrukcja modelu, który potrafi przewidzieć kolejne dane ze strumienia wejściowego na podstawie dostarczonej wcześniej informacji charakteryzującej kodowany zbiór danych. Informacja ta musi być zapisana w strumieniu wyjściowym kodera, aby umożliwić proces dekodowania. Do modelu przewidującego (predykcji) powtarzanego w dekoderze mogą być włączone jedynie dane już zdekodowane, czyli obowiązuje zasada przyczynowości.

Zasada przewidywania kolejnej wartości na wejściu kodera nie jest nowa i występuje właściwie we wszystkich prezentowanych dotąd technikach kodowania. Realizowana jest przez modele statystyczne koderów entropijnych, bądź też przez kontekstowo-treściowe przewidywanie przy pomocy słownika. Przewidywanie służy tutaj konstrukcji optymalnego kodu binarnego opisującego jedynie rzeczywistą informację, nie dającą się wydedukować na podstawie wiedzy dostępnej a priori i zakodowanych już sekwencji symboli. Efektywność kodowania gwałtownie rośnie, gdy przewidywanie jest poprawne, natomiast w fazie słabszej skuteczności predykcji generowany jest dłuższy strumień kodowy. Jednocześnie, w rozwiązaniach dynamicznych następuje adaptacja modelu do charakteru kodowanej właśnie partii danych. Wyróżnikiem metod predykcyjnych jest koncentracja jedynie na możliwie doskonałym modelowaniu kodowanego zbioru danych, bez propozycji skutecznej realizacji fazy binarnego kodowania w algorytmie kompresji. Przewidywanie to w fazie optymalizacji modelu predykcji nabiera w dużym stopniu charakteru statystycznego, w odróżnieniu od metod słownikowych z czysto deterministycznym określaniem identyczności ciągu symboli wejściowych z frazą słownika. Ponadto, poszukiwana jest zależność funkcyjna danych w pewnym kontekście, przy czym źródło informacji przybliżane jest modelem wieloparametrycznym, zamiast prostego mechanizmu rejestracji powtórzeń określonych ciągów danych lub też modelu statystycznego na bazie prawdopodobieństw warunkowych.

Przy budowie modelu predykcji liczy się w pierwszej kolejności jego niezawodność, tj. na ile proces przewidywania w każdym kroku kończy się sukcesem. Idealny model predykcji jest skuteczny w 100%, nie sposób go jednak zbudować dla rzeczywistego zbioru

danych z niezerową informacją. Drugą istotną kwestią jest bowiem możliwie mała złożoność takiego modelu, pozwalająca na oszczędny zapis jego parametrów jako wyjściowej reprezentacji kodera. Okazuje się, że taki idealny model kodowania predykcyjnego jest niepraktyczny ze względu na bardzo złożoną postać modeli cechujących się dużą niezawodnością. Rezygnacja z 'doskonałego' przewidywania powoduje konieczność oszacowania różnicy pomiędzy wartością rzeczywistą i przewidywaną, a następnie umieszczenia wartości różnicowych w strumieniu wyjściowym kodera, obok parametrów modelu predykcji. Jakkolwiek istotnym problemem teoretycznym jest efektywna równowaga pomiędzy złożonością opisu modelu a prostotą statystyki kodowanego zbioru wartości różnicowych, to jednak ze względów praktycznych przydatne okazują się przede wszystkim bardzo proste modele liniowe, a opis parametryczny tych modeli zajmuje pomijalnie małą część strumienia wyjściowego. Strumień ten zawiera więc przede wszystkim wartości różnicowe, przy czym dla dobrze dobranego modelu przeważają wartości bliskie zeru. Powoduje to, iż metody predykcyjne są często nazywane metodami kodowania różnicowego (ang. differential encoding).

*Predykcja ze statystycznym modelem prawdopodobieństw warunkowych*

Problem dużej złożoności modelu predykcji pojawia się przy wykorzystaniu prawdopodobieństw warunkowych do przewidywania kolejnej wartości strumienia. Rozważmy następujący przykład. W opisie zbioru danych przy pomocy źródła Markowa rzędu $m$ wartość danej ($b$ - bitowa) zależy jedynie od $m$ wartości ją poprzedzających, czyli prawdopodobieństwo wystąpienia każdego symbolu z alfabetu $A_X = \{a_1, a_2, ..., a_{2^b}\}$ jest określone przez zbiór prawdopodobieństw warunkowych, określonych na $m$ -elementowym kontekście $P(x_k = a_i \mid x_{k-1}, x_{k-2}...., x_{k-m}),\ 1 \le i \le 2^b$ Wykorzystano tutaj oznaczenie $x$ dla zmiennej losowej opisującej źródło informacji dla podkreślenia charakteru niewiadomej, która ma być przewidywana na podstawie schematu predykcji. Mając określony przez kontekst stan źródła, jedna z $2^b$ możliwych wartości alfabetu odpowiada wartości rzeczywistej, kodowanej w danym kroku algorytmu. Zbiór prawdopodobieństw wystąpienia każdej z tych wartości, określony rzetelnie w modelu prawdopodobieństw warunkowych pozwala zdecydowanie wskazać pewne symbole (dane, wartości), które mają największe szanse pojawienia się teraz w strumieniu, przy praktycznym braku jakichkolwiek możliwości wystąpienia innych elementów alfabetu. Gdyby ten model statystyczny był 'nieomylny', wówczas najbardziej prawdopodobny symbol alfabetu pokrywałby się dokładnie z wartością, która pojawi się za chwilę w strumieniu wejściowym. Model zawierałby wtedy pełną informację o źródle, a jego parametry stanowiłyby zakodowaną reprezentację strumienia danych.

W fazie kodowania można więc próbować zbudować taki model statystyczny, mając do dyspozycji wszystkie dane ze strumienia kodowanego, który w każdym kroku mając przyczynowy kontekst potrafi jednoznacznie przewidzieć jako najbardziej prawdopodobną daną, która za chwilę pojawi się na wejściu bez żadnej pomyłki. Można to uzyskać niemal w każdym przypadku, jednak kosztem znaczącego zwiększenia, odpowiednio do stopnia złożoności zawartej w zbiorze danych informacji, rzędu $m$ modelu. Wtedy jednak musimy zakodować początkowy stan modelu statystycznego o $2^{b \cdot m}$ wartościach prawdopodobieństw warunkowych, jak również pewne parametry ewentualnej adaptacyjnej modyfikacji modelu w trakcie procesu kodowania. Zapisanie tak dużej liczby danych charakteryzujących model czyni ten algorytm niepraktycznym. Ponadto, konstrukcja takiego modelu jest bardzo czasochłonna i nie zawsze pozwala uzyskać stuprocentową skuteczność. Przypadki nietrafnej predykcji trzeba dopisać do kodu wyjściowego, aby uzyskać odwracalną kompresję.

