Partner:

SYSTEMY TELEINFORMATYCZNE

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

Projekt Absolwent na miarę czasu współfinansowany przez Unię Europejską
w ramach Europejskiego Funduszu Społecznego
Nr umowy UDA-POKL.04.01.01-00-421/10-01

Politechnika Lubelska
Wydział Elektrotechniki i Informatyki
ul. Nadbystrzycka 38A
20-618 Lublin

SYSTEMY TELEINFORMATYCZNE

Waldemar Wójcik

Politechnika Lubelska
Lublin 2011

Recenzenci
dr hab. inż. Oleksandra Hotra, prof. PL
prof. dr hab. Volodymyr Harbarchuk

Publikacja finansowana z projektu „Absolwent na miarę czasu”

Projekt „Absolwent na miarę czasu” współfinansowany przez Unię Europejską w ramach Europejskiego Funduszu Społecznego. Nr umowy UDA-POKL.04.01.01-00-421/10-01

Ta publikacja odzwierciedla jedynie stanowiska jej autorów, a Komisja Europejska nie ponosi odpowiedzialności za informacje w niej zawarte

Publikacja dystrybuowana bezpłatnie

Publikacja wydana za zgodą Rektora Politechniki Lubelskiej

ISBN: 978-83-62596-64-5

Wydawca: Politechnika Lubelska
ul. Nadbystrzycka 38D, 20-618 Lublin
Realizacja: Biblioteka Politechniki Lubelskiej
Ośrodek ds. Wydawnictw i Biblioteki Cyfrowej
ul. Nadbystrzycka 36A, 20-618 Lublin
tel. (81) 538-46-59, email: wydawca@pollub.pl
www.biblioteka.pollub.pl
Druk: ESUS Agencja Reklamowo-Wydawnicza Tomasz Przybylak
www.esus.pl

Elektroniczna wersja książki dostępna w Bibliotece Cyfrowej PL www.bc.pollub.pl
Nakład: 100 egz.

Spis treści

1. Wstęp

Teleinformatyka wkracza do wszystkich dziedzin naszego życia, takich jak: edukacji, handlu, opieki medycznej, motoryzacji, bankowości, finansów, administracji, rozrywki i wielu innych. Usługi teleinformatyczne pozwalają na rozwój przedsiębiorstwa i na osiągnięcie bardzo dobrej pozycji konkurencyjnej. Z możliwości i zastosowań jakie oferuje teleinformatyka nie sposób nie skorzystać w dzisiejszych czasach. Możliwy jest elektroniczny obrót pieniądzem, przesyłanie dokumentów w formie elektronicznej, transmisja już nie tylko plików tekstowych, ale również dźwięku i obrazu.

Internet jest podstawowym źródłem pozyskiwania i przekazywania informacji. Coraz częściej mówi się o tym, że Internet będzie medium przyszłości, łączącym w doskonały sposób możliwości prasy, radia, telewizji. Już korzystamy z internetowych sklepów, bibliotek, kawiarni, banków i archiwów, elektronicznych giełd.

W myśl ustawy o świadczeniu usług drogą elektroniczną z dnia 18 lipca 2002 r. System teleinformatyczny to zespół współpracujących ze sobą urządzeń informatycznych i oprogramowania, zapewniający przetwarzanie i przechowywanie, a także wysyłanie i odbieranie danych poprzez sieci telekomunikacyjne za pomocą właściwego dla danego rodzaju sieci urządzenia końcowego w rozumieniu ustawy z dnia 16 lipca 2004 r. – Prawo telekomunikacyjne (Dz.U. z 2004 r. Nr 171, poz. 1800, z późn. zm.).

Teleinformatyka jako nauka, to jedna z dziedzin informatyki, która zajmuje się wykorzystaniem sieci telekomunikacyjnej do przekazywania informacji między komputerami. Jej najpopularniejszą jak dotąd realizacją jest Internet.

Teleinformatyka, ICT (akronim od ang. Information and Communication Technologies) – dział telekomunikacji i informatyki, zajmujący się technologią przesyłu informacji oraz narzędziami logicznymi do sterowania przepływem oraz transmisją danych za pomocą różnych medium. W rozdziale 2 przedstawiono zagadnienia związane z zastosowaniem światłowodów jako medium w teleinformatyce. Po przedstawieniu idei telekomunikacyjnego łącza światłowodowego przeanalizowano jego podstawowe elementy takie jak źródła światła, kable i złącza światłowodowe, sprzęgacze oraz urządzenia służące do zwielokrotniania kanałów informacyjnych.

Przykładem zastosowania teleinformatyki i połączenia przesyłu telefonii i zastosowań informatycznych, jest powszechnie stosowana technologia VOIP, która bez ograniczeń i limitów może śmiało konkurować z tradycyjnym telefonem i przekazywaniem dźwięku/obrazu. W rozdziale 3 zawarto definicję systemu multimedialnego opisano standardy kompresji wideo i dźwięku oraz przedstawiono zagadnienia przesyłu głosu w sieciach IP. Po zapoznaniu czytelnika z podstawowymi pojęciami szeroko pojętych multimediów w rozdziale 4 przedstawiono zagadnienia związane z ich strumieniowaniem. Rozdział 5 zawiera natomiast opis algorytmów kompresji i rozpoznawania obrazów.

System informatyczny jest to zbiór powiązanych ze sobą elementów, którego funkcją jest przetwarzanie danych przy użyciu techniki komputerowej. Na systemy informatyczne składają się obecnie takie elementy jak sprzęt – obecnie głównie komputery, oraz urządzenia służące do przechowywania danych, urządzenia służące do komunikacji między sprzętowymi elementami systemu, urządzenia służące do komunikacji między ludźmi a komputerami, urządzenia służące do odbierania danych ze świata zewnętrznego – nie od ludzi (na przykład czujniki elektroniczne, kamery, skanery). W odpowiedzi na powyższą definicję systemów informatycznych powstał rozdział 6 niniejszego podręcznika taktujący o sieciach komputerowych i zagadnieniach dotyczących ich administrowania. Oprócz przedstawionych metod i technologii zarządzania sieciami poruszono również tematy związane z monitorowaniem sieci, zarządzaniem ich awariami oraz oceny ich niezawodności z perspektywy usług sieciowych.

Na systemy informatyczne składają się również takie elementy jak urządzenia służące do wywierania wpływu przez systemy informatyczne na świat zewnętrzny – elementy wykonawcze (na przykład silniki sterowane komputerowo, roboty przemysłowe, podłączony do komputera ekspres do kawy, sterowniki urządzeń mechanicznych) oraz urządzenia służące do przetwarzania danych niebędące komputerami. Rozdział 7 porusza zatem zagadnienia sztucznej inteligencji oraz systemów ekspertowych, w którym czytelnik odnajdzie wprowadzenie w problematykę związaną z sztucznymi sieciami neuronowymi, systemami rozmytymi a także algorytmy ewolucyjne i technologie agentowe.

2. Technika światłowodowa w teleinformatyce

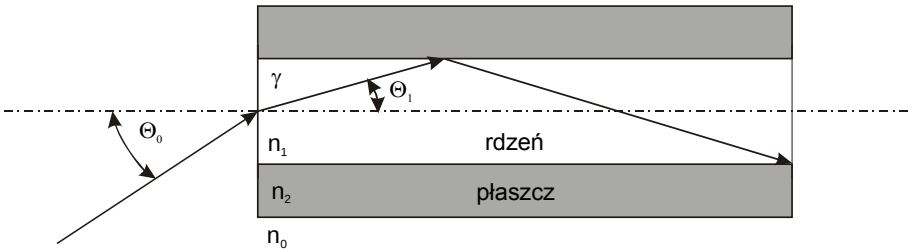
Światłowody są to kanały do prowadzenie fali świetlnej. Mogą mieć one różną geometrię – wyróżnia się światłowody planarne (w postaci pasków prowadzących światło) oraz włókniste (inaczej cylindryczne – w postaci koncentrycznych warstw prowadzących promieniowanie).
Światłowody są medium transmisyjnym, bez którego współczesna teleinformatyka nie może się obyć. Wynika to z ich właściwości transmisyjnych, jakich nie posiadają żadne inne kanały fizyczne, wśród których można wymienić:

- mała tłumienność (rzędu 0,2dB/km),
- bardzo szerokie pasmo częstotliwości(ok. 0,1THz),
- możliwość zwielokrotniania kanałów w jednym włóknie (ponad 400),
- wysoka trwałość i niezawodność (ok. 25lat),
- niewrażliwość na zakłócenia elektromagnetyczne,
- niski koszt medium (poniżej 1$ za 1m),
- niska waga medium (kilometr samego włókna waży ok. 30g).

Pierwsze włókno optyczne o właściwościach pozwalających na wykorzystanie go do transmisji danych zostało wytworzone w 1968 roku w Japonii, a w Polsce pod koniec lat 70 XX wieku (w Lublinie). Od tamtego czasu trwają prace zmierzające do poprawy jego właściwości transmisyjnych i użytkowych, lepszego wykorzystania go do celów transmisyjnych oraz prace nad aktywnymi (źródła światła, detektory, wzmacniacze) i pasywnymi (złącza, sprzęgacze, cyrkulatory, filtry) elementami niezbędnymi do realizacji transmisji światłowodowej. Dzięki temu w 2010 roku osiągnięto transmisję pojedynczym włóknem o przepustowości 69,1Tb/s na odległości 240km przy wykorzystaniu 432 kanałów WDM [13].

2.1. Budowa i zasada działania światłowodu

Zasada działania światłowodu związana jest ze zjawiskiem całkowitego wewnętrznego odbicia. Zjawisko to zachodzi, gdy światło biegnące z ośrodka o większym współczynniku załamania, pada na granicę ośrodków pod kątem większym od tzw. kąta granicznego.
Klasyczny światłowód koncentryczny składa się więc z dwóch materiałów różniących się wartością współczynnika załamania. Materiał o większym współczynniku załamania (rdzeń, ang. core) otoczony jest materiałem o współczynniku załamania mniejszym (płaszcz, ang. cladding). W niektórych światłowodach rolę płaszcza pełni otoczenie, np. powietrze.
Przekrój światłowodu o skokowej zmianie współczynnika załamania oraz zasadę propagacji w nim światła przedstawia rys. 2.1.



Rys. 2.1. Budowa światłowodu skokowego (n_0, n_1, n_2 – odpowiednio współczynniki załamania otoczenia, rdzenia i płaszcza) oraz zasada propagacji światła (Θ_0, γ – kąt padania światła na czoło światłowodu oraz granicę rdzeń-płaszcz, Θ_1 – kąt załamania światła w rdzeniu) [16]

Rodzaj materiału, z którego wykonany jest rdzeń i płaszcz, rozkład współczynnika załamania materiałów wzdłuż średnicy światłowodu, a także ilość propagowanych modów światła są podstawowymi kryteriami, według których klasyfikuje się światłowody.
Najbardziej ogólny podział światłowodów włóknistych:

- ze względu na wykorzystywane materiały:
  - światłowody szklane,
  - światłowody plastikowe;
- ze względu na rozkład współczynnika załamania:
  - światłowody skokowe,
  - światłowody gradientowe;
- pod względem liczby prowadzonych modów:
  - światłowody jednomodowe,
  - światłowody wielomodowe;
- pod względem oddziaływania na transmitowany sygnał:
  - światłowody pasywne,
  - światłowody aktywne.

2.2. Idea telekomunikacyjnego łącza światłowodowego

Telekomunikacyjne łącze światłowodowe składa się z:

- nadajnika (zwykle źródło światła z modulatorem), który przekształca sygnał elektryczny niosący informację źródłową na sygnał świetlny;
- toru telekomunikacyjnego, na który składa się światłowód wraz z niezbędnymi światłowodowymi elementami pasywnymi i aktywnymi;
- odbiornika (zwykle detektor ze wzmacniaczem), który konwertuje sygnał świetlny na elektryczny, przekazywany do odbiorcy.

Wykorzystanie światła jako nośnika informacji w miejsce sygnałów elektrycznych znacznie zmienia właściwości łącza. Oprócz zalet wymienionych wcześniej, łącze takie charakteryzuje się znacznie większymi odstępami między regeneratorami (co znacznie obniża koszty łącza), a także znacznie większy poziom bezpieczeństwa.

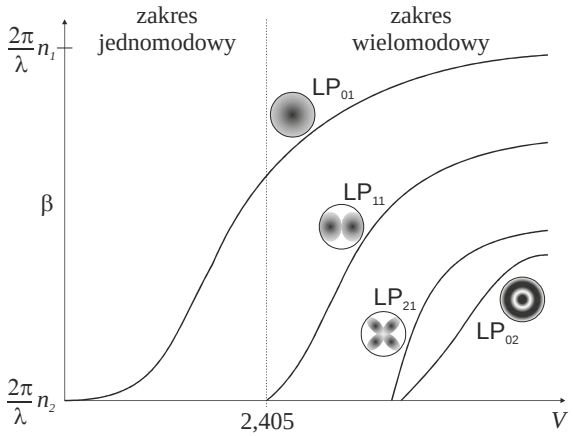
W łączach światłowodowych występuje znacznie mniej punktów dostępowych, w których potencjalny intruz mógłby przyłączyć się do sieci. Niezauważone włączenie się do linii światłowodowej jest praktycznie niemożliwe, gdyż z reguły wymaga przynajmniej chwilowego przerwania linii, co powinno być zauważone przez nadzór łącza. Ponadto obecność jakiegokolwiek dodatkowego elementu w torze światłowodowym wpływa na jego parametry, a więc jest łatwo wykrywalna, a ponadto łatwa do zlokalizowania.
Stosowanie światłowodów jako toru transmisyjnego ma też swoje wady. Zaliczyć do nich można łączenie światłowodów. Proces taki wymaga zaawansowanych technologicznie urządzeń oraz wykwalifikowanej obsługi. Stosunkowo kosztowne jest również takie łączenie źródeł światła ze światłowodami oraz światłowodów z detektorami, aby zapewnić jak najmniejsze straty.
Kolejną wadą jest występowanie stosunkowo dużych granicznych promieni gięcia światłowodów i kabli światłowodowych. Powoduje to trudności w ich układaniu oraz zwiększa straty występujące w linii. Problemy te mogą być częściowo rozwiązane poprzez stosowanie specjalnych konstrukcji włókien, tzw. włókien ze zminimalizowanym promieniem gięcia (ang. LowBend) [20].
Za jedną z największych wad łączy światłowodowych uważa się niedostępność optycznych przełączników, których stosowanie pozwoliłoby na pełniejsze wykorzystanie zalet światłowodów. Obecnie przełączanie sygnałów odbywa się na drodze elektrycznej, co wymaga wcześniejszej konwersji sygnału ze świetlnego na elektryczny. Proces taki znacznie ogranicza szybkość działania sieci. Zagadnienie to jest jednak w fazie intensywnych badań zmierzających do wytworzenia komercyjnych przełączników optycznych.

2.3. Światłowody jedno i wielomodowe

Modem światłowodowym nazywa się charakterystyczny rozkład pola elektromagnetycznego wzbudzany w światłowodzie. Ilość pobudzanych we włóknie modów uzależniona jest od tzw. częstotliwości znormalizowanej V, która jest funkcją parametrów światłowodu (materiałowych i geometrycznych) oraz zależy od długości fali propagującego światła.

$$V = \frac{2\pi a}{\lambda} NA,$$

gdzie a – promień rdzenia światłowodu, NA – apertura numeryczna światłowodu, λ - długość fali. Rys. 2.2 pokazuje mody występujące przy danej wartości częstotliwości znormalizowanej.



Rys. 2.2. Zależność występowania modów liniowospolaryzowanych LP od częstotliwości znormalizowanej

W przypadku gdy $V \leq 2{,}405$ to w światłowodzie propaguje się tylko jeden mod, którego średnicę pola w_0 (odległość między miejscami, w którym moc osiąga wartość e^2 razy mniejszą niż wartość maksymalna) można wyznaczyć z zależności:

$$w_0 = 2a\left(0{,}65 + 1{,}619V^{-3/2} + 2{,}879V^{-6}\right),$$

a rozkład mocy wzdłuż promienia r można opisać zależnością (krzywa Gaussa):

$$P(r) = P(0)e^{-\left(\frac{r}{w_0}\right)^2}$$

Długość fali świetlnej, dla której $V = 2{,}405$ nazywa jest długością fali odcięcia i jest jednym z ważniejszych parametrów światłowodów, który określa dla jakich długości światła światłowód jest wielomodowy. Liczbę propagowanych modów M określić można w tym przypadku w przybliżeniu na:

- $\mathrm{M} \approx \frac{\mathrm{V}^2}{2}$ dla światłowodów skokowych,
- $\mathrm{M} \approx \frac{\mathrm{V}^2}{4}$ dla światłowodów gradientowych.

Z przedstawionych zależności wynika, że dla światłowodów jednomodowych wraz ze spadkiem częstotliwości znormalizowanej (wzrostem długości fali światła) zwiększa się średnica modu przekraczając pole rdzenia. Oznacza to, że dla światłowodów jednomodowych o małej wartości V, znaczna część światła jest propagowana w płaszczu, np. dla $V = 2{,}4$ w rdzeniu propagowanych jest 80% mocy, a dla $V = 1{,}0$ tylko 20% prowadzonych jest w rdzeniu [15].

Stosowanie światłowodów do budowy sieci teleinformatycznych wymaga zrozumienia wpływu cech włókien optycznych na parametry toru telekomunikacyjnego. Do podstawowych parametrów światłowodów należą: apertura numeryczna, tłumienność i dyspersja.

2.3.1. Apertura numeryczna

Jednym z podstawowych parametrów charakteryzujących światłowód jest apertura numeryczna NA. Określa ona maksymalny kąt pod jakim światło wprowadzone do światłowodu będzie się w nim propagowało, czyli tak zwany kąt akceptacji światłowodu.

Światło wprowadzone do światłowodu pod kątem wyznaczonym względem osi światłowodu zawierającym się w stożku akceptacji ulega całkowitemu wewnętrznemu odbiciu na granicy rdzeń-płaszcz. Wynika stąd następująca zależność:

$$n_0 sin\alpha_{max} = \sqrt{n_r^2 - n_p^2} = NA,$$

gdzie n_0 – współczynnik załamania ośrodka, z którego pada promieniowanie, n_r, n_p – współczynnik załamania rdzenia i płaszcza odpowiednio, NA – apertura numeryczna, α_{max} – kąt akceptacji światłowodu. Apertura numeryczna zależy więc jedynie od stałych materiałowych światłowodu, a te zależą między innymi od długości fali światła, stąd aperturę numeryczną podaje się dla określonej długości fali.

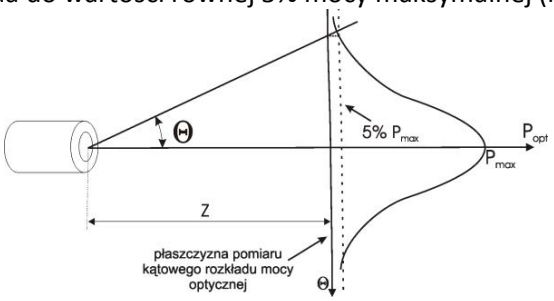
W sposób praktyczny aperturę numeryczną można wyznaczyć wykreślając zależność mocy wyjściowej światłowodu od kąta padania fali płaskiej – pomiary należy przeprowadzać w warunkach tzw. dalekiego pola.

Dalekie pole dla włókien optycznych to odległość z od czoła światłowodu spełniająca warunek:

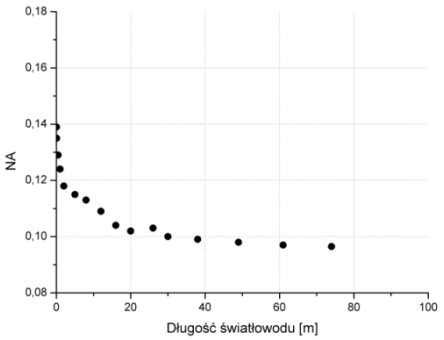
$$z \gg \frac{(2a)^2}{\lambda},$$

gdzie $2a$ - średnica rdzenia światłowodu, λ - długość fali świetlnej.

Dla tak wykonanych pomiarów za kąt graniczny przyjmowany jest kąt, dla którego moc optyczna spada do wartości równej 5% mocy maksymalnej (rys. 2.3).



Rys. 2.3. Wyznaczenie kąta akceptacji z pomiarów rozkładu mocy w polu dalekim światłowodu [17]



Rys. 2.4. Zależność apertury numerycznej od długości włókna pomiarowego obrazująca spadek udziału modów wyższych rzędów [18]

W sposób przybliżony można wyznaczyć aperturę numeryczną z pomiaru rozmiaru plamki świetlnej otrzymanej na ekranie umieszczonego w pewnej odległości od czoła światłowodu. W tym celu należy wyliczyć sinus połowy kąta rozwarcia otrzymanego stożka.
W zależności od warunków pobudzenia włókna światłowodowego przy pomiarach tego typu można otrzymać inne wartości apertury numerycznej. Wynika to z obecności we włóknie modów wyższego rzędu sztucznie zawyżających aperturę numeryczną. Zanikają one wraz z przebytą we włóknie odległością (rys. 2.4). Poprawne wyznaczenie apertury numerycznej wymaga więc zachowania na końcu włókna stanu modowego ustalonego. Można to osiągnąć albo mierząc rozkład mocy dla odpowiednio długiego odcinka światłowodu, albo stosując tuż za wejściowym końcem światłowodu mieszacz modów.

2.3.2. Tłumienie światłowodu

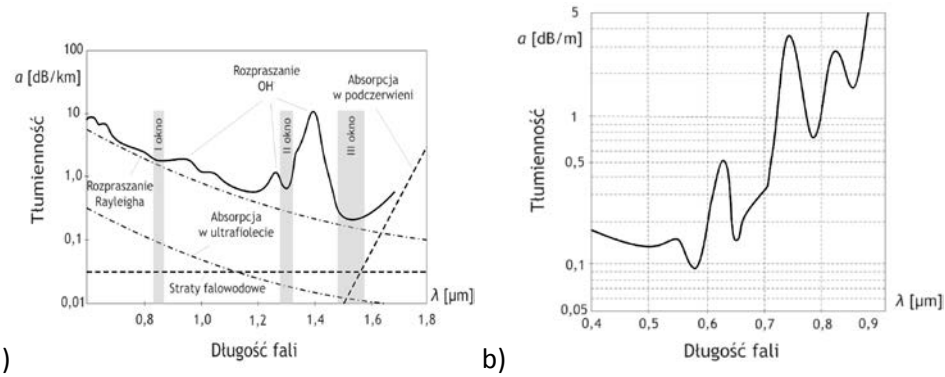
Parametrem, który warunkuje zakres transmisji światłowodu jest jego tłumienie. Parametr ten definiowany jest jako stosunek mocy wyjściowej P_{wyj} do mocy wejściowej P_{wej} i może być wyrażany w procentach lub decybelach, przy czym ta druga jednostka jest powszechnie stosowana w przemyśle. Wpływ na tłumienie odcinka światłowodu ma rodzaj materiału, z którego jest on wykonany oraz jego długość. Wyraża to wzór:

$$P_{wyj} = P_{wej} 10^{-\alpha L}$$

gdzie $\alpha = \frac{10}{L} log \frac{P_{wej}}{P_{wyj}}$ [dB/km] - współczynnik tłumienia jednostkowego światłowodu, L - długość światłowodu.
Współczynnik α zwany jest tłumiennością światłowodu i wyrażany jest w decybelach/kilometr. Jest on dla materiału, z którego wykonany jest światłowód, silnie zależny od długości fali świetlnej (rys. 2.5). Na wartość tego współczynnika mają wpływ takie czynniki jak rozpraszanie Rayleigha ($\sim\lambda^{-4}$), straty falowodowe (niezależne od długości fali), absorpcja w ultrafiolecie i podczerwieni oraz na grupach OH^- [3].

Lokalne minima tłumienności światłowodu telekomunikacyjnego wyznaczają tzw. okna transmisyjne. Są to (rys. 2.5a):

- I okno transmisyjne (krótkofalowe) – znajdujące się w otoczeniu długości fali 850nm,
- II okno transmisyjne (średniofalowe) – znajdujące się w otoczeniu długości fali 1300nm,
- III okno transmisyjne (długofalowe) – znajdujące się w otoczeniu długości fali 1550nm.



a) b)

Rys. 2.5. Widmowa zależność współczynnika tłumienności dla światłowodu krzemowego (a) oraz światłowodu polimerowego PMMA (b) [18]

III okno transmisyjne charakteryzuje się najmniejszą tłumiennością (rzędu 0,2dB/km) i w związku z tym umożliwia najdłuższe międzyregeneratorowe odcinki toru telekomunikacyjnego.
Porównanie pod względem tłumienności krzemionki z innymi materiałami (tabela 2.1) pokazuje zalety włókien światłowodowych.

Tabela 2.1. Straty mocy w różnych materiałach – na podstawie [4]

medium transmisyjne	tłumienność [dB/km]	odległość międzyregeneratorowa [km]
kabel koncentryczny	25	1,5
skrętka telefoniczna	12-18	2-3
szkło okienne	5	0,007
szkło krzemionkowe	0,16-1,00	50-250
szkło halogenkowe (w fazie badań)	0,01	3500

W wyniku prowadzonych badań opracowano również światłowody o zmodyfikowanej charakterystyce tłumienności, w których wyróżniane są dodatkowe okna: IV okno transmisyjne – dla długości fal większych niż III okno, do około 1610nm oraz V okno transmisyjne pomiędzy oknami II i III.

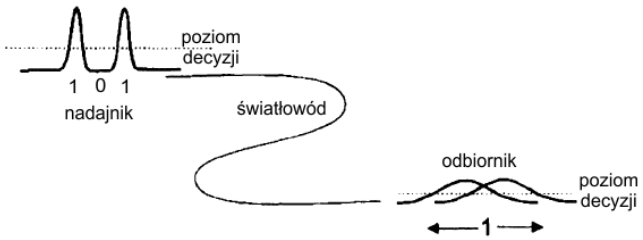
Międzynarodowy Związek Telekomunikacyjny (ang. International Telecommunication Union - ITU) wprowadził następujący podział okien transmisyjnych:

- pasmo O (ang. orginal) - od 1260nm do 1360nm,
- pasmo E (ang. extended) - od 1360nm do 1460nm,
- pasmo S (ang. short) - od 1460nm do 1530nm,
- pasmo C (ang. conventional) - od 1530nm do 1565nm,
- pasmo L (ang. long) - od 1565nm do 1625nm,
- pasmo U (ang. ultra-long) - od 1625nm do 1675nm.

Tłumienność włókna światłowodowego można wyznaczyć bezpośrednio korzystając ze wzoru definicyjnego. Wymaga to wyznaczenia mocy wejściowej i wejściowej światłowodu. O ile wyznaczenie mocy wyjściowej nie stanowi problemu, to pomiar mocy wejściowej tak. Nie można jej określić na podstawie mocy źródła światła, gdyż nie można z dostateczną dokładnością ocenić stopnia sprzężenia źródło światła – światłowód. Konieczne jest więc zastosowanie specjalnej metody pomiarowej.
Jako metodę wzorcową pomiaru tłumienia włókna podaje się tzw. metodę odcięcia, zwaną również metodą odcięcia wstecznego (ang. cut-back method). Polega ona na tym, że za moc wejściową przyjmuje się moc wyjściową z krótkiego (ok. 2m) odcinka światłowodu sprzężonego i pobudzonego w taki sam sposób, jak przy pomiarze mocy wyjściowej dla całego badanego światłowodu. Uzyskanie miarodajnego wyniku wymaga, aby pomiar był przeprowadzony dla ustalonego stanu modowego. Najlepiej zapewnić to umieszczając tuż za źródłem światła mieszacz modów.
Wadami metody odcięcia są: niszczący charakter metody (konieczność odcięcia kawałka włókna) oraz konieczność dostępu do obydwu końców światłowodu. Wady tej nie posiada metoda reflektometryczna, która jest powszechnie stosowana w przemyśle.

2.3.3. Dyspersja

Zależność parametrów ośrodka od częstotliwości (długości fali) nazywana jest dyspersją [14]. W przypadku szkieł, zmianie wraz z długością fali ulega współczynnik załamania, a więc prędkość rozchodzenia się światła w ośrodku. Dyspersja odgrywa szczególną rolę przy transmisji cyfrowej o dużej przepływności. Jej wpływ obrazowo przedstawia rysunek 2.6.



Rys. 2.6. Schemat wpływu dyspersji na transmisję cyfrową [2]

Na skutek niemonochromatyczności transmitowanej wiązki światła, na końcu światłowodu obserwuje się rozmycie czasowe impulsów. Dla impulsów krótkich, może to doprowadzić do ich nakładania się, a więc niejednoznaczności w interpretacji w odbiorniku. Oczywiście zakłóca to transmisję i aby temu zapobiec można wydłużyć czasy trwania bitów, ale jest to równoznaczne ze zmniejszeniem prędkości transmisji. Można również zastosować łącze o mniejszej dyspersji lub stosować włókna o różnych znakach dyspersji, tak by wypadkowa dyspersja była znacznie mniejsza.

Terminem *dyspersja prędkości grupowejświatłowodu* określa się jeden z zasadniczych parametrów określających właściwości transmisyjne światłowodu przy przesyłaniu impulsów świetlnych. Definiuje się ją następującym wzorem:

$$D = \frac{d\tau}{d\lambda}$$

gdzie τ - jednostkowy czas przejścia impulsu (czas przejścia impulsu przez odcinek światłowodu o długości 1 km $\tau = 1/v_g$), λ - długość fali świetlnej.

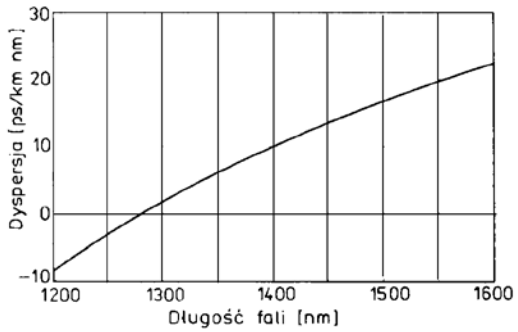
Jednostką dyspersji prędkości grupowej jest pikosekunda przez kilometr razy nanometr: ps/(km·nm).

Dystans L_D po jakim zjawisko dyspersji staje się znaczące w transmisji nazywany jest długością dyspersyjną i opisuje go wzór:

$$L_D = \frac{T_o^2}{|\beta_2|}$$

w którym przez T_o oznaczono szerokość propagowanego impulsu, a β_2 oznacza drugą pochodną stałej propagacji światła po pulsacji.

Zależność dyspersji od długości fali dla szkła kwarcowego przedstawia rysunek 2.7.



Rys. 2.7. Zależność dyspersji D od długości fali świetlnej dla szkła kwarcowego [9]

Z rysunku widać, że dla czystego szkła wartość dyspersji wynosi 0 dla długości fali 1280 nm. Dyspersja jest dodatnia dla fal dłuższych od 1280 nm. Dla długości fal świetlnych wykorzystywanych w transmisji światłowodowej (II i III okno transmisyjne) dyspersja wynosi odpowiednio +1,86 ps/(km·nm) i +19,7 ps/(km·nm).

Dyspersja związana z zależnością współczynnika załamania światła od długości fali świetlnej nazywana jest dyspersją materiałową. Jest to jeden ze składników dyspersji włókien światłowodowych. Należy pamiętać, że ten rodzaj dyspersji nie jest charakterystyczny tylko dla włókien światłowodowych, ale dla materiału, z którego są one wykonane. Kolejnym składnikiem jest dyspersja falowodowa. Jak wynika z nazwy, ten rodzaj dyspersji związany jest z prowadzeniem światła we włóknie optycznym. Związana jest ona z zależnością od częstotliwości efektywnego współczynnika załamania, co jest spowodowane zmianami podziału mocy pomiędzy rdzeniem i płaszczem światłowodu. Dyspersja falowodowa światłowodów telekomunikacyjnych zwykle ma ujemny znak dla wykorzystywanego zakresu spektralnego i maleje wraz z długością fali. W światłowodach dyspersja falowodowa zawsze towarzyszy dyspersji falowodowej i łącznie nazywane są dyspersją chromatyczną. Ten rodzaj dyspersji jest obserwowany w każdym typie światłowodu.

W światłowodach wielomodowych obserwowany jest jeszcze jeden typ dyspersji – dyspersja międzymodowa. Związany jest on z różnymi stałymi propagacji dla prowadzonych modów światła.