W celu praktycznej implementacji tej metody potrzebne są pewne uproszczenia. Mogą one prowadzić do wyznaczenia i zapisywania jako reprezentacji wyjściowej różnicy pomiędzy wartością najbardziej prawdopodobną dla danego kontekstu a wartością rzeczywistą. Inną modyfikacją powyższego mechanizmu przewidywania jest zapisanie w kolejnych krokach indeksów trafnej predykcji przy pomocy modeli statystycznych, co pozwoli jednoznacznie odtworzyć zbiór oryginalny w dekoderze. W tym przypadku konieczne jest jednak także przekazanie modelu źródła o $2^{b \cdot m}$ stanach (lub nieco uproszczonego), co bardzo ogranicza efektywność takiego rozwiązania. Budowanie adaptacyjne tak złożonego modelu, zaczynając od prostej postaci początkowej oraz wykorzystując jedynie kontekst przyczynowy, jest zbyt mało efektywne (rozrzedzenie kontekstu), by można było uzyskać wysoką jakość przewidywania.

Można jednak osłabiając nieco warunek niezawodności zbudować prostsze modele opisujące zależności danych w kompresowanym pliku, które znacznie redukują poziom nadmiarowości w zbiorze tworząc nową jego reprezentację. Kodowanie tej reprezentacji, zamiast oryginału, przy pomocy metod entropijnych czy słownikowych pozwala w wielu przypadkach zwiększyć skuteczność kompresji i jeszcze bardziej zbliżyć się do granicy efektywności wyznaczanej przez entropię łączną.

*Predykcja z funkcją zależności danych w lokalnym kontekście*

Pomysł polega na całkowitej rezygnacji ze znajomości prawdopodobieństw warunkowych. Przewidywanie kolejnego symbolu danego źródła na wejściu odbywa się w kontekście jego wystąpienia jako wyznaczenie wartości funkcji kilku poprzednich danych ze strumienia, należącej oczywiście do alfabetu źródła informacji. Potencjalnie wszystkie poprzednio kodowane dane mogą być wykorzystane w procesie predykcji, jednak w praktycznych rozwiązaniach okazuje się, że modele wykorzystujące zaledwie kilku najbliższych sąsiadów w przestrzeni określoności danych oryginalnych dają wystarczającą skuteczność, a powiększanie rzędu modelu nie zawsze powoduje znaczącą poprawę efektywności kompresji. Zwiększa się natomiast czas obliczeń i złożoność opisu modelu.

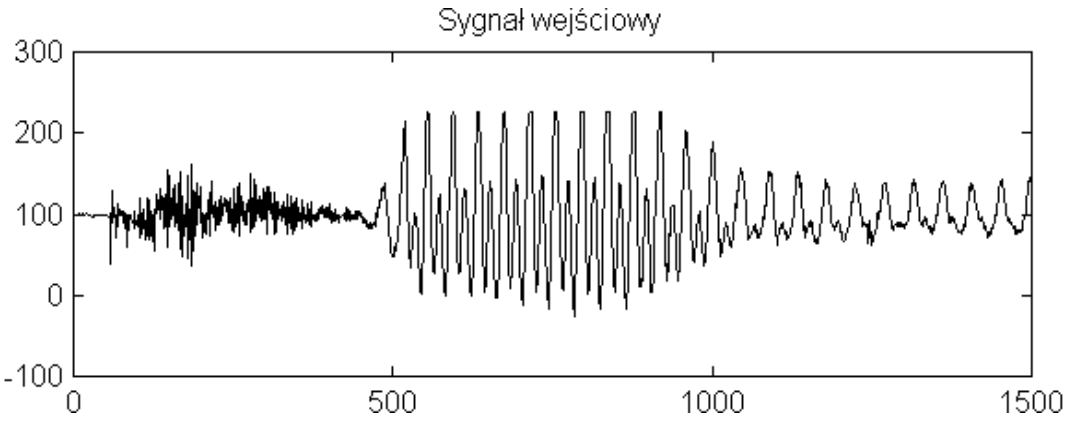
Wartość przewidywaną na podstawie modelu predykcji opisanego funkcją $f(\cdot)$ dla *k*-tej danej w strumieniu wejściowym, przy założeniu źródła informacji w postaci modelu Markowa, przedstawia następujące równanie

$$\hat{x}_k \doteq f(x_{k-1}, x_{k-2}, ..., x_0) = f(x_{k-1}, x_{k-2}, ..., x_{k-m}), \tag{6.1}$$

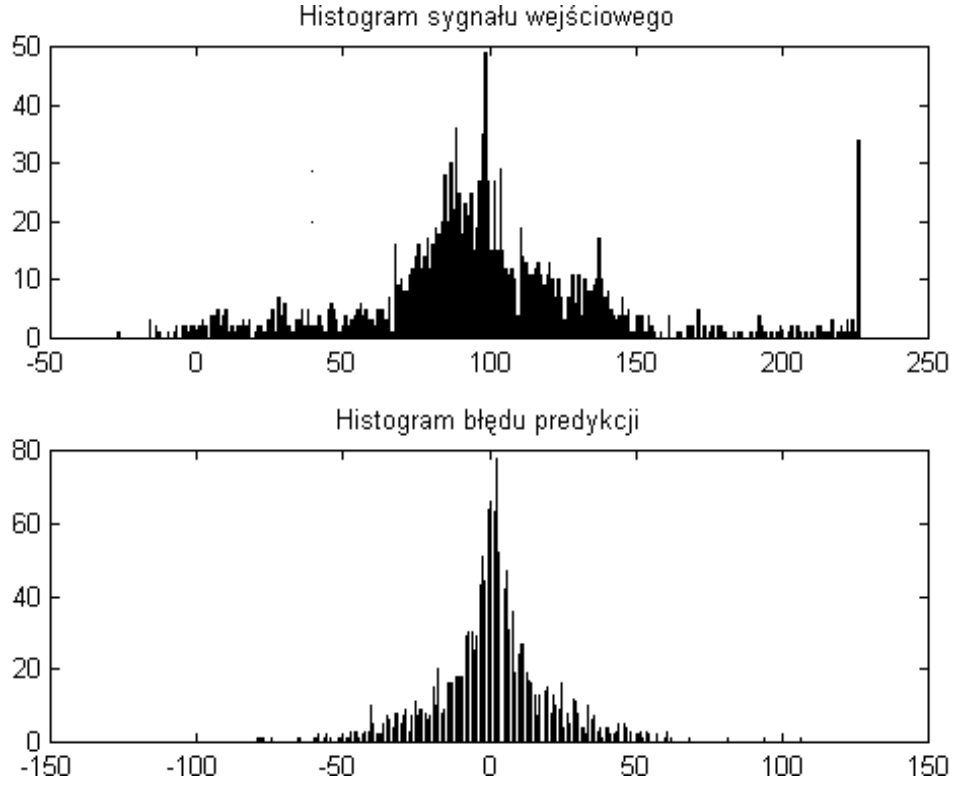
gdzie *m* - rząd modelu (rozmiar kontekstu), $\hat{x}_k \in A_X$ ($A_X$ - alfabet źródła danych wejściowych $x_i$). Oczywiście, przewidywana wartość danej $\hat{x}_k$ nie musi pokrywać się w każdym przypadku z wartością rzeczywistą $x_k$. Różnica tych wartości jest nazywana błędem predykcji $e_k = x_k - \hat{x}_k$. Zbiór wartości różnicowych stanowi nową reprezentację zbioru oryginalnego, bardziej podatną na binarne kodowanie. Rozkład wartości różnicowych przy poprawnie dobranym modelu predykcji przyjmuje charakterystyczną postać. Pokazuje to przykład 6.1. Trzeba zauważyć, że możliwa dynamika wartości błędu jest o jeden bit większa (czyli dwukrotnie) od dynamiki wartości $x_k$. Stąd też praktyczne algorytmy kodują np. oddzielnie wartość bezwzględną, oddzielnie znak lub też przy kodowaniu wartości różnicowych uwzględniają zwiększoną dynamikę budując np. koder Huffmana dla danych 9-cio bitowych.