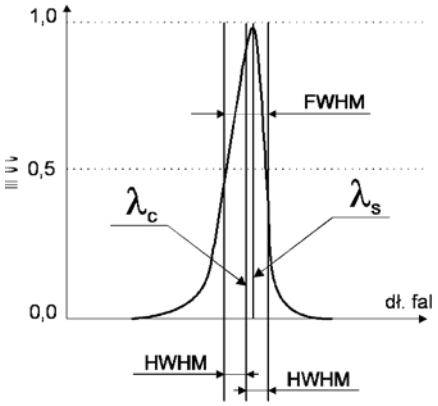
Do lat 90-tych XX wieku sądzono, że w światłowodach jednomodowych występuje tylko dyspersja chromatyczna. Okazało się jednak, że przy znacznych szybkościach transmisji, występuje zjawisko nazwane dyspersją polaryzacyjną (ang. polarization mode dispersion – PMD). Jej źródłem jest niestałość stałych propagacji dla prostopadłych składowych polaryzacji propagującego światła. Różnica między stałymi propagacji wynika z niejednorodności i losowych niedoskonałości światłowodu. Wprawdzie wartość tej dyspersji nie jest duża, jednak jej istnienie ogranicza prędkość transmisji danych.

- Możliwe jest kształtowanie charakterystyki dyspersyjnej światłowodów m.in. poprzez dobór ukształtowania profilu współczynnika załamania w światłowodzie. Dzięki takim działaniom uzyskano kilka typów światłowodów różniących się kształtem charakterystyki dyspersji, m.in.:
- światłowód z przesuniętą dyspersją (ang. dispersion shifted fiber – DSF),
- światłowód o płaskiej charakterystyce (ang. dispersion flattened fibre – DFF),
- światłowód o niezerowej przesuniętej dyspersji (ang. non-zero dispersion shifted fiber – NZDSF),
- światłowód kompensacyjny (ang. dispersion compesating fiber – DCF),
- światłowód o odwrotnej charakterystyce dyspersyjnej (ang. reverse dispersion fiber – RDF).

Poszerzenie impulsu na końcu włókna zależy zarówno od parametrów toru transmisyjnego (dyspersji wypadkowej toru) oraz od jego długości i szerokości spektralnej źródła światła. Wykorzystanie włókien o niestandardowych charakterystykach dyspersji w połączeniu z włóknami konwencjonalnymi pozwala na zarządzanie dyspersją, a więc kształtowaniu charakterystyki transmisyjnej całego toru. Zagadnienia z tym związane można znaleźć np. w pozycji [1].

2.4. Źródła światła

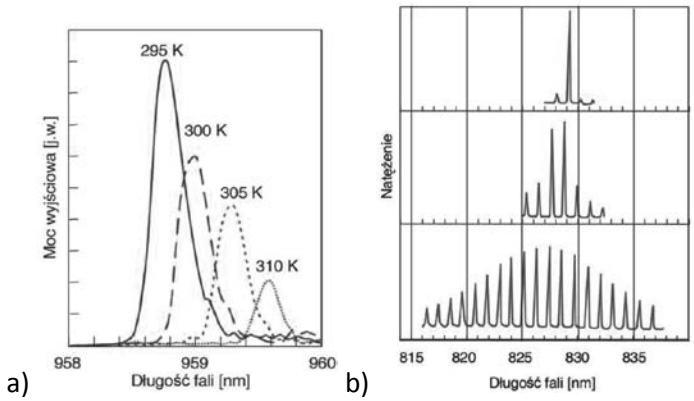
W technice światłowodowej najczęściej wykorzystywanymi źródłami światła są diody elektroluminescencyjne LED oraz diody laserowe LD. Zasada ich działania oparta jest o zjawisko rekombinacji promienistej nośników ładunku. Szczegółowe informacje na temat zasady działania źródeł półprzewodnikowych oraz ich parametrów można znaleźć w literaturze, np. [5, 6, 7, 10, 14]. Parametry źródeł światła są bardzo istotne z punktu widzenia łącza światłowodowego.
Ważnym aspektem jest m.in. kierunkowy rozkład mocy emitowanej przez te źródła. Ma on duże znaczenie przy sprzęganiu źródeł ze światłowodami i warunkuje ilość przekazanej mocy ze źródła do włókna optycznego. Im kąt rozsyłu światła jest mniejszy (wiązka lepiej skolimowana) tym większa jest skuteczność wprowadzania światła do światłowodu.



Rys. 2.8. Definicja parametrów charakterystyki spektralnej źródła światła

Kolejnym ważnym parametrem źródeł światła jest ich charakterystyka spektralna. Ma ona duże znaczenie ze względów transmisyjnych – wpływa m.in. na parametry dyspersyjne toru światłowodowego. Ważniejsze parametry widmowych źródeł światła zostały zdefiniowane na rys. 2.8.

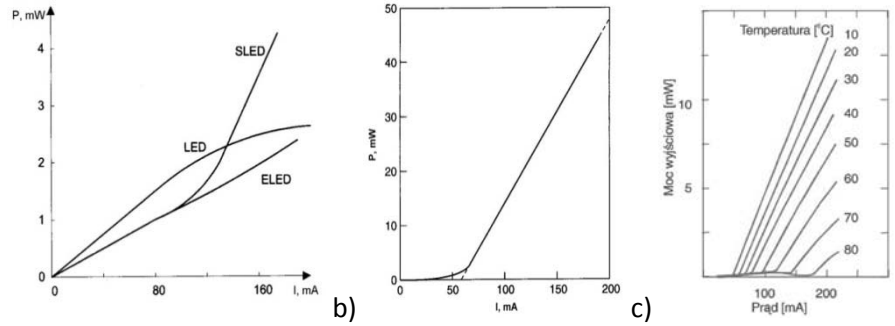
Należą do nich centralna długość fali λ_c (która zwykle pokrywa się ze szczytową długością fali λ_s), połówkowa szerokość spektralna dla połowy wysokości (HWHM – Half Width at Half-Maximum), całkowita szerokość spektralna dla połowy wysokości (FWHM – Full Width at Half-Maximum). Charakterystyka spektralna zależna jest w głównej mierze od materiału z jakiego wykonane jest źródło światła, jednak ulega ona zmianie m.in. pod wpływem temperatury oraz w zależności od wartości płynącego przez nie prądu – widoczne to jest np. na rys. 2.9. Zależność taka jest silniejsza dla laserów półprzewodnikowych niż diod LED.



Rys. 2.9. Charakterystyka widmowa przykładowych laserów półprzewodnikowych: a) wpływ temperatury, b) wpływ przepływającego prądu [19]

Wykorzystanie źródeł światła w telekomunikacji wiąże się z koniecznością modulacji ich natężenia światła. Najczęściej wykorzystywaną metodą modulacji jest modulacja płynącego przez diodę (laser) prąd. Aby było to możliwe oraz by wiedzieć jakie zakresy zmian prądu są możliwe do zastosowania konieczna jest

znajomość zależności mocy optycznej w funkcji prądu (rys. 2.10). Dla laserów półprzewodnikowych zależność ta charakteryzuje się tym, że po przekroczeniu pewnej granicznej wartości prądu (wartość progowa) następuje gwałtowny wzrost mocy.



a) b) c)

Rys. 2.10. Zależność mocy optycznej od prądu dla diod elektroluminescencyjnych [8] (a) i lasera półprzewodnikowego [6] (b) oraz wpływ na nią temperatury (c) [19] (LED – powierzchniowa dioda LED, ELED – krawędziowa dioda LED, SLED – superluminescencyjna dioda LED).

2.5. Łączenie światłowodów

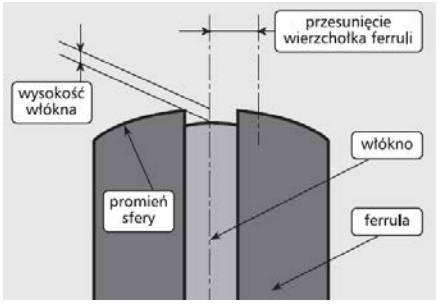
Światłowody można łączyć w sposób trwały (spawanie, klejenie) lub rozłączny. Norma zakładowa TPSA [21] podaje m.in. następujące definicje:

- Złącze światłowodowe - miejsce połączeń światłowodów.
- Złączka światłowodowa - element osprzętu służący do rozłączalnego połączenia światłowodów, składający się zazwyczaj z dwóch wtyków (półzłączek) i tulejki złączowej centrującej (coupler).
- Półzłączka - część wtykowa złączki światłowodowej stanowiąca zakończenie kabla stacyjnego (pigtaila, patchcordu).
- Tulejka centrująca (coupler) - część środkowa złączki światłowodowej służąca do centrycznego połączenia dwóch półzłączek, mocowana na polu przełącznicy.

Na parametry połączenia (straty wtrąceniowe) wpływ ma kilka czynników. Jednym z ważniejszych jest stan powierzchni czołowych obu łączonych włókien. Chodzi tu zarówno o stan powierzchni (jej chropowatość, występowanie rys, itp.), jej położenie względem osi włókna, jak i czystość powierzchni czołowej. Kluczowym zagadnieniem jest również prawidłowe ułożenie łączonych włókien względem siebie. Tłumienie połączenia rośnie w przypadku rozsunięcia powierzchni czołowych wzdłuż czy poprzek osi włókna lub ustawienia światłowodów w taki sposób, że ich osie tworzą niezerowy kąt.

Współczesne spawarki światłowodowe, wykonujące połączenia stałe, przed wykonaniem połączenia bardzo precyzyjnie ustawiają włókna między sobą, aby uniknąć wymienionych niedoskonałości połączeń. W przypadku połączeń rozłącznych właściwe centrowanie włókien jest zachowane poprzez prawidłowe

wykonanie półzłączek oraz couplerów oraz wykorzystanie do ich budowy wysokiej jakości materiałów. Widok zakończenia półzłączki przedstawiony został na rys. 2.11.



Rys. 2.11. Widok końcówki półzłączki światłowodowej z zaznaczonymi wybranymi wymiarami [20]

Zaletą połączeń spawanych nad połączeniami rozłącznymi są znacznie mniejsze straty wtrąceniowe oraz brak występowania odbić Fresnela. To drugie zjawisko może być znacznie ograniczone jeżeli w miejsce tradycyjnych półzłączek PC (z ang. physical contact) o włóknie zakończonym pod kątem prostym do osi włókna wprowadzi się półzłączki APC (ang. angled physical contact), które charakteryzuje się tym, że powierzchnia czołowa włókna tworzy z jego osią kąt 8°. Takie rozwiązanie sprawia, że sygnał ze źródła trafiając na koniec światłowodu nie wraca, a odbija się kierując się w stronę płaszcza i pokrycia.
W tabeli 2.2 zostały zestawione cechy wybranych typów półzłączek.

Tabela 2.2. Cechy wybranych półzłączek światłowodowych.

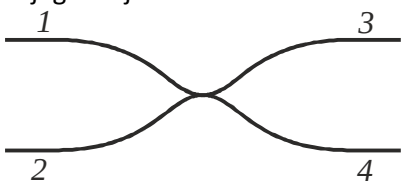
typ	Nazwa	wybrane cechy
SMA	Sub Miniature A	mocowanie gwintowane, bez pozycjonowania włókna
ST	Straight Tip	mocowanie bagnetowe, szybki montaż
FC	Fibre (Ferrule) Connector	mocowanie gwintowane, zabezpieczenie przed obrotem ferruli
SC	Subscriber Connector	lekkie, mechanizm zatrzaskowy
LC	Little (Local, Lucent) Connector	małe wymiary, mechanizm zatrzaskowy, ferrula 1,25mm
E2000		montaż z zatrzaskiem, osłona czoła włókna
LX-5		wąskie, ochrona czoła włókna

Na stosowanie danego typu złączek wpływ ma wiele czynników, począwszy od wymaganych parametrów połączenia (straty, odbicia wsteczne), ilości i częstotliwości wykonywania połączeń, warunków w których pracować będzie złącze (zakres temperatur, obecność drgań, itp.), poprzez możliwości dostępu do niego osób trzecich, aż po zasady przyjęte w firmie (rodzaj posiadanych urządzeń, dostawca sprzętu, itp.). Ważnym zagadnieniem jest również poziom mocy

optycznej jaka przepływa przez łączone światłowody, gdyż jak wykazuje praktyka, duża gęstość mocy może powodować uszkodzenia powierzchni czołowych półzłączy. Przy przesyłaniu wysokich mocy półzłącza muszą spełniać dodatkowe warunki [20]: być wyposażone w mechanizm zabezpieczający przed uszkodzeniem wzroku, elementy optyczne muszą być chronione przez bezpośrednim dostępem zanieczyszczeń i przed dotknięciem, w całej konstrukcji muszą być stosowane tylko materiały przezroczyste optycznie, złącze musi być zabezpieczone przed nieplanowanym rozłączeniem.

2.6. Sprzęgacze światłowodowe

Budowa linii światłowodowych wymaga m.in. stosowania w nich tzw. sprzęgaczy. Mają one za zadanie łączyć lub dzielić sygnał optyczny umożliwiając tym samym transmisję jednym włóknem wielu kanałów. Istnieje wiele konstrukcji sprzęgaczy światłowodowych różniących się od siebie technologią wykonania, parametrami oraz właściwościami. Klasycznym przykładem sprzęgacza jest sprzęgacz typu X, którego schemat ideowy przedstawiony jest na rys. 2.12, na którym oznaczono również jego wejścia.



Rys. 2.12. Sprzęgacz typu X

Zadaniem sprzęgacza jest przekazanie do wyjść 3 (wyjście główne) i 4 (wyjście sprzęgane) sygnału optycznego wprowadzonego do wejścia 1. Przy takim pobudzeniu sprzęgacza na wyjściach 2 oraz 1 pojawią się sygnały wynikające z rozproszenia wstecznego. Oczywiście do dowolnego z wejść sprzęgacza może zostać wprowadzony sygnał optyczny, wtedy na odpowiednich wyjściach pojawią się sygnały wyjściowe oraz rozproszone.
Podział mocy między wyjścia (główne i sprzężone) określa tzw. współczynnik sprzężenia (wyrażony w % lub dB):

$$k_s = \frac{P_4}{P_3}100\% \text{ lub } k_s = 10log\frac{P_4}{P_3}\,[dB]$$

Do innych parametrów sprzęgacza należą [15, 18]:

- straty wtrąceniowe określające poziom strat mocy w wyniku przejścia przez sprzęgacz

$$\alpha_w = 10log\frac{P_3+P_4}{P_1}\,[dB],$$

- straty przejścia określające spadek mocy przy przejściu sygnału z wejścia do wyjścia głównego

$$\alpha_p = 10log\frac{P_3}{P_1}\,[dB],$$

- straty sprzężenia (podziału) wyrażające stosunek mocy na wyjściu sprzężonym do mocy na wejściu

$$\alpha_s = 10log\frac{P_4}{P_1}\,[dB],$$

- współczynnik izolacji (zwany również kierunkowością sprzężenia) opisujące w dB ilość mocy przenikającej do wyjścia 2 przy pobudzeniu wejścia 1

$$\alpha_i = 10log\frac{P_2}{P_1}\,[dB].$$

Można wykonać tak sprzęgacz, że wymienione parametry, a zwłaszcza współczynnik sprzężenia będą silnie zależeć od długości fali. W takim przypadku otrzymuje tzw. sprzęgacz WDM. Głównym ich zadaniem jest kierowanie sygnałów o określonych zakresach widmowych do określonych portów. Pozwala to np. na łączenie w jednym włóknie sygnałów z kilku źródeł, przesyłanie ich bez wpływu na pozostałe kanały, a następnie rozdzielanie bez obecności przesłuchów międzykanałowych.

2.7. Łącze światłowodowe

Głównym obszarem wykorzystania światłowodów jest telekomunikacja. Dzięki włóknom optycznym można przesyłać sygnały analogowe lub cyfrowe na znacznie większe odległości i z o wiele większymi przepływnościami, niż dla łączy miedzianych.

Tor transmisyjny ma m.in. zapewnić na wyjściu sygnał o mocy wystarczającej do jego poprawnego odebrania. W tym celu, podczas projektowania łącza konieczne jest przeprowadzenie tzw. bilansu mocy. W celu przeprowadzenie bilansu mocy należy obliczyć straty całkowite występujące w łączu światłowodowym obejmujące straty wynikające z tłumienności włókien, tłumienia wnoszonego przez połączenia (stałe i rozłączne), straty sprzężenia źródła światła z włóknem oraz włókna z detektorem oraz margines bezpieczeństwa (uwzględniający przyszły wzrost tłumienia, np. na skutek zjawisk starzeniowych).

Znajomość strat występujących w łączu pozwala na:

- wybór detektora przy danym źródle światła (o określonej mocy),
- wybór źródła światła przy danym odbiorniku (o określonej czułości),
- obliczenie długości odcinków międzyregeneratorowych.