PRZYKŁAD 6.1. Analiza histogramu sygnału oryginalnego i błędu predykcji.

Zarejestrowany sygnał ma postać jak na rys. 6.1. Narysujmy histogram wartości tegoż sygnału, jak również histogram wartości błędu predykcji przy zastosowaniu prostego modelu predykcji $\hat{x}_k = 0.89x_{k-1} + 0.63x_{k-2} - 0.54x_{k-3}$. Histogramy te zostały przedstawione na rys.6.2.




Rys. 6.1. Przykładowy sygnał dźwięku ze składową stałą na poziomie wartości 100, poddany predykcji.




Rys. 6.2. Histogram sygnału wejściowego z rys. 6.1 oraz histogram błędu predykcji przy zastosowaniu modelu predykcji liniowej: $\hat{x}_k = 0.89x_{k-1} + 0.63x_{k-2} - 0.54x_{k-3}$.

Histogram błędu predykcji z rys. 6.2 ma charakterystyczną postać z dominującą liczbą zer i wartości bliskich zeru, co świadczy o niewielkich błędach przewidywania w większości przypadków. Występują jednak także większe wartości błędu o wartości bezwzględnej ponad 50, lecz są to jedynie pojedyncze przypadki. Histogram wartości błędu różni się wyraźnie od histogramu sygnału oryginalnego, a jego symetryczność względem zera świadczy o tym, że model predykcji jest nieobciążony.

Generalnie rozkład wartości błędu predykcji można skutecznie przybliżyć wykładniczym rozkładem Laplace'a postaci: $h_\sigma(e) = \frac{1}{\sqrt{2}\sigma} \exp\left(\frac{-\sqrt{2}\,|e|}{\sigma}\right)$. Zakładamy wartość średnią rozkładu błędu równą 0, a miarą kształtu rozkładu jest wartość odchylenia standardowego. Można więc rozkład błędu predykcji dla różnych modeli predykcji scharakteryzować przy pomocy jednego parametru, co jest bardzo wygodne w kodowaniu ze względu na oszczędną postać opisu takiego rozkładu.

Najczęściej stosowanym schematem predykcji, który pozwala przy dużej prostocie modelu i małych kosztach obliczeniowych uzyskać zadawalający efekt przewidywania jest model predykcji liniowej. Określa go równanie:

$$f(x_{k-1}, x_{k-2}, ..., x_{k-m}) = \sum_{i=1}^{m} \alpha_i x_{k-i} \tag{6.2}$$

Jest on podatny na dynamiczne dopasowanie do lokalnych własności zbioru, łatwy w realizacji algorytmicznej. Wyznaczanie wartości przewidywanych jako liniowej kombinacji *m* poprzednich wartości danych jest chwilami dalekie od optymalnego, jednak dla zbiorów danych o znaczącym stosunku sygnału do szumu i istotnych korelacjach występujących w znacznych obszarach kodowanego strumienia danych może dać zaskakująco dobre rezultaty.

Nieliniowa postać funkcji $f(\cdot)$ o charakterze wielomianowym stopnia *n*, określona w sposób następujący:

$$f(x_{k-1}, x_{k-2}, ..., x_{k-m}) = \sum_{i=1}^{m} (\alpha_i^{(1)} x_{k-i} + \alpha_i^{(2)} x_{k-i}^2 + ... + \alpha_i^{(n)} x_{k-i}^n), \tag{6.3}$$

pozwala teoretycznie dobrze w większości przypadków opisać zależności w zbiorze danych i dokonać predykcji z niewielkim błędem. Okazuje się jednak, że często jest to model zbyt złożony i trudny w projektowaniu, a zadanie optymalizacji takiego predyktora, przy ewentualnej próbie adaptacyjnej jego modyfikacji staje się trudnym i kosztownym zadaniem obliczeniowym. Uzyskiwane rezultaty zaś nie są adekwatne do nakładów.

## 6.2. Liniowa predykcja DPCM

Schemat kodowania predykcyjnego prowadzący do praktycznych implementacji nazywany jest DPCM (ang. differential pulse code modulation), opracowany w Bell Laboratories [1]. W zdecydowanej większości wykorzystuje się tutaj prosty model liniowy [2,3], chociaż predyktory nieliniowe niskich rzędów także mogą znaleźć praktyczne zastosowanie [4].

Wyznaczona przez model predykcji wartość jest odejmowana od rzeczywistej wartości kolejnych danych tworząc różnicowy strumień danych o dużo mniejszej korelacji pomiędzy wartościami sąsiednich danych w stosunku do zbioru oryginalnego. Konkretny model predykcji liniowej (ustalone wartości współczynników) może być stały dla całej grupy zbiorów danego typu (globalna predykcja), może zmieniać się w zależności od zbioru

(lokalna predykcja), co jest szczególnie skuteczne w przypadku zbiorów charakteryzowanych stacjonarnymi źródłami informacji, a także może być dostosowany do lokalnej statystyki zbioru (predykcja adaptacyjna) w przypadku źródeł niestacjonarnych. Otrzymane w wyniku predykcji różnicowe zbiory danych koduje się wykorzystując najczęściej algorytmy Huffmana i kodowania arytmetycznego. Ponieważ rozkład wartości różnicowych może być dobrze przybliżony rozkładem Laplace'a, można z dużą efektywnością zastosować statyczny model entropijnego kodowania przy znacznych oszczędnościach czasowych.

Projektowanie efektywnych algorytmów DPCM sprowadza się do optymalizacji procesu predykcji. Kluczowym zagadnieniem staje się tutaj odpowiedni dobór kontekstu, tj. jego kształtu i rzędu, a także określenie wag poszczególnych elementów kontekstu przy wyznaczaniu wartości przewidywanych.

## Optymalizacja modelu predykcji liniowej

Model predykcji liniowej może być realizowany w zarówno w wersji statycznej, jak i adaptacyjnej. W algorytmie statycznym parametry modelu są stałe dla wielu zbiorów danych, pojedynczego zbioru czy jego znaczącej części i muszą być przekazane dekoderowi. Przy predykcji adaptacyjnej współczynniki są na bieżąco (punkt po punkcie) modyfikowanie w zależności od skuteczności przewidywania aktualnego modelu. Algorytm śledzi więc lokalne zmiany zależności danych w kodowanym strumieniu dobierając możliwie najlepszy schemat predykcji liniowej. Mechanizm zmian jest przyczynowy, co pozwala powtórzyć cały proces podczas dekodowania, bez konieczności zapisywania dodatkowej informacji w pliku wyjściowym.

*Wyznaczanie współczynników predykcji*

Przy wyznaczaniu współczynników predykcji $\alpha_i$ często wykorzystywane jest kryterium minimalizacji błędu średniokwadratowego. Formułując takie kryterium dla liniowego modelu predykcji określonej równaniem: $\hat{x}_k \doteq \sum_{i=1}^{m} \alpha_i x_{k-i}$, sprowadza się ono do zagadnienia minimalizacji sumy błędów predykcji w poszczególnych punktach danych $\varepsilon_p$ określonej wyrażeniem:

$$\varepsilon_p = \sum_{k=1}^{K} e_k{}^2 = \sum_{k=1}^{K} (x_k - \hat{x}_k)^2 = \sum_{k=1}^{K} \left( x_k - \sum_{i=1}^{m} \alpha_i x_{k-i} \right)^2, \tag{6.4}$$

gdzie *K* oznacza liczbę danych w zbiorze z mieszczącym się kontekstem rzędu *m*. Dla *m* pierwszych elementów zbioru stosuje się zazwyczaj uproszczony model predykcji odpowiednio niższego rzędu, a pierwsza dana ze zbioru kompresowana jest bez predykcji.

Tradycyjnie rozwiązanie takiego zagadnienia sprowadza się do konstrukcji układu równań normalnych, utworzonych poprzez przyrównanie do zera pochodnych cząstkowych wyrażenia na błąd predykcji po wszystkich współczynnikach $\alpha_i$:

$$\begin{aligned} \frac{\partial \varepsilon_p}{\partial \alpha_1} &= -2 \cdot x_{k-1} \cdot \sum_{k=1}^{K} \left( x_k - \sum_{i=1}^{m} \alpha_i x_{k-i} \right) = 0 \\ &\vdots \\ &\vdots \end{aligned} \tag{6.5}$$

$$\frac{\partial \varepsilon_p}{\partial \alpha_m} = -2 \cdot x_{k-m} \cdot \sum_{k=1}^{K} \left( x_k - \sum_{i=1}^{m} \alpha_i x_{k-i} \right) = 0$$

Porządkując nieco elementy po obu stronach równań można je zapisać w sposób następujący:

$$\begin{aligned} \sum_{k=1}^{K} \left( x_{k-1} \cdot \sum_{i=1}^{m} \alpha_i x_{k-i} \right) &= \sum_{k=1}^{K} x_k \cdot x_{k-1} \\ &\vdots \\ \sum_{k=1}^{K} \left( x_{k-m} \cdot \sum_{i=1}^{m} \alpha_i x_{k-i} \right) &= \sum_{k=1}^{K} x_k x_{k-m} \end{aligned} \qquad (6.6)$$

Trzeba więc rozwiązać układ $m$ równań liniowych z $m$ niewiadomymi. Aby łatwiej było znaleźć rozwiązanie, scharakteryzujmy to zagadnienie nieco inaczej.

Formułując zagadnienie minimalizacji błędu w kategoriach statystycznych należy stwierdzić, że analogicznie do (6.4) minimalizowane jest wyrażenie: $\varepsilon_p = E\left\{ \left( x_k - \sum_{i=1}^{m} \alpha_i x_{k-i} \right)^2 \right\}$, a układ równań wygląda następująco:

$$E\left\{ \left( x_k - \sum_{i=1}^{m} \alpha_i x_{k-i} \right) \cdot x_{k-j} \right\} = 0 \quad \text{dla} \quad j = 1,...,m. \qquad (6.7)$$

Przekształcając ten układ równań wykorzystamy pojęcie funkcji autokorelacji zmiennej losowej $X$ przyjmującej wartości $x_k$, która określona jest następująco: $R_X(i) = E\{x_k \cdot x_{k-i}\}$. Porządkując nieco wyrażenia w układzie równań 6.7 możemy zapisać je w sposób następujący:

$$\sum_{i=1}^{m} \alpha_i R_X(i-j) = R_X(i) \quad \text{dla} \quad i = 1,...,m \qquad (6.8)$$

lub w jeszcze bardziej wygodnej postaci macierzowej:

$$\boldsymbol{R_M} \cdot \boldsymbol{A} = \boldsymbol{R_K}, \qquad (6.9)$$

gdzie $\boldsymbol{R_M} = \begin{bmatrix} R_X(0) & R_X(1) & ... & R_X(m-1) \\ R_X(1) & R_X(0) & ... & R_X(m-2) \\ \vdots & \vdots & & \vdots \\ R_X(m-1) & R_X(m-2) & ... & R_X(0) \end{bmatrix}$, $\boldsymbol{A} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \end{bmatrix}$ i $\boldsymbol{R_K} = \begin{bmatrix} R_X(1) \\ R_X(2) \\ \vdots \\ R_X(m) \end{bmatrix}$.

Wykorzystano przy tym własność funkcji autokorelacji, taką że $R_X(-i) = R_X(i)$. Znając więc wartości funkcji autokorelacji $\{R_X(i)\}$ dla $i = 1,...,m$ można wyznaczyć wartości współczynników modelu liniowej predykcji z zależności:

$$\boldsymbol{A} = \boldsymbol{R_M}^{-1} \cdot \boldsymbol{R_K} \qquad (6.10)$$

Minimalizowaliśmy wariancję błędu przy założeniu stacjonarności kompresowanego sygnału, czyli niezmienności cech statystycznych (rozkładu, momentów) zbioru wartości tegoż błędu. To założenie jednak rzadko jest spełnione przy kompresji realnych zbiorów danych. Można by więc podzielić przykładowy zbiór na kilka części o własnościach w przybliżeniu stacjonarnych i poszukać metodą regresji kilku modeli predykcji odpowiednio przełączanych przy kompresji kolejnych partii danych. Jest to tzw. predykcja przełączana, kiedy to przy zastosowaniu pewnego kryterium (np. błąd predykcji większy od *T*) następuje przełączanie pomiędzy tymi modelami zwiększające skuteczność przewidywania w poszczególnych fragmentach sygnału wejściowego. Przykłady takich rozwiązań można znaleźć w [5,6]. Dalsza modyfikacja statycznego modelu predykcji prowadzi do koncepcji modeli adaptacyjnych, w których początkowa, np. optymalna w sensie globalnym, postać predyktora jest dynamicznie modyfikowana w zależności od lokalnych własności strumienia kodowanego lub też bieżącej efektywności predykcji.