Bilans mocy przy znanych parametrach źródła i detektora, ilości i rodzaju połączeń, stratach innych elementów toru (np. sprzęgaczy) pozwala na dobór właściwych światłowodów (pod względem tłumienności).

W przypadku obliczania bilansu mocy, moc źródła światła dogodnie jest podawać nie w watach, a w decybelach odniesionych do 1mW, czyli tzw. dBm. Przeliczenie mocy źródła światła z W na dBm odbywa się na podstawie wzoru:

$$P[dBm] = 10log\frac{P[W]}{1mW}.$$

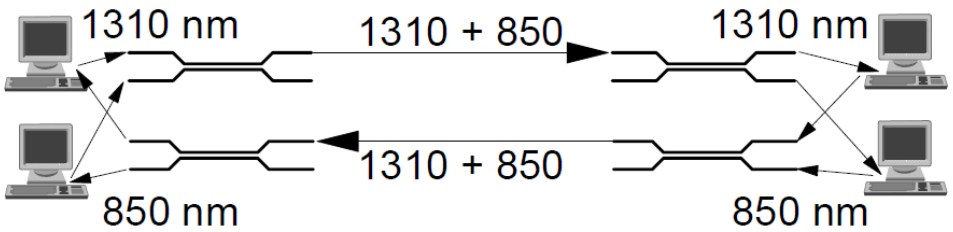
Pozwala to na łatwe obliczenia, gdyż od tak wyrażonej mocy można wprost odejmować tłumienie odcinka światłowodu, czy tłumienie złączy. Również czułość detektorów podaje się w tych jednostkach.

2.8. Techniki zwielokrotniania kanałów komunikacyjnych

W chwili obecnej transmisja sygnału tylko między dwoma odbiorcami nie jest w stanie w pełni wykorzystać możliwości światłowodu. Z tego powodu stosowane są techniki zmierzające do lepszego wykorzystania go. W tym celu stosuje się tzw. techniki zwielokrotniania, czyli metody pozwalające na realizację większej liczby kanałów komunikacyjnych w bazie jednego kanału fizycznego. W stosunku do światłowodu mogą to być m.in. techniki:

- TDM – (ang. time division multiplexing) – multipleksacja w dziedzinie czasu – sygnały z różnych nadajników są przesyłane tym samym kanałem w wyznaczonych chwilach (szczelinach) czasowych,
- FDM – (ang. frequency division multiplexing) – multipleksacja w dziedzinie częstotliwości – sygnały z różnych źródeł są przesyłane w tym samym czasie na różnych falach nośnych,
- WDM – (ang. wavelength division multiplexing) – multipleksacja w dziedzinie długości fali – sygnały z różnych źródeł są przesyłane w tym samym czasie na różnych falach nośnych,
- CDM – (ang. code division multipleksing) – multipleksacja kodowa – sygnały z wielu nadajników są kodowane unikatowymi kodami i przesyłane w jednym czasie przez medium transmisyjne.

Techniki WDM oraz FDM są podobne w idei, ale ze względu na różnice w wielkości odstępu między transmitowanymi kanałami dokonuje się ich rozróżnienia.
Od lat 90-tych XX wieku w technice światłowodowej zwielokrotnienie oparte jest głównie o technikę WDM, ewentualnie WDM wraz z TDM [12], co związane jest głównie z dwoma czynnikami: ograniczeniem uzyskiwanych przepływności w systemach TDM związanym z ograniczoną szybkością pracy układów elektronicznych oraz znacznie mniejszymi kosztami wdrażania systemów WDM niż TDM przy zwiększających się przepływnościach.



Rys. 2.13. Idea systemu WDM [4]

Ideę systemu WDM przedstawia rysunek 2.13. W jednym włóknie optycznym przesyłanych jest wiele (nawet do tysiąca) długości fal, przy czym każda jest nośnikiem sygnału z innego źródła. Sygnały te mogą być przesyłane w światłowodzie w tą samą stronę (jak na rys. 2.13), ale możliwe jest rozwiązanie, w którym fale biegną w strony przeciwne. Stosowanie techniki WDM wymaga

stosowania specjalizowanych elementów toru komunikacyjnego, np. źródeł światła (wysokie wymagania odnośnie monochromatyczności i stałości parametrów w czasie i z temperaturą), sprzęgaczy (czułe na długość fali), złączek (pozwalających na transmisję dużych mocy), wzmacniaczy optycznych (charakterystyka wzmocnienia skorelowana z charakterystyką tłumienia toru dla wszystkich kanałów). Włókna wykorzystywane w tych torach muszą być optymalizowane m.in. pod względem wartości dyspersji, i to dla wszystkich kanałów.

Istnieje wiele odmian techniki WDM różniące się ilością kanałów optycznych, odległością między nimi, a co za tym idzie wymaganiami co do elementów toru światłowodowego. Można więc wyróżnić systemy [12]: NWDM (narrow WDM – wąskopasmowe), CWDM (coarse WDM – zgrubne), DWDM (dense WDM – gęste), UDWDM (ultradense WDM – ultragęste) zwane również HDWDM (high DWDM – wysoki/silny DWDM). Najpowszechniejsze stosowanie mają techniki CWDM oraz DWDM.

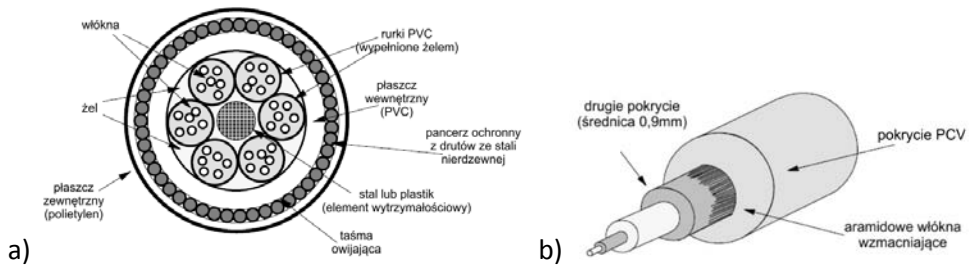
Zgodnie ze standardem ITU G.694.2 w technice CWDM wykorzystuje się 18 kanałów w zakresie długości fal od 1270 nm do 1610 nm. Zwykle wykorzystuje się 8 kanałów o odstępie 20nm w zakresie III okna transmisyjnego. Takie rozwiązanie pozwala na wykorzystanie źródeł światła niestabilizowanych temperaturowo (a więc stosunkowo tanich). Ze względu na wykorzystywane pasmo optyczne nie jest możliwe stosowanie wzmacniaczy optycznych – do wzmacniania sygnału stosuje się wzmacniacze elektroniczne.

Technika DWDM wykorzystuje znacznie węższe kanały i leżące w mniejszych odległościach od siebie. Standardowy odstęp międzykanałowy wynosi 100GHz (około 0,8nm), dopuszcza się odstępy stanowiące wielokrotność 100GHz. np. 200GHz, 400GHz oraz odstępy połówkowe: 50GHz, 25GHz. Do transmisji tego typu wykorzystuje się pasmo C. Ze względu na mniejsze odstępy międzykanałowe, wymagania stawiane źródłom światła są wyższe niż dla CWDM – lasery i elementy filtrujące muszą mieć bardziej stabilne charakterystyki widmowe. Ze względu na przesyłanie jednym włóknem większej mocy (wynikającej z dużej liczby kanałów) konieczne jest uwzględnianie w tych systemach efektów nieliniowych (m.in. mieszanie czterofalowe).

2.9. Kable światłowodowe

Telekomunikacyjne włókno optyczne, ze względu na swoje rozmiary i właściwości materiału, z którego jest wykonane, jest stosunkowo nietrwałe. Każdy światłowód jest na etapie produkcji pokrywany polimerem, który stanowi pierwszą ochronę włókna przed oddziaływaniem środowiska zewnętrznego oraz ochronę mechaniczną (średnica zewnętrzna 250μm). Ochrona ta jednak nie jest wystarczająca. Konieczne jest więc jego dodatkowe zabezpieczanie, stąd włókna optyczne otacza się dodatkowymi warstwami tworząc kable. Istnieje wiele rodzajów kabli światłowodowych różniących się między sobą ilością włókien optycznych, ilością warstw, ich grubością, wykorzystywanymi materiałami, sposobem wypełnienia przestrzeni między włóknami, warunkami pracy, itp. Można

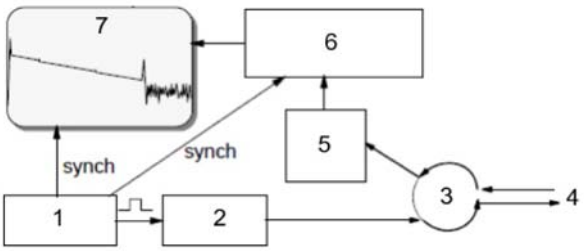
powiedzieć, że konstrukcja kabla może być dowolna – zależna od potrzeb użytkownika.
Kable można podzielić ogólnie na wewnątrzobiektowe i zewnątrzobiektowe. W zależności od miejsca montażu kabla występują inne czynniki zagrażające, np. wahania temperatur, możliwość wystąpienia obciążeń mechanicznych, zawilgocenie, niebezpieczeństwo pożaru, itp.
Kable zewnątrzobiektowe (rys. 2.14 a) wykonuje się jako tzw. kable z luźną tubą. W kablach takich włókna ułożone są swobodnie. Pozwala to na unikanie naprężeń we włóknach wywołanych wydłużaniem kabla pod wpływem zmiany temperatury zewnętrznej (np. sezonowych lub związanych z porą dnia). Przestrzenie między nimi wypełnione są cieczą (żelem) o właściwościach hydrofobowych – chroniących przed degradującym działaniem wody w przypadku uszkodzenia kabla, oraz tiksotropowych co zapewnia ochronę włókien przed mikrozgięciami. Kable te z racji możliwych zagrożeń mogą być otaczane kolejnymi warstwami, włączając w to warstwy metalowe. Warstwy zewnętrzne kabli mogą być również wykonywane w wersji odpornej na gryzonie (specjalnie dobrane rodzaje PCV lub ocynkowana stal).
Kable wewnątrzobiektowe (rys. 2.14 b) to zwykle kable ze ścisłą tubą. Włókna w tych kablach umieszczane są w ściśle przylegających rurkach, których liczba i grubość zależy od innych wymagań.



Rys. 2.14. Schemat budowy kabla światłowodowego: luźna tuba (a), ścisła tuba (b) [4]

2.10. Reflektometr i pomiary reflektometryczne

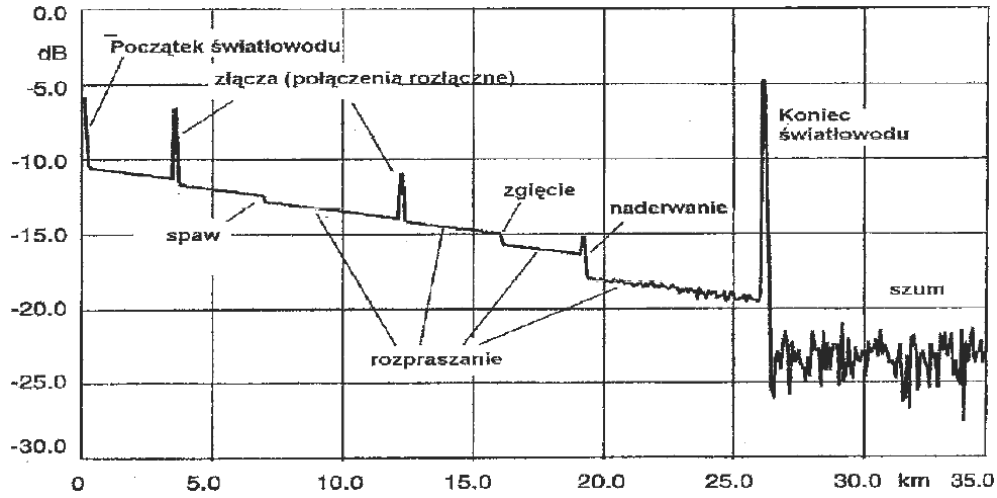
Reflektometr światłowodowy (ang. Optical Time-Domain Reflectometer – OTDR) jest urządzeniem, które pozwala na nieniszczące badania torów optycznych, a zwłaszcza ich długości i tłumienia oraz wykrywania zdarzeń (takich jak spawy, złączki) i uszkodzeń w analizowanych torach. Schemat budowy reflektometru przedstawiony został na rys. 2.15.



Rys. 2.15. Schemat budowy reflektometru; (1 – generator ipmlulsów, 2 – laser, 3 – cyrkulator, 4 – badany światłowód, 5 – detektor ze wzmacniaczem, 6 – przetwornik a/c z integratorem, 7 - wyświetlacz) [4]

Zasada działania reflektometru jest następująca. Impuls świetlny z lasera, poprzez sprzęgacz optyczny (lub cyrkulator) zostaje wprowadzony do badanego światłowodu. Rozproszone wstecznie promieniowanie powraca do reflektometru i przez sprzęgacz optyczny (cyrkulator) trafia do detektora. Sygnał rozproszenia jest bardzo słaby więc sygnał wyjściowy detektora (elektryczny) jest wzmacniany, a następnie przetwarzany na postać cyfrową. Kolejnym członem reflektometru jest integrator, którego zadaniem jest sumowanie sygnałów pochodzących ze wstecznego rozproszenia kolejnych impulsów światła. Niekiedy przetwarzanie sygnału na postać cyfrową i jego uśrednianie są realizowane we wspólnym układzie nazywanym po angielsku boxcar averager.

Oprócz pomiaru mocy sygnału powracającego mierzony jest czas t jaki upływa od chwili wygenerowania impulsu lasera do momentu odebrania sygnału rozproszonego. Czas ten jest równy pokonaniu przez impuls światła drogi L od źródła do miejsca, w którym impuls się odbił, i z powrotem. Przy znajomości czasu t oraz prędkości v rozchodzenia się światła w światłowodzie $\left(v = \frac{c}{n}\right)$ możliwe jest obliczenie odległości L, w której nastąpiło wsteczne rozproszenie światła. Zależność różnicy mocy wejściowej i powracającej (wyrażonej w decybelach) od odległości nazywana jest reflektogramem (rys. 2.16).



Rys. 2.16. Reflektogram przykładowej linii światłowodowej z zaznaczeniem rozpoznanych zjawisk [11]

Widoczne na reflektogramie podbicia mocy są wynikiem silnych odbić Fresnela powstających na granicy szkło-powietrze, a wynikających z różnicy współczynników załamania obu tych ośrodków. Świadczą więc one o występowaniu w torze światłowodowym miejsc z przerwą powietrzną, czyli złącza rozłącznego, naderwania włókna oraz zakończenia światłowodu. Rozróżnienie, która z tych sytuacji ma miejsce może jest dokonywane na podstawie straty mocy przed i po zlokalizowanym zdarzeniu.

Dla rozłącznego złącza światłowodowego (dwie powierzchnie odbijające) straty te wyraża wzór:

$$\alpha_z = 2 \cdot 10log\left[1 - \left(\frac{n_r - n_o}{n_r + n_o}\right)^2\right][dB],$$

gdzie n_r - współczynnik załamania rdzenia, n_o - współczynnik załamania ośrodka między czołami światłowodów. Typowo strata ta wynosi ok. 0,3dB. W przypadku naderwania włókna strata ta jest większa. Dla końca włókna, po wystąpieniu odbicia Fresnela obserwuje się szumy o poziomie mocy znacznie niższym niż przed końcem światłowodu (rys. 2.16). Odbicia Fresnela mogą nie być obserwowane dla końca włókna w przypadku, gdy jest ono urwane. Nie ma wtedy powierzchni prostopadłej do biegnącego sygnału, a więc odbite promieniowanie nie wraca do reflektometru.