W odniesieniu do wartości współczynników modeli predykcji istotnym ze względów praktycznych jest spełnienie dodatkowego warunku, wynikającego ze wspomnianej wcześniej cechy modeli przewidywania. Powinny być one budowane jako estymator nieobciążony, co w rozważanym przypadku sprowadza się do następującego równania: $E\{e_k\}=0$. Przekształćmy nieco lewą stronę tego wyrażenia w następujący sposób:

$$E\{e_k\}=E\{x_k-\hat{x}_k\}=E\{x_k-\sum_{i=1}^{m}\alpha_i\cdot x_{k-i}\}=E\{x_k\}-\sum_{i=1}^{m}\alpha_i\cdot E\{x_{k-i}\}=E\{x_k\}(1-\sum_{i=1}^{m}\alpha_i)\,. \quad (6.11)$$

Widać, że warunek nieobciążalności sprowadza się do alternatywy jak niżej:

$$E\{x_k\}=0 \qquad \text{lub} \qquad \sum_{i=1}^{m}\alpha_i=1. \quad (6.12)$$

Można więc usunąć składową stałą sygnału bez nakładania jakichkolwiek warunków na współczynniki. Nie jest to jednak możliwe w każdym przypadku. Wykonanie algorytmu kompresji dwuprzebiegowo, kiedy to następuje wstępny przegląd danych i obliczanie wartości średniej w pierwszym etapie oraz właściwa predykcja i kodowanie przesuniętego o wartość składowej sygnału w drugim, wymaga bowiem pełnej dostępności do strumienia danych przed kompresją. Dla przypadków, gdy nie jest to możliwe, np. przy kodowaniu na bieżąco transmitowanego strumienia danych z telekonferencji, pozostaje drugi warunek, że suma współczynników powinna być równa jeden. W zastosowaniach transmisyjnych w celu ograniczenia propagacji błędu zapewnia się zazwyczaj wartość sumy współczynników modelu predykcji nieco mniejszą od jedynki.

*Poszukiwanie odpowiedniego kontekstu*

Kontekst stosowany w modelu predykcji liniowej może również być zmieniany adaptacyjnie w procesie kodowania kolejnych fragmentów strumienia danych, jest to jednak rozwiązanie rzadziej spotykane. Ogólnie rzecz biorąc, postać kontekstu, zarówno co do wielkości jak i kształtu, powinna wynikać jednoznacznie z poziomu korelacji pomiędzy sąsiednimi wartościami danych w strumieniu. Zwiększania rozmiaru kontekstu, czyli rzędu modelu predykcji pociąga za sobą wzrost kosztów obliczeniowych, a także bufora do przechowywania splatanego wektora danych - obliczanie wartości przewidywanych może być bowiem realizowane jako splot strumienia danych z przesuwanym filtrem o współczynnikach $\alpha_i$. Jeżeli towarzyszy temu wyraźne obniżenie wariancji wartości błędów predykcji, to jest to uzasadnione. Pokazuje to przykład 6.2.

PRZYKŁAD 6.2. Dobór wielkości kontekstu w liniowym modelu predykcji.

Sygnał mowy z rys. 6.1 poddano kompresji bezstratnej z wykorzystaniem modeli predykcji liniowej o współczynnikach wyznaczonych metodą minimalizacji błędu średniokwadratowego (regresji liniowej). Zaobserwujemy wpływ rzędu predykcji na postać nowej reprezentacji - błędu predykcji - oraz na skuteczność kodowania wejściowego sygnału dźwięku.

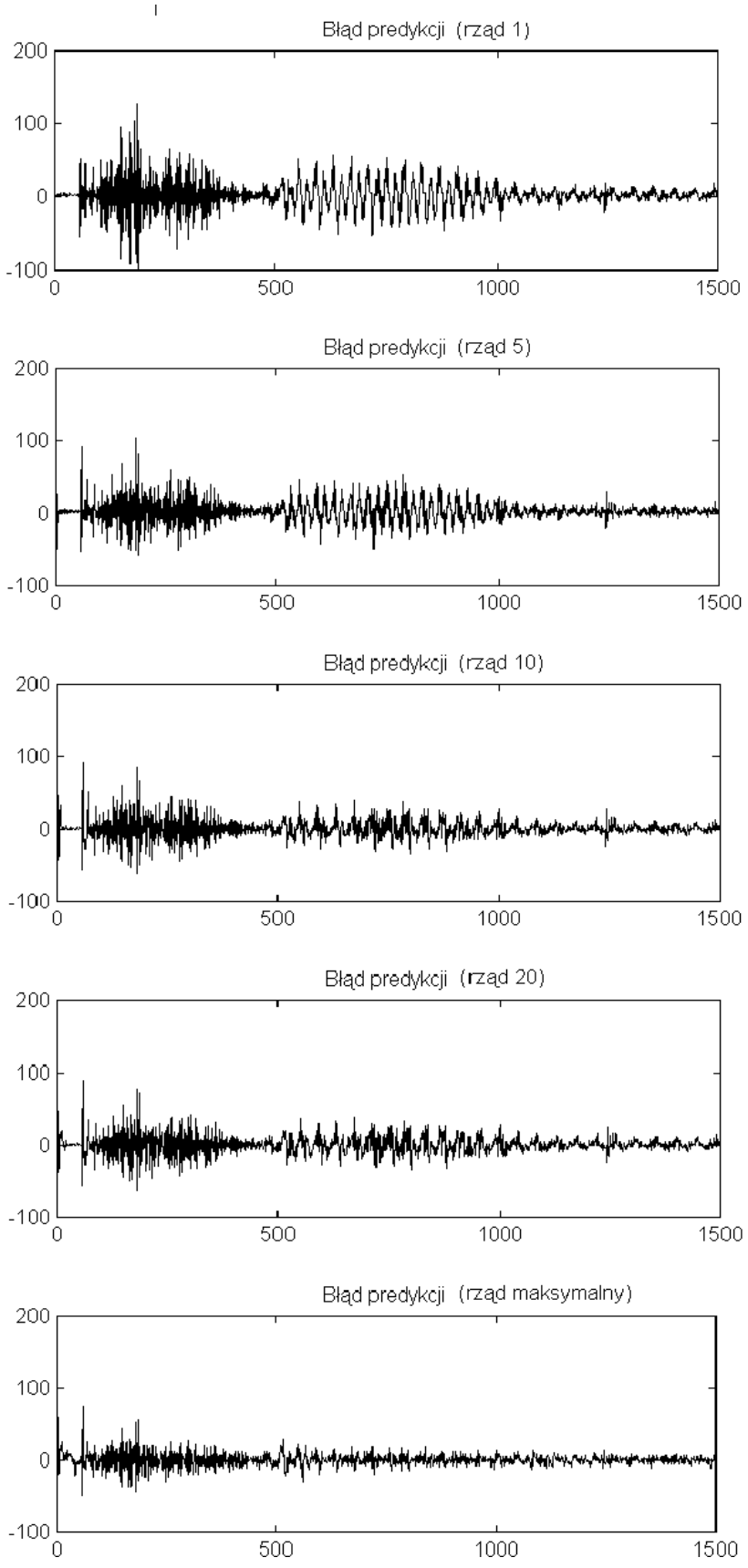
Minimalizując wariancję błędu predykcji obliczono współczynniki modeli różnych rzędów. Przykładowo, modele rzędu 1, 2, 3 i 5 wyglądały następująco:

$$\begin{aligned}
&\text{rząd 1: } \hat{x}_k = 0.98x_{k-1} \\
&\text{rząd 2: } \hat{x}_k = 0.77x_{k-1} + 0.21x_{k-2} \\
&\text{rząd 3: } \hat{x}_k = 0.88x_{k-1} + 0.63x_{k-2} - 0.54x_{k-3} \\
&\text{rząd 5: } \hat{x}_k = 0.74x_{k-1} + 0.75x_{k-2} - 0.27x_{k-3} - 0.17x_{k-4} - 0.08x_{k-5}
\end{aligned} \tag{6.13}$$

Wariancja sygnału wejściowego (rys. 6.1) wynosi 1930.4. Sygnał ten składa się jakby z trzech części, które znacznie różniąc się od siebie podważają sensowność założeń o stacjonarności kompresowanego sygnału. Pierwszy fragment (próbki mniej więcej od 1 do 500) to mało skorelowany ciąg wartości, nie rokujący raczej wyraźnej poprawy predykcji przy zwiększaniu kontekstu, podczas gdy w dwu kolejnych częściach (próbki 500-1000 i 1000-1500) powtarzają się cyklicznie regularne kształty sygnału sugerując zastosowanie wyższego rzędu modelu predykcji.

Z przeprowadzonych testów wynikło, że w miarę zwiększania rozmiaru kontekstu malała wartość wariancji błędu predykcji, przyjmując dla rzędów 1, 2, 3, 5, 10 i 20 wartości odpowiednio: 398.0, 379.7, 256.4, 234.4, 169.4, 162.7. Widać więc wyraźnie mniejszą wariancję nowej reprezentacji oryginalnego sygnału już przy najprostszym modelu predykcji rzędu 1. Zwiększenie kontekstu do dwóch poprzednich elementów niewiele wpłynęło na wartość wariancji podczas, gdy kolejny kontekst rzędu 3 już wyraźnie zmniejszył wartość wariancji - o 35% w stosunku do rzędu 1, potwierdzając opłacalność zwiększania w tym konkretnym przypadku rozmiaru kontekstu. Przy przewidywaniu danej wartości na podstawie dziesięciu poprzednich, widać kolejne wyraźne obniżenie wartości wariancji, natomiast dwukrotnie większy kontekst nie wpływa już praktycznie na jakość predykcji. Używając cały dostępny zbiór zakodowanych poprzednio wartości z wejściowego strumienia danych do przewidywania wartości następnych uzyskano wartość wariancji błędu predykcji na poziomie 79. Taki kontekst nazwijmy modelem rzędu maksymalnego.

Wpływ rzędu predykcji na kształt nowej reprezentacji sygnału wejściowego, tj. błąd predykcji, przedstawiono na rys. 6.3. Potwierdzają one nasze przypuszczenia, że zwiększanie rozmiaru kontekstu poprawia znacząco predykcję jedynie w drugiej i trzeciej części sygnału wyraźnie słabiej wpływając na rozkład wartości błędu w pierwszej części kodowanego sygnału.



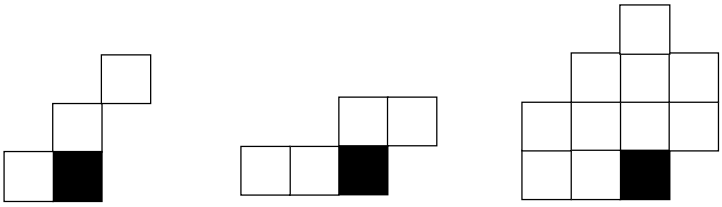

Rys. 6.3. Wykresy błędów predykcji przy zastosowaniu optymalnych w sensie kryterium średniokwadratowego modeli predykcji liniowej rzędu 1, 5, 10, 20 i maksymalnego.

W naszym eksperymencie dotyczącym kształtowania kontekstu w modelu predykcji zbadano również korelacje pomiędzy wartością wariancji błędu predykcji (skutecznością predykcji), a efektywnością kodowania sygnału wejściowego. Zastosowano statyczny koder arytmetyczny bez pamięci. Kodując oryginalny zbiór danych uzyskano średnią bitową 8 bitów na symbol (bps). Przy kodowaniu błędu predykcji z modelu rzędu 1, wyraźnie mniejszej wariancji odpowiadało zmniejszenie średniej do 6.66 bps. Okazuje się, że przy kontekście dwuelementowym następuje nieznaczne pogorszenie skuteczności kodowania do wartości 6.71 bps pomimo zmniejszenia wariancji błędu predykcji. Znajdujemy tutaj potwierdzenie poprzedniej tezy, że nie jest to jedyny czynnik wpływający na efektywność kompresji. Dla dalszych modeli predykcji, rzędu 3, 5, 10, 20 i maksymalnego uzyskano następujące wartości średniej bitowej: 6.28, 6.23, 5.96, 5.94 i 5.25. Przykład ten pokazuje więc przydatność metody predykcji liniowej w zastosowaniu do bezstratnej kompresji dźwięku. Przydatność tę potwierdza również fakt, że najskuteczniejsza z przedstawionych dotąd metod bezstratnej kompresji - technika arytmetycznego kodowania w wersji adaptacyjnej z optymalnie dobranym rzędem modelu statystycznego – użyta bez wstępnego etapu predykcji pozwala uzyskać jedynie 7.9 bps średniej bitowej.

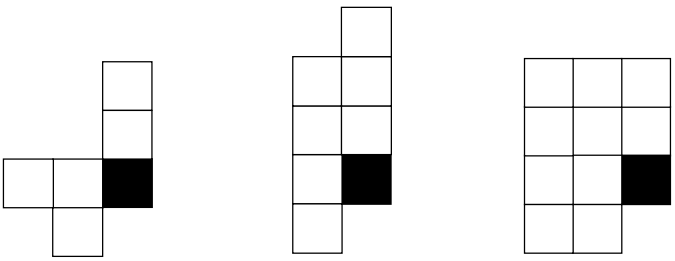
Kształt kontekstu w przypadku strumienia danych tekstowych czy innych danych jednowymiarowych jest jednoznaczny i sprowadza się do kilku danych bezpośrednio poprzedzających kodowany właśnie symbol. W niektórych przypadkach uzasadnione jest sięganie w sposób nieciągły do poprzednich wartości, np. jeśli kodowana jest sekwencja powtarzających się, bardzo podobnych fraz o znanej długości, ale są to przypadki szczególne, w których na skutek wiedzy dostępnej *a priori* modyfikowane są ogólne schematy kompresji, aby dostosować je do oszczędnego zapisu rzeczywistej informacji zawartej w strumieniu danych.

Dla obrazów wybór kształtu kontekstu jest już bardziej złożony i musi uwzględniać zależności pomiędzy pikselami w kierunku poziomym, pionowym i ukośnym. Przykładowe kształty kontekstów wykorzystywane w technikach predykcji danych obrazowych przedstawia rys. 6.4. W przypadku kompresji danych trójwymiarowych możliwe jest budowanie kontekstów w przestrzeni 3-D w sposób analogiczny.

a)



b)



Rys. 6.4. Przykłady kontekstów w predykcyjnym kodowaniu obrazów: a) przy przeglądaniu obrazów wiersz po wierszu, b) przy przeglądaniu obrazów kolumna po kolumnie. Kształt kontekstu ograniczony jest warunkiem przyczynowości.