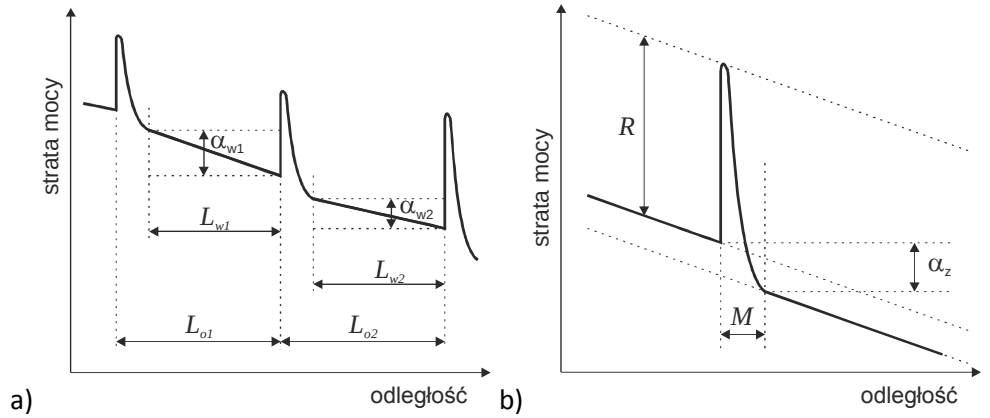
Należy zauważyć, że tuż za miejscami, dla których następuje znacząca zmiana mocy (zwłaszcza dla odbić Fresnela), przebieg reflektogramu ma kształt malejącej funkcji wykładniczej. Jest to wynikiem „olśnienia" detektora po odebraniu przez niego dużej mocy. Detektor po oświetleniu go znaczną mocą wchodzi w stan nasycenia, a wyjście z tego stanu trwa pewien czas zależny od jego parametrów i czasu trwania impulsu laserowego. W czasie nasycenia detektor nie rejestruje docierających do niego sygnałów, co przekłada się na występowanie tzw. strefy martwej, w której

reflektometr nie jest w stanie wykrywać żadnych zdarzeń. Strefa ta może mieć różną długość (zależną od czasu odzyskiwania właściwości przez detektor) od kilku metrów do nawet kilometra. Strefa martwa jest również obserwowana na początku toru światłowodowego i jest wynikiem pobudzenia we włóknie modów płaszczowych i wyciekających. Aby dokonać pomiarów, na początkowym odcinku toru światłowodowego stosuje się więc tzw. włókno rozbiegowe (potocznie nazywane: rozbiegówka), o długości większej niż strefa martwa.

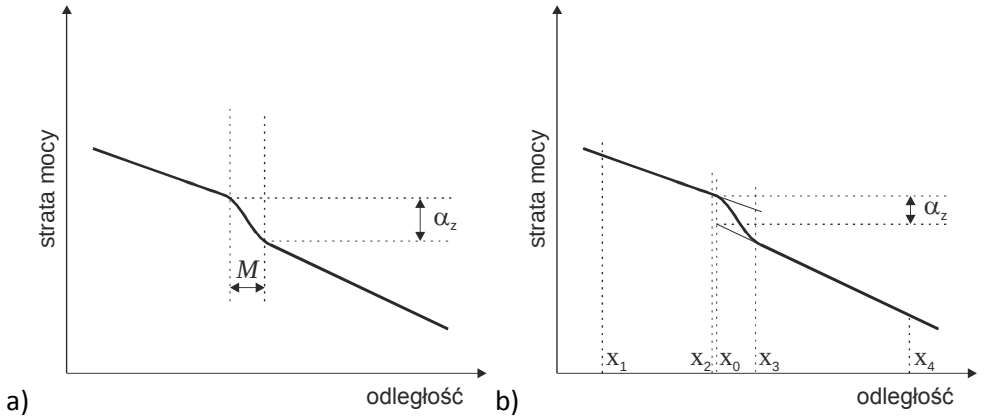
Sposób wyznaczania na podstawie pomiarów tłumienia częściej spotykanych zdarzeń oraz innych ich parametrów, został przedstawiony na rys. 2.17-2.19. Rys. 2.17a przedstawia jak dokonuje się pomiaru długości światłowodu oraz jak wyznacza się jego tłumienie i tłumienność. Należy zwrócić uwagę na to, że mierząc tłumienność włókna należy dokonywać pomiarów poza strefą martwą występującą na złączach końcowych tego włókna.

Sposób oceny reflektancji złącza oraz strefy martwej i tłumienia złącza rozłącznego przedstawiono na rys. 2.17b. Natomiast rysunek 2.18 pokazuje dwie metody wyznaczenia tłumienia połączenia. Przykład dotyczy połączenia nierozłącznego, ale takie same metody można stosować do połączeń rozłącznych, a także pomiaru tłumienia odcinka światłowodu.

Metoda dwupunktowa polega na wyborze 2 punktów na reflektogramie i obliczeniu różnicy ich tłumienia. Metoda najmniejszych kwadratów (LSA) polega na przybliżeniu charakterystyk włókien przed i za złączem liniami prostymi wyznaczonymi przez pary punktów x_1 i x_2 oraz x_3 i x_4 oraz obliczenia różnicy tłumienia punktów powstałych z przecięcia ekstrapolacji tych prostych i prostej przechodzącej przez punkt x_0. Uzyskanie poprawnych wyników wymaga by punkty x_2 i x_0 leżały blisko siebie i blisko początku złącza, a punkt x_3 blisko końca tego złącza. Punkty x_1 i x_4 powinny leżeć możliwie daleko od złącza, jednak nie mogą leżeć za kolejnymi zdarzeniami w torze.

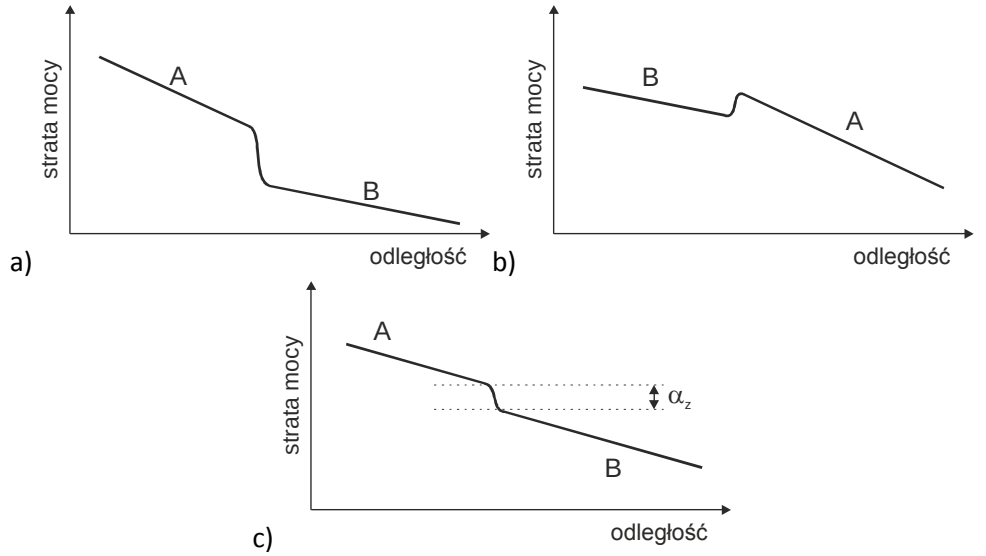


Rys. 2.17. Sposób wyznaczenia: a) długości odcinków światłowodowych L_{o1} i L_{o1}, tłumienia tych odcinków α_{w1} i α_{w2} oraz ich tłumienności równej: α_{w1}/L_{w1} i α_{w2}/L_{w2}; b) tłumienia połączenia rozłącznego α_z, reflektancji R i długości strefy martwej M.



Rys. 2.18. Sposób wyznaczenia tłumienia α_z połączenia spawanego oraz strefy martwej M: a) metodą dwupunktową (2p) i b) metodą najmniejszych kwadratów (LSA)

Ciekawy przypadek jest zaprezentowany na rys. 2.19. Podczas pomiarów może zdarzyć się taka sytuacja, że w miejscu spawu zamiast spadku mocy może pojawiać się jej podbicie. Świadczy to o połączeniu dwóch światłowodów o różnych właściwościach rozproszeniowych. Może to być, np. połączeni światłowodu o małej i dużej średnicy rdzenia. Prawidłowe wyznaczenie tłumienia takiego spawu jest możliwe dopiero po wyznaczeniu reflektogramu dla pomiaru z jednego (rys. 2.19a) i drugiego (rys. 2.19b) końca włókna. Wyniki tak uzyskanych pomiarów należy uśrednić.



Rys. 2.19. Sposób wyznaczenia tłumienia α_z połączenia spawanego dla włókien o różnych współczynnikach rozproszenia wstecznego: a) reflektogram dla pomiaru z jednego końca włókna, b) pomiar dla drugiego końca włókna, c) wynik prawidłowy (średnia arytmetyczna z pomiarów z punktu a i b)

Literatura

1. Bjarklev A. at all: Fiber Based Dispersion Compensation, Springer 2007
2. Crisp J.: Introduction to Fiber Optics (second edition). Newnes, 2001
3. Dorosz J.: Technologia światłowodów włóknistych. Ceramika Vol. 86, Kraków 2005
4. Dutton H.: Understanding Optical Communications, International Technical Support Organization, 1998
5. Einarsson G.: Podstawy telekomunikacji światłowodowej. WKŁ, Warszawa 1998
6. Gupta M. C., Ballato J.: The handbook of photonics- second edition. CRC Press 2007
7. Holejko K.: Optyczne sieci telekomunikacyjne. Podstawy telekomunikacji światłowodowej. XIII Krajowa Szkoła Optoelektroniki. Wydawnictwo POLSOFT, Poznań 1998
8. Kaczmarek Z.: Światłowodowe czujniki i przetworniki pomiarowe. Agenda Wydawnicza PAK, Warszawa 2006
9. Majewski A.: Rozchodzenie się ultrakrótkich impulsów w światłowodzie. Przegląd Telekomunikacyjny 9/97
10. Marciniak M.: Łączność światłowodowa. WKŁ, Warszawa 1998
11. Perlicki K.: Pomiary w optycznych systemach telekomunikacyjnych. WKŁ, Warszawa 2002
12. Perlicki K.: Systemy transmisji optycznej WDM, WKŁ, Warszawa 2007
13. Sano A. at all: 69.1-Tb/s (432 × 171-Gb/s) C- and extended L-band transmission over 240 km Using PDM-16-QAM modulation and digital coherent detection, Optical Fiber Communication Conference (OFC), San Diego 2010, paper PDPB7
14. Siuzdak J.: Systemy i sieci fotoniczne. WKŁ, Warszawa 2009
15. Szustakowski M.: Elementy techniki światłowodowej. WNT Warszawa 1992
16. Wójcik W.: Optoelectronic Diagnostics Of Combustion Processes – Instruments, Methods, Applications, Polska Akademia Nauk, Komitet Inżynierii Środowiska, Monografie Nr 48, 2008
17. Intercorporation International Journal Of Optical Sensors. vol. 3 no. 1, 1988, pp 58
18. Materiały do laboratorium: http://wwwold.wemif.pwr.wroc.pl/swiatlowody/
19. Materiały do wykładów http://www.fizyka.umk.pl/~bezet/pdf/FL
20. Materiały firmy FCA Sp. z o.o. (www.fca.com.pl)
21. Norma zakładowa TPSA ZN-96

3. Systemy multimedialne

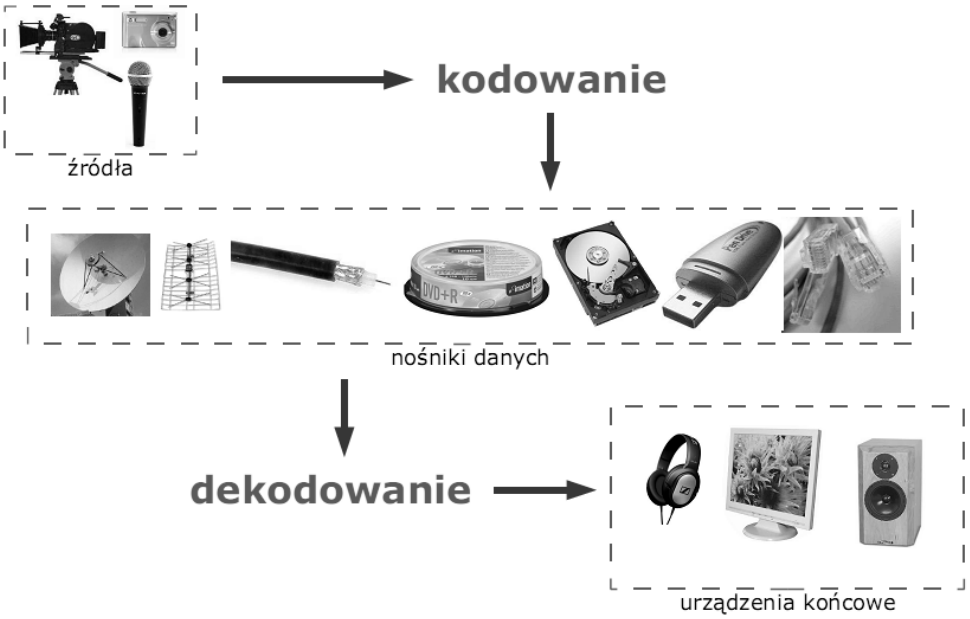
3.1. Co to jest system multimedialny?

Rozwój usług sieciowych w dużej mierze wynika z wprowadzenia elementów przesyłu danych o charakterze multimedialnym. Oprócz normalnych danych przesyłanych w sieci, takich jak ftp czy też WWW, coraz większą rolę odgrywają właśnie dane multimedialne. Możliwość przesyłu głosu stała się podstawą do powstania nowej usługi VoIP (ang. Voice over IP), a z punktu widzenia usługodawców sieciowych, do zaoferowania nowej formy komunikacji – telefonii internetowej. To co jest atrakcyjne z punktu widzenia klienta (znaczne obniżenie kosztów rozmów telefonicznych, „mobilność" terminali telefonicznych), nie zawsze jest proste do zaoferowania przez operatorów sieciowych.

Multimedia stanowią zatem połączenie wielu mediów, a jedną z istotnych cech przekazu multimedialnego jest interakcja z użytkownikiem, pojawiająca się w większości aplikacji. Komputer multimedialny, odtwarzający wysokiej jakości dźwięk, grafikę i filmy, stanowi doskonałe narzędzie wspomagające działalność edukacyjną i biznesową, a jednocześnie oferuje wiele możliwości rozrywki, np. poprzez efektowne gry komputerowe [1].

W słowie „multimedialny" zawiera wyraz „multi", zatem zakłada ono wielość mediów w takim systemie (np. obraz i dźwięk). Systemy dostarczające tylko jedno medium (np. tylko dźwięk), również można nazwać podsystemami multimedialnymi, bo słowo „medialny" znaczy w naszym języku coś zupełnie innego [2].

Aplikacje multimedialne znajdują coraz więcej zastosowań i można się z nimi spotkać praktycznie wszędzie. W zastosowaniach komercyjnych multimedia służą m.in. do przedstawiania informacji klientom firm, np. w formie multimedialnych kiosków informacyjnych z wirtualnymi przyciskami na monitorze dotykowym i wirtualną klawiaturą. Aplikacje multimedialne ułatwiają naukę personelu, a także promocję różnego rodzaju produktów. Wiele aplikacji korzysta z Internetu, np. telewizja interaktywna, sklepy wirtualne czy telefon przez Internet.



Rys. 3.1. Elementy systemu multimedialnego [2]

Zakupy internetowe czy wybieranie filmów na życzenie przez Internet stanowią istotne udogodnienie, szczególnie dla osób, dla których wyjście z domu jest problemem. Wideokonferencje umożliwiają przeprowadzenie dyskusji między zainteresowanymi osobami mimo dzielących uczestników odległości, których pokonanie byłoby kosztowne, niepraktyczne czy wręcz niemożliwe. Multimedia na pokładach samolotów podczas dłuższych lotów pozwalają na obserwację przebiegu lotu, a także relaks przy oglądaniu filmów, wiadomości, słuchaniu audycji muzycznych itp. [3].

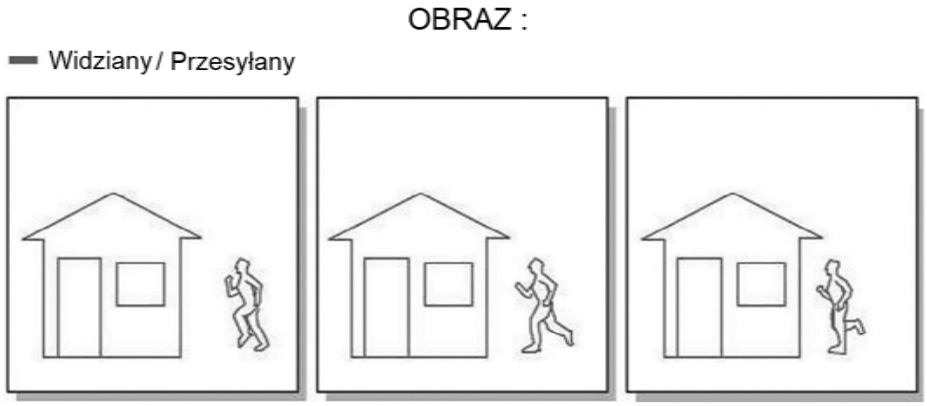
3.2. Standardy kompresji wideo

Jakość obrazu jest jednym z ważniejszych parametrów każdej kamery, jeśli nie najważniejszym. Ma to duże znaczenie szczególnie w przypadku dozoru i zdalnego monitoringu w miejscach, gdzie zagrożone może być czyjeś życie i mienie. W przeciwieństwie do tradycyjnych kamer analogowych, kamery sieciowe mają moc obliczeniową wykorzystywaną nie tylko do rejestracji i wyświetlania obrazu, ale też do zarządzania obrazem i cyfrowego kompresowania go w celu przesyłania przez sieć. Podstawą wszystkich metod kompresji stosowanych dzisiaj w systemach rejestrujących jest standard kompresji JPEG (ang. Joint Photographics Expert Group). Bezpośrednim „potomkiem" kompresji JPEG stosowanym do zapisu materiału video jest standard M–JPEG, w którym przedrostek M oznacza słowo

Ruchomy (Motion). Wszystkie pozostałe mechanizmy kompresji, w tym również szeroko stosowany format MPEG (ang. Moving Pictures Expert Group), bazują na algorytmach stosowanych w standardzie JPEG wzbogacając go o szereg nowych rozwiązań. Poniżej opisano kilka standardów najczęściej stosowanych w sieciowych systemach dozoru wizyjnego.

JPEG – algorytm kompresji JPEG, jest zaliczany do algorytmów stratnych, w wyniku działania których pewna liczba pierwotnych informacji jest bezpowrotnie tracona. Przy zastosowaniu niskiego poziomu kompresji ilość traconych danych jest na tyle niewielka, że oko ludzkie nie jest w stanie dostrzec ich braków [4]. Kompresja JPEG może być wykonywana na różnych, określanych przez użytkownika poziomach, precyzujących stopień skompresowania obrazu. Wybrany stopień kompresji ma bezpośredni związek z jakością obrazu. Także sam obraz ma wpływ na wynikowy współczynnik kompresji. Na przykład biała ściana może dać w efekcie względnie mały plik obrazu (z wyższym współczynnikiem kompresji), natomiast ten sam stopień kompresji zastosowany do bardzo zróżnicowanej sceny da w efekcie plik o większym rozmiarze i z mniejszym stopniem kompresji [5].

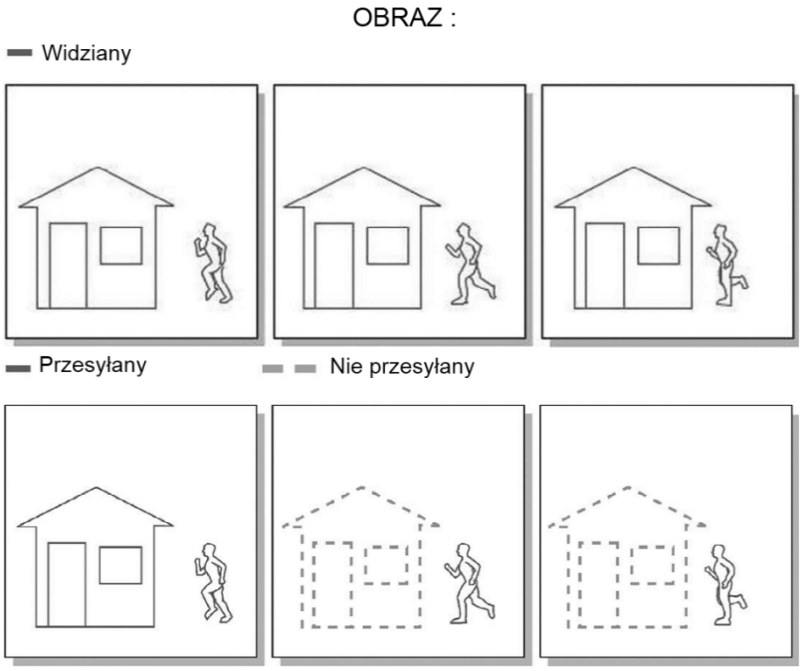
Motion JPEG (M–JPEG) – jest standardem najczęściej używanym w sieciowych systemach dozoru wizyjnego. Kamera cyfrowa rejestruje pojedyncze obrazy i kompresuje je do formatu JPEG. Może np. rejestrować i kompresować 30 pojedynczych zdjęć na sekundę (30 fps – obrazów na sekundę), a następnie udostępniać je jako ciągły strumień obrazów w sieci lub stacji wyświetlającej. Przy szybkości ok. 16 fps i więcej, obserwator ma wrażenie oglądania filmu dlatego ta metoda została nazwana Motion JPEG [6]. Na rys. 3.2 przedstawiono przykładową sekwencję trzech pełnych obrazów JPEG.



Rys. 3.2 Sekwencja trzech pełnych obrazów JPEG [6]

Poszczególne zdjęcia są skompresowanymi zdjęciami JPEG, więc mają tą samą gwarantowaną jakość określoną przez stopień kompresji wybrany dla kamery sieciowej lub serwera wizyjnego. Powszechnie stosowany w wielu systemach standard Motion JPEG ze względu na swoją prostotę najczęściej stanowi dobre rozwiązanie. Występuje w nim jedynie niewielkie opóźnienie między rejestracją obrazu przez kamerę, kodowaniem, przesyłaniem przez sieć i wreszcie wyświetlaniem w stacji monitorowania. Innymi słowy, Motion JPEG zapewnia małe opóźnienie dzięki swojej prostocie (kompresja obrazu i kompletowanie pojedynczych zdjęć), jest więc użyteczny do przetwarzania obrazu, np. wykrywania ruchu lub śledzenia obiektów. System ten zapewnia jakość obrazu niezależną od ruchu lub złożoności obrazu, a zarazem możliwość wyboru wysokiej (z niską kompresją) lub niskiej (z wysoką kompresją) jakości obrazu, z mniejszymi rozmiarami plików obrazu, a więc mniejszą szybkością transmisji i mniejszym wykorzystaniem pasma. Wadą standard Motion JPEG jest generowanie dość dużej ilość danych obrazów przesyłanych przez sieć.

Podstawową zasadą grupy standardów MPEG jest porównywanie dwóch skompresowanych obrazów, które mają być przesłane przez sieć. Pierwszy skompresowany obraz jest używany jako obraz referencyjny (odniesienia). Następnie przesyłane są tylko te fragmenty kolejnych obrazów, które różnią się od obrazu referencyjnego. Sieciowa stacja wyświetlania rekonstruuje wszystkie obrazy na podstawie obrazu referencyjnego oraz „danych różnicowych”[6]. Zilustrowano to na rys. 3.3.



Rys. 3.3. Sekwencja trzech obrazów w standardzie MPEG [5]

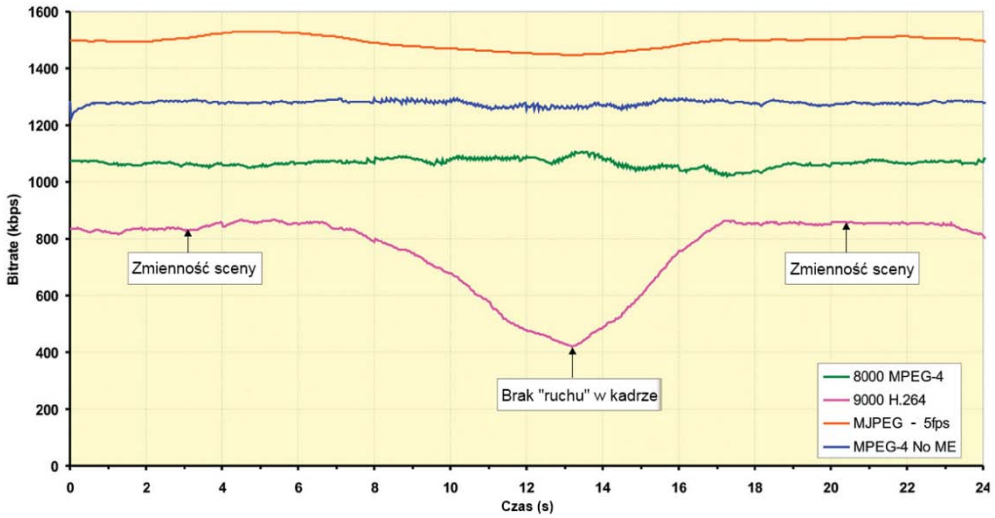
W sieci przesyłane są tylko informacje o różnicach w drugiej i trzeciej klatce. Standard ten jest dużo bardziej skomplikowany niż M–JPEG często wykorzystuje dodatkowe techniki lub narzędzia dla takich parametrów jak przewidywanie ruchu w danej scenie czy identyfikowanie obiektów. Zaletą standardu MPEG–4 jest zastosowanie o wiele większej liczby narzędzi obniżających szybkość transmisji potrzebnej do osiągnięcia określonej jakości obrazu dla danego zastosowania lub scenerii obrazu. Ponadto liczba klatek na sekundę nie jest ustalona na 25/30 fps jak to było np. w standardzie MPEG–1 czy MPEG–2. Większość narzędzi obniżających szybkość transmisji nadaje się jednak obecnie tylko do zastosowań „nie w czasie rzeczywistym". Dzieje się tak dlatego, że niektóre z tych nowych narzędzi wymagają tak dużej mocy obliczeniowej, że ogólny czas kodowania i dekodowania (tj. opóźnienie) sprawia, że nie sprawdzają się w zastosowaniach innych niż kodowanie filmów studyjnych, animowanych itp. Kolejną zaletą MPEG jest przesyłanie przez sieć mniejszej ilości danych w jednostce czasu (szybkość transmisji) w porównaniu z Motion JPEG. Wyjątkiem jest mniejsza liczba obrazów na sekundę. Jeżeli przepustowość sieci jest ograniczona lub film ma być nagrany z dużą liczbą obrazów na sekundę przy występujących ograniczeniach miejsca na zapis danych, MPEG może stanowić najlepsze rozwiązanie. Standard ten zapewnia względnie wysoką jakość obrazu przy mniejszej szybkości transmisji (wykorzystania przepustowości).

Mniejsza przepustowość wymaga jednak większej złożoności kodowania i dekodowania, co z kolei przyczynia się do większego opóźnienia w porównaniu z Motion JPEG [6]. Oczywiście przy mniejszych liczbach obrazów na sekundę, gdy kompresja MPEG–4 nie może wykorzystać w większym stopniu podobieństw między sąsiednimi klatkami, jak też z powodu obciążenia powodowanego przez format przesyłania strumieniowego, wykorzystanie przepustowości przez MPEG–4 jest porównywalne z Motion JPEG. Poważną wadą standardu MPEG–4 w porównaniu do Motion JPEG jest fakt, że podlega on opłatom licencyjnym.

H.264 – Dwie grupy H.263 i MPEG–4 połączyły się w celu stworzenia standardu kompresji wideo następnej generacji. Jest to standard AVC (Advanced Video Coding – zaawansowane kodowanie sygnałów wizyjnych), nazywany też H.264 lub MPEG–4 Part 10. Jego zadaniem jest uzyskiwanie bardzo dużej kompresji danych. Główną korzyścią ze stosowania H.264 jest ograniczenie pasma i zajętości miejsca na dysku w określonych, raczej dobrych warunkach [7]. Najistotniejszą różnicą pomiędzy opisywanymi dotychczas metodami kompresji obrazów a standardem H.264 jest wysoka wydajność tego procesu. W stosunku do metod MPEG–2 i MPEG–4 H.264 zapewnia lepszą jakość obrazu przy porównywalnym strumieniu zakodowanych danych wyjściowych, albo skuteczniejszą kompresję strumienia danych przy podobnej jakości obrazu [8]. Standard H.264 jest raczej rozszerzeniem istniejących standardów niż nową, rewolucyjną koncepcją kodowania danych. Różnica polega na dalszej komplikacji działania algorytmu, na zwiększeniu różnorodności ramek pośrednich (operuje się tu pojęciem plastra – z angielskiego slice), a także na zastosowaniu bloków o dynamicznie zmienianych rozmiarach (od 4x4 do 16x16) i różnego zakresu przeszukiwania wektorów ruchu [9]. Uzyskuje się tym samym lepsze wykorzystanie możliwości kompresji w zależności od konkretnej treści obrazu. Działając z osobna, każde ze wspomnianych usprawnień poprawia skuteczność kompresji o niewielki procent, lecz wszystkie razem połączone w jeden system zapewniają zauważalną poprawę [10]. Niestety nawet tak dobry standard posiada pewne wady. Pojawia się problem, gdy w kadrze zaczyna pojawiać się nadmierny ruch. W przypadku H.264, im więcej rzeczy dzieje się w kadrze, tym trudniej o wydajną kompresję [11], np. gdy interpretowany będzie korytarz szkolny, gdzie podczas przerwy w zasadzie nic się nie dzieje, lecz podczas alarmu pożarowego zmienność sceny będzie bliska 100%, co dla H.264 stanowi spore wyzwanie. Chociaż format ten wymaga większej mocy obliczeniowej niż jego poprzednicy, to wydajność urządzeń jest na tyle duża, że prognozuje się całkowite zastąpienie algorytmów MPEG przez format H.264.

Firma Indigo Vision, która jest wiodącym ogólnoświatowym producentem systemów bezpieczeństwa (m.in. dla lotnisk, portów, kolei, komunikacji, miast,

handlu, bankowości, górnictwa, edukacji, kasyn, policji, więziennictwa i administracji) przeprowadziła badania mające na celu zilustrowanie oszczędności pasma, które można uzyskać stosując standard H.264. Badania polegały na zakodowaniu tej samej sekwencji wideo przy użyciu takich standardów kompresji jak: IndigoVision 8000 MPEG–4, IndigoVision 9000 H.264, MPEG–4 bez przewidywania ruchu oraz M–JPEG. Wszystkie kompresowały obraz z szybkością 25 fps (z wyjątkiem M–JPEG – 5 fps) przy założeniu uzyskania jednakowej jakości obrazu.



Rys. 3.4. Porównanie standardów kompresji [12]

W porównaniu z MPEG–4 standard H.264 może osiągnąć oszczędności pasma od 20% do 25%, a w okresie małej aktywności w kadrze (brak ruchu) nawet o 50 %. Obniżenie bitrate oznacza, że przesyłanych może być więcej danych, co zwiększa szybkość transmisji. Jest to idealne rozwiązanie dla aplikacji zabezpieczających, które potrzebują dużej częstotliwości odświeżania obrazu, jak np. kasyna. Dzięki H.264 nie tylko zmniejszają się wymagania związane z przepustowością pasma, ale również z ilością wymaganego miejsca na nagrania wideo, co często stanowi najdroższy element systemu.

Podsumowując opisane standardy nasuwa się na myśl pytanie: Czy jeden standard kompresji może spełniać wszystkie wymagania? Na takie pytanie oczywiście trudno jednoznacznie odpowiedzieć, gdyż jak wcześniej wspomniano – każdy system monitoringu należy projektować pod własne oczekiwania. Jednak stosownym wydaje się korzystanie z dwóch standardów kompresji na raz, co oczywiście umożliwiają kamery IP. Niezaprzeczalnym wydaje się wybór kodowania

MPEG–4 wersji 10-tej (czyli h.264) do zapisu obrazu pełnoklatkowego. Dodatkowo należy jednak zastanowić się nad zapisem obrazu o najlepszej jakości z ilością 2–4 klatki na sekundę np. w standardzie JPEG czy MJPEG. Użycie obu standardów pozwoli po pierwsze na dokładne i płynne odtworzenie wybranej sytuacji, a po drugie na identyfikację osób bądź obiektów zarejestrowanych na rejestratorze.

3.3. Pliki dźwiękowe i kompresja dźwięku.

Z punktu widzenia użytkownika ważne są różne wielkości charakteryzujące pliki dźwiękowe, takie jak na przykład głośność i jej wahania, barwa dźwięku, czas trwania, występowanie okresów ciszy, fragmenty, w których dźwięk jest zbyt głośny (przesterowanie), itd. Przykładowo: jeśli plik ma być ścieżką dźwiękową do filmu, to trzeba zadbać o odpowiednią głośność, tak aby stanowił tło dla innych dźwięków lub odwrotnie – aby był słyszalny w pierwszym planie. Analiza plików dźwiękowych może ujawnić niesłyszalne lub słabo słyszalne dźwięki w pliku, na przykład rozmowę którą p pozoru zagłuszają całkowicie szumy i inne dźwięki [13].