## 6.3. Adaptacyjne modele predykcji

Wśród adaptacyjnych schematów predykcji można wyróżnić dwa podstawowe rodzaje: wprzód (ang. forward adaptation) i wstecz (ang. backward adaptation). W przypadku adaptacji wprzód do uaktualniania modelu wykorzystywane są wszystkie informacje na temat kodowanego zbioru, które są dostępne na wejściu kodera, przy czym szczególnie chodzi o dostępność całego kompresowanego zbioru danych. Informacje te są niedostępne dla dekodera, więc dla wiernego odtworzenia oryginału konieczne jest przesłanie wartości zmienianych adaptacyjnie parametrów modelu predykcji oraz wszystkich innych modyfikacji do dekodera, jako informacja dodatkowa. W takich rozwiązaniach nie obowiązuje zasada przyczynowości przy konstruowaniu optymalnego lokalnie modelu. Można więc stosować pełny kontekst otaczający kodowany symbol z dowolnej strony, gdzie tylko potencjalnie mogą wystąpić zależności pomiędzy danymi. Ponieważ proces optymalizacji nie może być powtórzony w dekoderze, wykorzystuje się zatem już gotowe parametry modelu predykcji przesłane z kodera. Czyni to algorytm kompresji/dekompresji niesymetrycznym - faza dekompresji ma znacznie niższy koszt obliczeniowy.

Drugim schematem predykcji jest metoda adaptacji wstecz, gdzie poszukiwanie najskuteczniejszego modelu odbywa się jedynie na podstawie informacji już zakodowanej (koder) lub już zdekodowanej (dekoder) z zachowaniem zasady przyczynowości. W tym przypadku żadna dodatkowa informacja nie jest przesyłana do dekodera, ale oczywiście kontekst modelu predykcji jest uboższy. Możliwe jest także łączenie obu rozwiązań przesyłając niektóre parametry modelu predykcji optymalnie dobrane w koderze jako informacje dodatkowe i modyfikując pozostałe na podstawie uboższej ‘wiedzy przyczynowej’.

*Adaptacja wprzód*

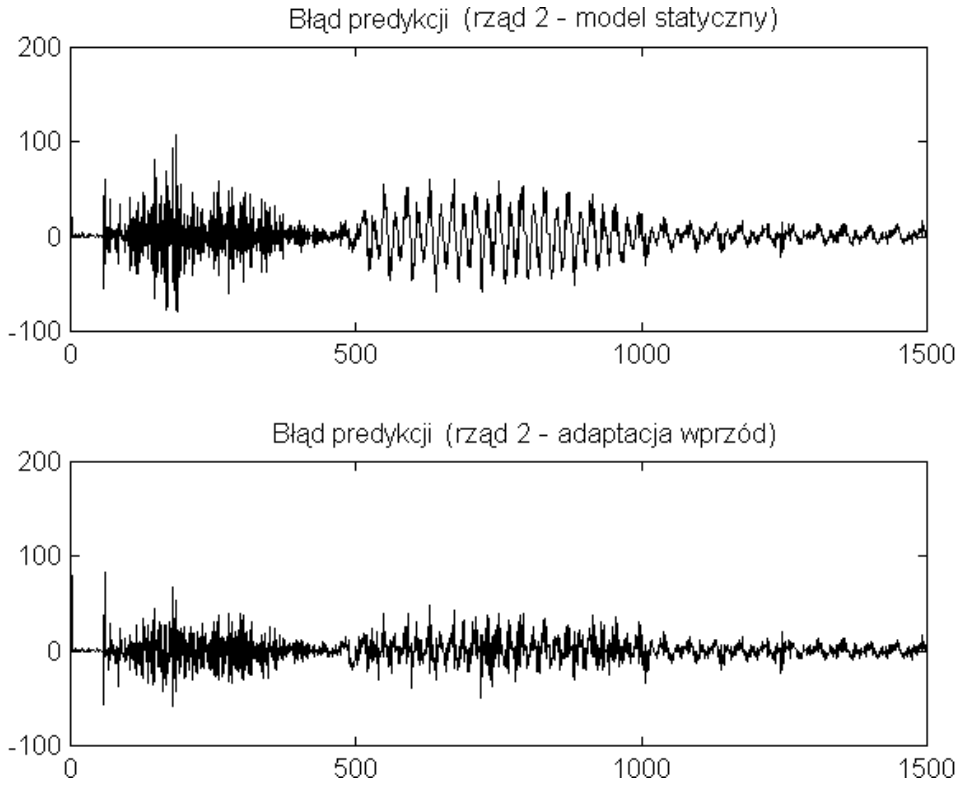
Zbiór danych wejściowych przeznaczonych do kompresji jest dzielony na kilka części. Są to zazwyczaj sąsiednie segmenty danych czy też bloki w przypadku obrazów. W koderach mowy na przykład takie segmenty zawierają około 16 ms zapisu mowy, co przy częstości próbkowania 8000 próbek na sekundę daje 128 próbek na segment. Bloki 16×16 w przypadku obrazów to 256 pikseli potencjalnie silnie skorelowanych ze sobą. Można także dokonać innego podziału zbioru danych w zależności od wstępnej analizy, np. poprzez wybrany algorytm segmentacji zbioru danych (segmentacja drzewa czwórkowego w przypadku obrazów). Parametry podziału wejściowego strumienia danych muszą być przesłane do dekodera.

Współczynniki predykcji są następnie wyznaczane dla każdego bloku oddzielnie i przesyłane do dekodera. Dla oszczędności zapisu współczynniki te mogą być określone z założoną z góry dokładnością.

Dla sygnału przedstawionego na rys. 6.1. widać nieprawdziwość założeń o stacjonarności. Można by w tej sytuacji podzielić sygnał na trzy części w przybliżeniu stacjonarne i dla każdej z nich wyznaczyć optymalny koder, np. przedstawioną metodą minimalizacji wariancji błędu. Zaobserwujmy rezultaty takiej adaptacji - zostały one przedstawione w tabeli 6.1.

Tabela 6.1. Zestawienie wyników optymalizacji schematu predykcji sygnału z rys. 6.1. przy pomocy statycznego modelu predykcji - stałego w całym zakresie sygnału wejściowego - oraz modelu dopasowanego do lokalnych własności sygnału według schematu adaptacji wprzód. Wyznaczono niezależnie trzy modele predykcji dla zakresów próbek: 1-500, 501-1000 i 1001-1500.

| Rząd modelu predykcji | Statyczny model predykcji | | Adaptacyjny model predykcji (wprzód) | |
|---|---|---|---|---|
| | Wariancja błędu | Średnia bitowa | Wariancja błędu | Średnia bitowa |
| 1 | 398.0 | 6.66 | 394.5 | 6.66 |
| 2 | 379.7 | 6.71 | 166.4 | 5.98 |
| 3 | 256.4 | 6.28 | 150.8 | 5.87 |
| 5 | 234.4 | 6.23 | 140.9 | 5.82 |
| 10 | 169.4 | 5.96 | 122.9 | 5.70 |
| 20 | 162.7 | 5.94 | 118.0 | 5.66 |




Rys. 6.5 Poprawa efektywności predykcji modelu adaptacyjnego w stosunku do modelu statycznego.