Szumem akustycznym nazywamy dźwięk, którego widmo jest płaskie, pozbawione lokalnych maksimów lub minimów, których moc znacznie odbiega od mocy innych częstotliwości. Przykładem szumów są naturalne dźwięki występujące w przyrodzie (np. szum drzew). Pojedyncze dźwięki jeżeli występują jednocześnie sprawiają często wrażenie szumu. Istnieją liczne filtry eliminujące lub w większości redukujące szumy dla ścieżki dźwiękowej. Są powszechnie wykorzystywane w profesjonalnych narzędziach pozwalających na obróbkę i montaż dźwięku. Szumy w ścieżce dźwiękowej są niekorzystne dla kompresji [14]. Wiele bitów zostaje utraconych co prowadzi do pogorszenia jakości nagrania. Istnieją narzędzia do kompresji zawierające filtry usuwające np. przydźwięk sieciowy [15].

Przedstawienie sygnału w dziedzinie częstotliwości lub pulsacji, otrzymane przy pomocy transformaty Fouriera [16] nazywa się widmem częstotliwościowym sygnału. Widmem sygnału nazywa się zarówno samą transformatę Fouriera (wynik transformacji Fouriera), jak i wykres przedstawiający tę transformatę [17].

Metodą przechowywania nieskompresowanego dźwięku w postaci cyfrowej jest modulacja kodowo-impulsowa PCM (ang.Pulse Code Modulation) [18]. Modulacja zawarta w każdej próbce stanowi migawkę fali dźwiękowej w określonym punkcie czasu. Sposobem zapisu danych dźwiękowych jest PCM Waveform firmy Microsoft, którego główną zaletą jest możliwość jednoczesnej pracy z większą liczbą strumieni danych, co jest szczególne przydatne przy miksowaniu nagrań wielośladowych. Jest to również format bezstratny, a więc jakościowo najbliższy źródłu dźwięku [19]. Wadą powyższej metody jest

niewątpliwie objętość plików ze względu na to, że w pliku zapisywana jest informacja o amplitudzie wszystkich próbek niezbędnych do odtworzenia dźwięku. Często spotykanym formatem plików audio (szczególnie na komputerach z systemem Windows) jest format WAVE. Wykorzystuje on jako metodę kodowania właśnie PCM.

Wraz z początkiem gwałtownego rozwoju technologii cyfrowych, rozpoczęto intensywne badania nad metodami kompresji danych. Głównym zadaniem kompresji jest zapewnienie odpowiedniej wydajności kompresji, zapewnienie odpowiedniej jakości i wielkości pliku. Każda minuta nieskompresowanego pliku w formacie PCM zajmuje 10 MB. Zasadne jest zatem szukanie wydajnej kompresji przy zachowaniu wiernej jakości dźwięku. Proces dekompresji zawsze musi odbywać się w czasie rzeczywistym.

Przykładem formatu kompresji dźwięku stworzonego przez firmę RealNetworks jest ReaAudio. Kodek został opracowany głównie z myślą o wykorzystaniu go w strumieniowaniu dźwięku przy łączu internetowym o niskiej przepustowości. Wiele internetowych stacji radiowych korzysta z RealAudio przy transmitowaniu audycji przez Internet. Główną wadą standardu jest złe przenoszenie niskich tonów, oraz brak szerokich możliwości wyboru oprogramowania obsługującego ten format.

Standard AAC jest stosowany jako format zapisu dźwięku w wielu różnych mediach - telewizji cyfrowej (formaty MPEG, MPEG2 MPEG4), plikach audio sklepu internetowego ITunes (współpracujących z odtwarzaczem IPod), w telefonach komórkowych z opcją odtwarzania dźwięku (np. część modeli firmy Nokia) i innych. Wiosną 2004 r. organizacja DVD forum (http://www.dvdforum.org) zaleciła format AAC jako uzupełniający dla formatu DVD Audio (ma być stosowany podczas odtwarzania płyt za pomocą komputerów PC). Rozszerzenie format AAC o nazwie AAC plus zapewnia jakość na poziomie CD audio już przy 48 kbps, a przy 32 kbps jakość dźwięku jest wciąż bardzo dobra. 128 kbps pozwala na przesłanie dźwięku wielokanałowego w standardzie 5.1. Ogólnie format AAC określa się jako następcę mp3. Format AAC umożliwia zabezpieczanie plików przed nieuprawnionym kopiowaniem.

Format mp3PRO stanowi ulepszenie starego mp3 polegające głównie na zwiększeniu wydajności kodowania dla małych plików. Zapewnia też ogólnie lepsze upakowanie danych i jakość dźwięku osiąganą przy tym samym strumieniu bitów. Nie udostępnia zabezpieczeń plików przed nieuprawnionym kopiowaniem. Ze względu na sztuczne generowanie brakujących wysokich częstotliwości, utwory z gatunku np. rock, metal, jazz brzmią metalicznie i nienaturalnie.

Standard Windows Media Audio - WMA w najnowszych wersjach kompresuje dane znacznie lepiej niż mp3. Jest to jak na razie jedyny z popularnych formatów

z kompresją, który posiada opcję zapisu dźwięku próbkowanego z częstotliwością 96 kHz / 24-bit (w wersji WMA PRO). Ma on także możliwość zapisu dźwięku wielokanałowego. Ogólniejszy standard "Windows Media" pozwala na kompresję zarówno audio, jak i video, w tym zapis video wysokiej rozdzielczości (HDTV). Posiada jako opcję bezstratny format zapisu dźwięku. Standard WMA został zatwierdzony jako oficjalny kodek HD DVD - następcy formatu DVD. Umożliwia on też zabezpieczanie plików przed nieuprawnionym kopiowaniem.

Format Ogg Vobis charakteryzuje się bardzo dobrą, dobrą, lub niską jakością (zależną od stopnia kompresji). Jest to format kompresji stratnej, który wykorzystują dowolne nośniki cyfrowe (podobne jak mp3). Wysoka jakość dźwięku zapisanego (pod warunkiem, że stopień kompresji nie jest zbyt silny) jest niewątpliwą zaletą standardu ogg. Format jest obsługiwany przez część odtwarzaczy dźwięku z pamięci flash.

Format MP3 został stworzony w roku 1991 w Niemczech w Instytucie Fraunhofera, instytucji zajmującej się obróbką dźwięku. Przy tworzeniu jego pierwszej implementacji wykorzystywany był m.in. utwór Suzanne Vegi Tom's Diner w celu dostosowania kompresji do brzmienia ludzkiego głosu. Pliki w tym formacie otrzymują rozszerzenie .mp3. Standard opisuje jedynie format zapisu. Wszystkie urządzenia i programy potrafiące zapisać dźwięk w tym formacie lub odczytać go, są zgodne z tym formatem. Kodek analizuje głośność dźwięków w poszczególnych pasmach i na podstawie dokładnego modelu psychoakustycznego określa, z jaką dokładnością należy zakodować każde z pasm i czy w ogóle jest potrzeba zapisu danego pasma. Dopasowania takie są wykonywane na bieżąco w sposób ciągły, odrębnie dla każdej chwili. W ten sposób znacznie ograniczono ilość bitów potrzebnych do przesłania danych, a szum kwantyzacji, mimo że znacznie większy niż w oryginale, został tak ukształtowany, aby był niesłyszalny. Dodatkowo liczby opisujące dźwięk zapisuje się w postaci zmiennoprzecinkowej, co również zmniejsza ilość danych.

Tak przygotowane dane łączy się, dodając dane sterujące, umożliwiające dekoderowi odkodowanie dźwięku (które z pasm z jaką dokładnością zostały zakodowane) i dalej kompresuje za pomocą kompresji bezstratnej [20].

3.4. Transmisja głosu w sieciach IP

Ze względu na to, że w sieciach pakietowych nie ma możliwości przesyłania dźwięku w postaci analogowej, tak jak odbywa się to w starszych sieciach POTS (ang. Plain Old Telephone Service), dźwięk jest odpowiednio próbkowany oraz kwantyzowany. Wymusza to przyjęcie za podstawę transmisji sygnałów w postaci

cyfrowej, co ma swoje dobre strony. W przypadku telefonii analogowej, można zauważyć duże pogorszenie jakości dźwięku podczas odbywania rozmów na duże odległości. Wiąże się to z faktem, że stosowane wzmacniacze oprócz wzmocnienia samego głosu powodują również wzmocnienie szumów i innych zakłóceń pojawiających się na łączu. Jest to spowodowane ich ograniczoną zdolnością odróżniania zakłóceń od sygnałów mowy.

Najbardziej podstawową formą cyfrowej modulacji, jest modulacja impulsowo-kodowa PCM (ang. Pulse Code Modulation). W modulacji tej sygnał informacyjny jest reprezentowany w postaci ciągu zakodowanych impulsów. Uzyskiwane jest to przez przetworzenie sygnału do postaci dyskretnej zarówno w czasie jak i amplitudzie [21]. Charakteryzuje się ona 8 bitowym próbkowaniem z częstotliwością 8000 Hz. W Japonii oraz Ameryce Północnej jeden z ośmiu bitów pełni funkcję sterującą, w telefonii europejskiej wszystkie osiem bitów to bity danych. Powoduje to powstanie różnic w wartości wymaganej przepływności toru transmisyjnego dla obu modulacji. W Europie wynosi ona 64 kbps a w Ameryce Północnej i Japonii 56 kbps. Modulacja ta jest dość prosta do zastosowania. Ważnymi zaletami cechującymi modulację PCM są:

- odporność na szumy kanałowe i interferencje,
- duża odporność na ilość straconych pakietów w przypadku sieci komputerowych,
- efektywna regeneracja zakodowanych sygnałów w medium transmisyjnym,
- jednolity format transmisji różnych rodzajów sygnałów, co daje nam możliwość integracji modulacji PCM z innymi rodzajami danych cyfrowych, w obrębie wspólnych sieci teleinformatycznych,
- możliwość zapewnienia bezpiecznej komunikacji z użyciem protokołów szyfrujących,

Ciągły rozwój metod przetwarzania i kompresji doprowadził do zmniejszenia wymaganych przepływności w stosunku do tej, jaka wynikała ze stosowania modulacji PCM. Powodem motywującym zmniejszenie wymagań odnośnie przepływności są oszczędności pasma oraz lepsze jego wykorzystanie (większa liczba strumieni w jednym kanale).

Koncepcja przesyłu głosu przez sieci komputerowe jest znana od lat. W zasadzie nie rożni się ona znacznie od przesyłania innych danych. W ogólnym zarysie, koncepcja ta polega na zastosowaniu następującego algorytmu: odpowiednio przekształcony głos do postaci cyfrowej pakuje się w ramki, dołącza adres IP docelowy i wysyła poprzez sieć. Urządzenie odbiorcze identyfikuje dostarczone w ten sposób ramki, ustawia pakiety w odpowiedniej kolejności (jeśli używamy odpowiednich do tego protokołów np. RTP) dekompresuje otrzymane dane

z postaci cyfrowej na analogową oraz przeprowadza inne zabiegi, takie jak wstawianie ciszy w przypadku utraconych pakietów lub też wygładza dźwięk zapewniając jego odpowiednią wyrazistość.

Najczęściej spotykanym terminem opisującym metody przesyłania pakietowego dźwięku jest VoIP (ang. Voice over IP). Wynika to z powszechności stosowania protokołu IP, szczególnie w odniesieniu do sieci Internet. Pojęcie to nie zawsze odwzorowuje prawdziwy stan rzeczy. Dzieje się tak z tego względu na fakt, że sieć Internet wykorzystuje szereg innych protokołów warstw niższych i równoważnych w stosunku do IP. W szkielecie sieci mogą występować takie technologie jak Frame Relay, ATM, MPLS, xDSL czy technologie bezprzewodowe pracujące w standardzie WiFi (ang. Wireless Fidelity) lub pokrewnych. Możemy zatem spotkać się z takimi określeniami jak VoFR (Voice over Frame Relay), VoATM (Voice over ATM), VoDSL (Vocie over DSL) czy też VoX (ang. Voice over Anything). Wszystkie te rozwiązania możemy określić skrótem VoP (ang. Voice over Packet), który oddaje całą koncepcje przesyłu głosu w sieciach IP.

Ze względu na oszczędność pasma i możliwość przesłania większej liczby strumieni przez ten sam kanał, w połowie lat 90-tych został ustandaryzowany przez organizację ITU-T (ang. International Telecommunication Union - Telecommunication Standardization Sector) szereg kodeków zapewniających transmisję głosu w torach transmisyjnych o niskich przepływnościach bitowych.

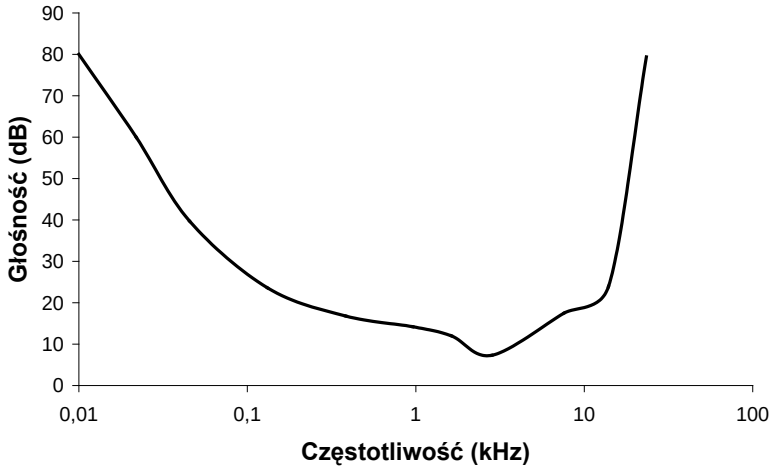
Najpowszechniejszą oraz najstarszą modulacją używaną w telekomunikacji jest PCM. W nomenklaturze organizacji ITU nosi ona symbol G.711.

Kodek G.711 charakteryzuje się przepływnością 64 kbps oraz standardami μ-law i a-law. Metody są do siebie podobne pod tym względem, że wykorzystują od 13 do 14 bitów liniowej jakości PCM w 8 bitach [22]. Użycie kodeków jest uwarunkowane regionalnie. Standard a-law stosowany jest w Europie a μ-law w Ameryce Północnej i Japonii. Zasadniczo μ-law przewyższa odrobinę a-law w kwestii utrzymania stosunku sygnału do zakłóceń na korzyść tego pierwszego.

Inną często stosowaną metodą kompresji głosu jest adaptacyjno-różnicowa modulacja kodem impulsowym ADPCM G.726 (ang. Adaptive Differential Pulse Code Modulation). Kodek ten szyfruje głos wykorzystując próbki 4-bitowe, co daje prędkość 32 kbps. W przeciwieństwie do G.711, nie jest szyfrowana sama amplituda a jedynie różnice w jej poziomach.

G.711 oraz G.726 są to tak zwane kodeki „przebiegu fali", czyli techniki kompresji wykorzystujące powtarzalne właściwości samej fali akustycznej. W ostatnich latach opatentowano szereg kodeków wykorzystujących niedoskonałości ludzkiego słuchu. Najczęściej opierają się one na wykorzystaniu kodowania percepcyjnego. Kodowanie to polega na eliminowaniu częstotliwości,

dla których czułość ucha ludzkiego jest najmniejsza. Służy do tego tzw. krzywa progu słyszalności (rys. 3.5). Wyznacza ona dla danej częstotliwości najmniejszą głośność, przy której ton o danej częstotliwości zaczyna być słyszalny. Dla każdego człowieka krzywa taka ma inną postać, jednakże na potrzeby kompresji stosuje się wartości uśrednione.



Rys. 3.5. Przebieg progu słyszalności w funkcji częstotliwości

Techniki te zgrupowano jako kodeki źródłowe, do których należą takie odmiany jak: liniowe kodowanie przewidujące LPC (ang. Linear Predictive Coding), pobudzona kodem kompresja liniowej predykcji CELP (ang. Code Excited Linear Prediction Compression), kwantyzacja wieloimpulsowa i wielopoziomowa MP-MLQ (ang. Multipulse, Multilevel Quantization) czy też CS-ACELP (ang. Conjugate-Structure Algebraic Code Excited Linear Prediction). Wszystkie te algorytmy kodowania głosu są ustandaryzowane przez organizacje ITU-T w rekomendacjach serii G. Najpopularniejsze z nich wraz z omówionymi wcześniej kodekami G711 oraz G.726, są zestawione w tabeli 3.1.