Wyniki pokazują poprawę skuteczności kodowania dla modelu adaptacyjnego, szczególnie dla modelu rzędu 2. Potwierdza to wyraźnie także rysunek 6.5. Znaczne obniżenie wariancji błędu predykcji (ponad dwukrotne) w stosunku do modelu adaptacyjnego rzędu 1 oraz zmniejszenie średniej bitowej o ponad 10% jest uzyskane niewielkim kosztem

obliczeniowym. Do zaakceptowania jest również konieczność dopisania parametrów trzech modeli predykcyjnych (po dwa współczynniki każdy) do strumienia wyjściowego. Dalsze zwiększania rzędu predykcji przynosi już niewielką korzyść - zmniejszenie średniej bitowej wartości błędu predykcji o kolejne 5% przy rzędzie 20 plus konieczność dopisania 3 razy po 20 współczynników modeli predykcji.

*Adaptacja wstecz*

Stosując to samo kryterium minimalizacji błędu średniokwadratowego można zbudować algorytm adaptacyjnej korekcji modelu predykcji w miarę kodowania kolejnych symboli wejściowych (metoda najmniejszych kwadratów – ang. LMS). Należy podkreślić, że w większości metod adaptacji wstecz modyfikacja modelu jest wykonywana po każdym kroku predykcji wartości symbolu wejściowego, analogicznie jak w adaptacyjnych metodach entropijnego kodowania.

Rozważmy schemat predykcji rzędu 1 elementu strumienia wejściowego $x_k$: $\hat{x}_k = \alpha_1 \cdot x_{k-1}$. Kwadrat błędu predykcji w tym przypadku ma postać $\varepsilon_k = e_k^2 = (x_k - \alpha_1 \cdot x_{k-1})^2$. Jest to funkcja kwadratowa względem $\alpha_1$ mające minimum dla pewnej wartości $\alpha_1 = \alpha_{opt}$. Mniejsza lub większa wartość współczynnika od $\alpha_{opt}$ daje większy błąd predykcji. Dobrą miarą zbliżania lub oddalania się wartości $\alpha_1$ względem wartości optymalnej jest pochodna wyrażenia na kwadrat błędu: $\frac{\partial \varepsilon_k}{\partial \alpha_1} = -2 \cdot (x_k - \alpha_1 \cdot x_{k-1}) x_{k-1}$. Jeśli wartość $\alpha_1$ jest mniejsza od $\alpha_{opt}$, wówczas pochodna jest ujemna i rośnie w miarę oddalania się od wartości optymalnej. Z kolei zbyt duża wartość współczynnika daje dodatnią wartość pochodnej o wartości określającej odległość od $\alpha_{opt}$. Rosnąca wartość pochodnej oznacza więc konieczność zmniejszenia wartości współczynnika $\alpha_1$ o pewną wartość, a ujemna wartość pochodnej wskazuje na konieczność zwiększenia jego wartości w celu przybliżenia się do optimum. Oczywiście, skutki lepszego doboru wartości współczynnika będą widoczne w następnym kroku predykcji kolejnej wartości ze strumienia wejściowego.

Taki mechanizm modyfikacji wartości współczynników można zapisać następująco: $\alpha_1^{(k+1)} = \alpha_1^{(k)} - \beta' \frac{\partial \varepsilon_k}{\partial a_1}$, co po podstawieniu wyrażenia na pochodną błędu sprowadza się do następującej postaci schematu adaptacyjnej predykcji:

$$\alpha_1^{(k+1)} = \alpha_1^{(k)} + 2\beta' \cdot (x_k - \alpha_1 \cdot x_{k-1}) x_{k-1} = \alpha_1^{(k)} + \beta \cdot e_k \cdot x_{k-1}, \qquad (6.14)$$

gdzie $\beta = 2\beta'$. Do modyfikacji wartości współczynnika predykcji wykorzystuje się więc jedynie kodowaną (odtworzoną) poprzednio wartość danej $x_{k-1}$ oraz wyznaczoną (odczytaną) ostatnio wartość błędu predykcji $e_k$. Współczynnik równania $\beta$, który określa szybkość adaptacji przyjętego schematu, powinien odpowiadać charakterowi zmian lokalnych kompresowanego zbioru danych. Jeśli zmiany te są gwałtowne (z dużą wartością gradientu, częstą zmianą kierunku gradientu), wtedy wartość współczynnika powinna być mniejsza, jeśli zaś zmiany są łagodne, to wartość współczynnika może być większa. Wartość współczynnika $\beta$ może być więc wyznaczona w koderze w konwencji adaptacji wprzód i jako informacja dodatkowa przesłana do dekodera tworząc schemat łączony adaptacji wprzód - wstecz.

Można też ten prosty schemat rozciągnąć na model predykcji dowolnego rzędu *m*. Stosując takie same przekształcenia wyrażenia na kwadrat błędu predykcji: $\varepsilon_k = \left( x_k - \sum_{i=1}^{m} \alpha_i x_{k-i} \right)^2$, licząc pochodne cząstkowe po współczynnikach $\alpha_i$ i tworząc schemat adaptacji (6.14) niezależnie dla każdego z nich, otrzymujemy:

$$\alpha_i^{(k+1)} = \alpha_i^{(k)} + \beta \cdot e_k \cdot x_{k-i} \quad \text{dla} \quad i=1, ..., m \tag{6.15}$$

Ta metoda wstecznej adaptacji modelu predykcji może być łatwo zaimplementowana w algorytmie bezstratnej kompresji, a koszty obliczeniowe sekwencyjnej modyfikacji wartości współczynników są stosunkowo niewielkie. Nie w każdym jednak przypadku zastosowanie takiego schematu predykcji daje oczekiwaną poprawę skuteczności kompresji.

W eksperymencie testowano prosty schemat predykcji w wersji statycznej, jak i adaptacyjnej wstecz. Korzyść jest niewielka, a w najlepszym przypadku dla kontekstu rzędu 1 uzyskano zmniejszenie średniej bitowej o 0.05 bita/symbol w stosunku do statycznej predykcji liniowej.

## Bibliografia:


1. C. C. Cutler, Differential Quantization for Television Signals, U.S. Patent 2,605,361, July, 1952.
2. T. V. Ramabadran, K. Chen, *The Use of Contextual Information in the Reversible Compression of Medical Images*, IEEE Trans. Medical Imaging, 11(2): 185-195, 1992.
3. G. R. Kuduvalli, R. M. Rangayyan, *Performance Analysis of Reversible Image Compression Techniques for High-Resolution Digital Teleradiology*, IEEE Trans. Medical Imaging, 11(3):430-445, 1992.
4. B. Townshend, *Nonlinear Prediction in Speech*, Proceedings of the International Conference on Acoustics, Speech and Signal Processing, 425-428, 1991.
5. W. Zschunke, *DPCM Picture Coding with Adaptive Prediction*, IEEE Trans. Communications, 25:1295-1302, 1977.
6. S. A. Martucci, *Reversible Compression of HDTV Images Using Median Adaptive Prediction and Arithmetic Coding*, IEEE International Symposium on Circuits and Systems, 1310-1313, IEEE Press, 1990.