Tabela3.1. Porównanie różnych kodeków sygnałów mowy

Standard	Algorytm	Przepływność [kb/s]	Wymagana moc obliczeniowa [MIPS]
G.711	PCM	64	0,34
G.726	ADPCM	32	14
G.728	LD-CELP	16	33
G.729	CS-ACELP	8	20
G.729a	CS-ACELP	8	10,5
G.723.1	MP-MLQ	6,3	16
G.723.1	ACELP	5,3	16

Jakość transmitowanego głosu jest subiektywnym odczuciem dla słuchacza. Cyfryzacja głosu uśrednia ciągły sygnał analogowy. Efektem tych uśrednień jest deformacja sygnałów odbieranych na drugim końcu kanału. Każdy kodek zapewnia pewien poziom jakości. Oprócz samego szumu kwantyzacji na jakość głosu wpływa szereg innych czynników, takich jak opóźnienie, echo, ilość straconych pakietów.

Kiedy mówimy o jakości rozmowy, możemy wyróżnić trzy podstawowe kategorie:

- jakość odbioru – odnosi się do tego jak użytkownicy oceniają to co słyszą podczas rozmowy,
- jakość konwersacji – ocena jakości rozmowy przez użytkowników bazująca na jakości odebranej rozmowy i zdolności do konwersacji z drugą osobą podczas połączenia (wliczane są w to takie zakłócenia jak echo, opóźnienie itp.),
- jakość transmisji – określa jakość połączenia sieciowego używanego do przenoszenia głosu,

Celem pomiaru jakości przeprowadzanej rozmowy jest uzyskanie rzetelnej oceny jednej lub więcej wymienionych kategorii, przy użyciu subiektywnych lub obiektywnych metod testowania. Mamy do czynienia z różnorodnością form mierzenia jakości głosu, które zostały opracowane przez szereg lat. Do najbardziej intuicyjnych należą metody polegające na subiektywnej ocenie głosu. Polegają one najczęściej na ocenie przez grupę osób kodeków w idealnych warunkach i nadaniu im pewnej oceny. Testy subiektywne są kosztowne i zabierają dużo czasu. Jednym z takich testów jest ACR (ang. Absolute Category Rating). W teście tym pula słuchaczy ocenia pliki dźwiękowe za pomocą pięciostopniowej skali. Ocena równa 5 oznacza najlepszą jakość głosu natomiast 1 - najgorszą. Po uzyskaniu ocen dla poszczególnych próbek dźwiękowych obliczana jest średnia arytmetyczna. W ten sposób uzyskujemy skale MOS (ang. Mean Opinion Score). Test ten jest bardziej

wiarygodny im większa liczba słuchaczy oceni poszczególną próbkę. Skala MOS jest ustandaryzowana przez organizację ITU, odpowiednio rekomendacja ma numer P.800. Oceny skali MOS przedstawione są w tabeli 3.2.

Tabela3.2. Skala ocen MOS

Wynik	Jakość	Opis spadku jakości
5	Wyśmienita	Niedostrzegalny
4	Bardzo dobra	Dostrzegalny ale nie irytujący
3	Dobra	Dostrzegalny i nieznacznie irytujący
2	Słaba	Denerwujący ale do przyjęcia
1	Zła	Nie do przyjęcia

Kategoria jakości konwersacji jest bardziej skomplikowana i w związku z tym rzadko przeprowadzana. W teście konwersacji pula słuchaczy przeprowadza interaktywne rozmowy z innymi słuchaczami w przeciwieństwie do testu odbioru MOS-LQ (ang. Mean Opinion Score – Listening Quality). Dodatkowo są wprowadzane takie zakłócenia jak opóźnienie, echo, odrzucanie pakietów.

W celu zmniejszenia kosztów oraz poprawienia szybkości określania jakości głosu, zostały opracowane metody nie bazujące na subiektywnych ocenach słuchaczy. Organizacja ITU opracowała odpowiednio rekomendacje P.861 oraz nowszą P.862 bazujące na algorytmie PESQ (ang. Perceptual Evaluation of Speech Quality). Powyższe techniki pomiarów określają zniekształcenia wprowadzone przez system transmisyjny bądź też sam kodek stosowany w transmisji. Metody te polegają na przesyłaniu przez sieć znanej próbki dźwiękowej, a następnie porównywaniu otrzymanej próbki przepuszczonej przez kanał transmisyjny z oryginalną, nie posiadająca uszkodzeń. Wyniki obliczeń algorytmu P.862 mają podobny zakres jak skala MOS. Nie jest to jednak dokładne takie samo odwzorowanie. Obecnie istnieją rekomendacje, które obliczają skalę MOS jedynie na podstawie otrzymanej ścieżki audio. Metodę tą możemy określić mianem pasywnej, w przeciwieństwie do aktywnej, takiej jak P.862, która wymaga przesłania określonego pliku dźwiękowego i porównania go z oryginalnym. Do metod pasywnych należy rekomendacja ITU P.563. Wadą tej metody jest znaczne wykorzystanie zasobów procesora i pamięci urządzenia które odbiera dane głosowe. Spowodowane jest to przetwarzaniem danych w czasie rzeczywistym (obliczanie skali MOS na podstawie aktualnie otrzymywanych próbek).

Ostatnią omawianą metodą obliczania jakości głosu jest tak zwany model E. Jak większość metod pomiaru jakości głosu, model E również posiada swoją

rekomendację organizacji ITU – G.107. Celem powstania tego modelu było subiektywne przewidywanie uszkodzeń próbek głosowych, przy użyciu wcześniej zachowanych informacji na temat poszczególnych efektów powodujących pogorszenie się jakości głosu. Skala jakości dźwięku jest opisana przez czynnik R. Jego przedział zawiera się od 0 do 100 przy czym 100 określa najlepszą jakość dźwięku. W praktyce nie da się uzyskać lepszego wyniku niż 94 w przypadku telefonii wąskopasmowej. Jakość głosu o wartościach poniżej 50 jest nie do zaakceptowania. Czynnik R może być oszacowany w przypadku używania kategorii jakości odbioru czy też jakości konwersacji (odpowiednio MOS-LQ oraz MOS-CQ). Obliczanie czynnika R w modelu E jest oparte na przekonaniu, że efekty wpływające na pogorszenie się transmisji mogą być sumowane. Prezentuje to poniższe równanie:

$$\mathbf{R = Ro - Is - Id - Ie + A}$$

Poszczególne współczynniki oznaczają:

- Ro – opisuje stosunek sygnału do szumu
  - No – szum z różnych źródeł, wyrażany w dBm
- Is – ciągły współczynnik pogorszenia jakości
  - Iolr – niska głośność wyjściowa
  - Iq – szum kwantyzacji
  - Ist – Nieoptymalny sidetone
- Id – współczynnik opóźnienie punkt-punkt
  - Id,te – echo osoby mówiącej
  - Id,le – echo osoby słuchającej
  - Id,d – całkowite opóźnienie
- Ie – pogorszenie głosu wprowadzone przez sprzęt (np. stratny kodek)
- A – współczynnik oczekiwań przez użytkownika (oczekiwania użytkownika co do możliwości użytkowania telefonu w różnych miejscach budynku – mobilność telefonu, obszarów geograficznych lub pojazdów)

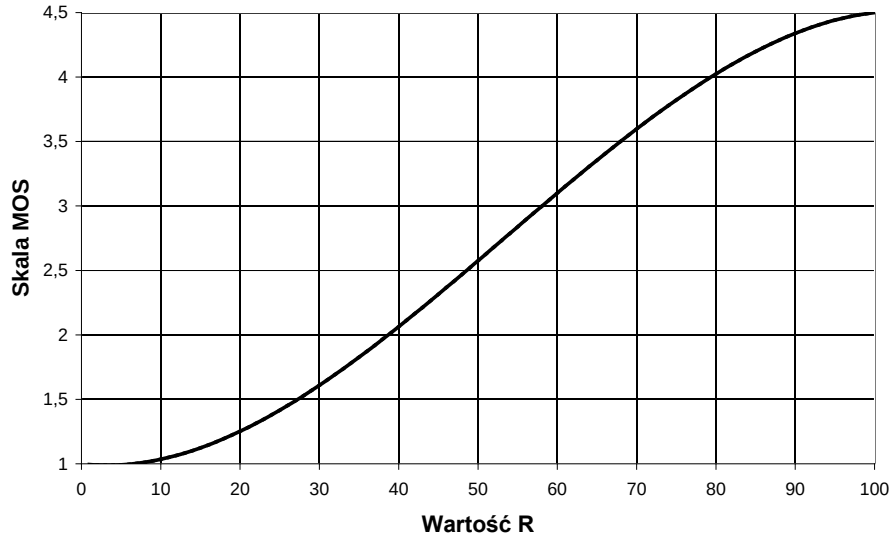
Czynnik Ie (ang. Equipment Impairment) w podanym wzorze odzwierciedla efekt stosowania technologii VoIP. Można to zobrazować za pomocą kodeka G.729, który domyślnie wprowadza pewne pogorszenie jakości gdyż jest kodekiem stratnym. Stąd wartość Ie równa 11. Stosunek sygnału do szumu w przeciętnej sieci wynosi Ro=94. W idealnych warunkach (bez opóźnień echa itp.) czynnik R byłby równy:

$$\mathbf{R = Ro – Ie = 94 – 11 = 83}$$

Uzyskany czynnik R możemy aproksymować do skali MOS za pomocą wzoru:

$$\mathbf{MOS = 1 + 0.035R + 7x10^{-6}R(R-60)(100-R)}$$

Powyższa metoda estymacji jakości dźwięku jest korzystna, gdyż ułatwia zastosowanie nowych czynników określających jakość głosu. Na wykresie (rys 3.6.) można porównać skale MOS i modelu E.



Rys. 3.6. Wzajemna relacja skali MOS i czynnika R

Wyniki poszczególnych testów subiektywnych oraz komputerowych opartych na percepcyjnych modelach ludzkiego słuchu nie zawsze odpowiadają tym samym wartościom skali od 1 do 5. Komputery nie słyszą tego co ucho ludzkie, które można oszukać tak, aby odbierało głos sądząc że ma on wysoką jakość. Testy percepcyjne obliczają najczęściej jakość kompresji dokonywanej przez sam kodek, nie uwzględniając przy tym opóźnienia echa i innych zakłóceń mających ujemny wpływ na jakość połączenia. W tabeli 3.3 zestawione są subiektywna ocena MOS i oceny metod percepcyjnego pomiaru jakości mowy.

Tabela3.3. Porównanie wyników w zależności od rodzaju testu

Kodek	MOS	PESQ(P.862)	Wartość R
G.711	4,1	4,1	83
G.728	3,61	3,8	70
G.729	3,92	3,6	78
G.723.1	3,9	3,5	77

Uzyskiwana jakość głosu w transmisji VoIP jest wypadkową szeregu czynników. Na jakość transmisji składają się takie czynniki jak opóźnienie, zmienność opóźnień (ang. jitter), parametr odrzucenia pakietów (ang. packet loss) czy też dostępna przepustowość. Ważnymi elementami są też czynniki, które nie są bezpośrednio powiązane z samym charakterem sieci komputerowych. Możemy zaliczyć do nich takie elementy jak echo, pogorszenie jakości głosu ze względu na zastosowanie stratnego kodeka (szumy kwantyzacji itp.). W sieciach pakietowych możemy również zauważyć niektóre niepożądane czynniki powodujące pogorszenia głosu znane z sieci PSTN. Należą do nich przesłuchy (mające miejsce wyłącznie w przypadku łączenia sieci VoX i PSTN), szumy otoczenia (spowodowane nieprawidłowym wyregulowaniem mikrofonów), obniżenie lub też podwyższenie głośności transmisji.

Opóźnienie w sieciach VoX to ilość czasu, jaka jest potrzebna na wymówienie słów i dotarcie ich do słuchacza. Jest to jeden z głównych czynników określających możliwość przeprowadzenia rozmowy interaktywnej w czasie rzeczywistym. Rozróżniamy dwa typy opóźnień: stałe i zmienne. Stałe opóźnienie może być wprowadzane przez takie urządzenia jak rutery tzw. opóźnienie serializacji (ang. Router urządzenie, którego zadaniem jest wybór ścieżki dla pakietu). Zmienne opóźnienie jest najczęściej wynikiem szeregowania pakietów, lub zbyt małej przepustowości toru transmisyjnego (zbyt mała przepustowość prowadzi do buforowania pakietów i wysyłania ich w nierównomiernych odstępach czasu co prowadzi do zmienności opóźnienia ang. jitter). Przyjmuje się że przy łącznym opóźnieniu jednokierunkowym rzędu 150 ms jest ono prawie niezauważalne. Dokładne określenia co do opóźnień są zawarte w rekomendacji G.114 ITU-T. Opóźnienia przy dobrej jakości samego dźwięku mogą być akceptowalne przez użytkowników. Zależy to od tolerancji osób przeprowadzających rozmowę. W transmisji satelitarnej dotarcie transmisji do satelity trwa około 250 ms, a następnie drugie tyle samo jej powrót na Ziemie. Choć rekomendacja ITU-T stwierdza iż jest to nieakceptowane, każdego dnia odbywa się wiele takich rozmów o nawet większym opóźnieniu.

W dzisiejszych sieciach komputerowych, oprócz podziału na opóźnienia stałe i zmienne, możemy wprowadzić podział na: opóźnienie propagacji, opóźnienie serializacji oraz opóźnienie obsługi. Opóźnienie propagacji jest powodowane przez prędkość światła w sieciach światłowodowych lub przez prędkość elektronów w sieciach miedzianych oraz prędkość fal elektromagnetycznych w powietrzu, w przypadku sieci bezprzewodowych. Opóźnienie obsługi zwane także opóźnieniem przetwarzania definiuje wiele różnych przyczyn opóźnienia. Należą do nich

opóźnienia pakietowania, kompresji, szeregowania czy też serializacja pakietów. Wprowadzane są przez urządzenia przekazujące datagramy w sieci.

Opóźnienie kompresji jest czasem potrzebnym dla procesora sygnałowego DSP (ang. Digital Signall Procesing) do skompresowania bloku próbek PCM. Opóźnienie te zależy od użytego kodeka oraz szybkości przetwarzania procesora (to drugie w dzisiejszych czasach może zostać pominięte). Czas zdekodowania jest równy około jednej dziesiątej czasu potrzebnego na zakodowanie próbek. W tabeli3.4 zestawiono czasy najkrótszego i najdłuższego opóźnienia kompresji.

Tabela3.4. Opóźnienia kompresji

Kodek	Przepływność [kbps]	Opóźnienie kompresji [ms]
G.711 PCM	64	0.75
G.726 ADPCM	32	1
G.728 LD-CELP	16	Od 3 do 5
G.729 CS-ACELP	8	10
G.723.1 MP-MLQ	6.3	30
G.723.1 ACELP	5.3	30

Ze względu na fakt przesyłania danych w pakietach istnieje pojęcie tzw. opóźnienia pakietyzacji. Określa ono czas potrzebny do zapełnienia skompresowanym głosem pola danych w pakiecie IP. Kodek G.729 domyślnie tworzy ramki dźwięku o długości 10 ms czyli 10 bajtów.

Tabela3.5. Opóźnienie pakietyzacji w zależności od rozmiaru ładunku

Kodek	**Przepływność**	**Rozmiar ładunku (domyślny) [byte]**	**Opóźnienie pakietyzacji [ms]**	**Rozmiar ładunku [byte]**	**Opóźnienie pakietyzacji [ms]**
PCM, G.711	64 Kbps	160	20	240	30
ADPCM, G.726	32 Kbps	80	20	120	30
CS-ACELP, G.729	8.0 Kbps	20	20	30	30
MP-MLQ, G.723.1	6.3 Kbps	24	24	60	48
MP-ACELP, G.723.1	5.3 Kbps	20	30	60	60

Ze względu, że takie rozwiązanie byłoby marnowaniem przepustowości (sam nagłówek IP ma długość 20 bajtów plus 8 bajtów nagłówka UDP i 12 bajtów nagłówka RTP), tworzone są próbki o długości co najmniej 20 ms. Opóźnienie to jest dosłownie rozmiarem bloku próbek, które są umieszczane w jednej ramce. Opóźnienie pakietyzacji jest często nazywane opóźnieniem akumulacji, ponieważ próbki z głosem są buforowane dopóki nie osiągną odpowiedniego rozmiaru, który można przesłać. W tabeli 3.5 zestawiono czasy opóźnień w zależności od rozmiarów pakietów.

Opóźnienie serializacji jest to stała ilość czasu potrzebna na rzeczywiste przesłanie pakietu przez łącze. Opóźnienie odgrywa dużą rolę na łączach o małej przepustowości. Im większy pakiet tym dłuższy czas jego przesyłu. Serializacji używa się najczęściej do zmniejszania rozmiaru pakietów z innymi danymi niż głos.

Tabela3.6. Opóźnienie serializacji

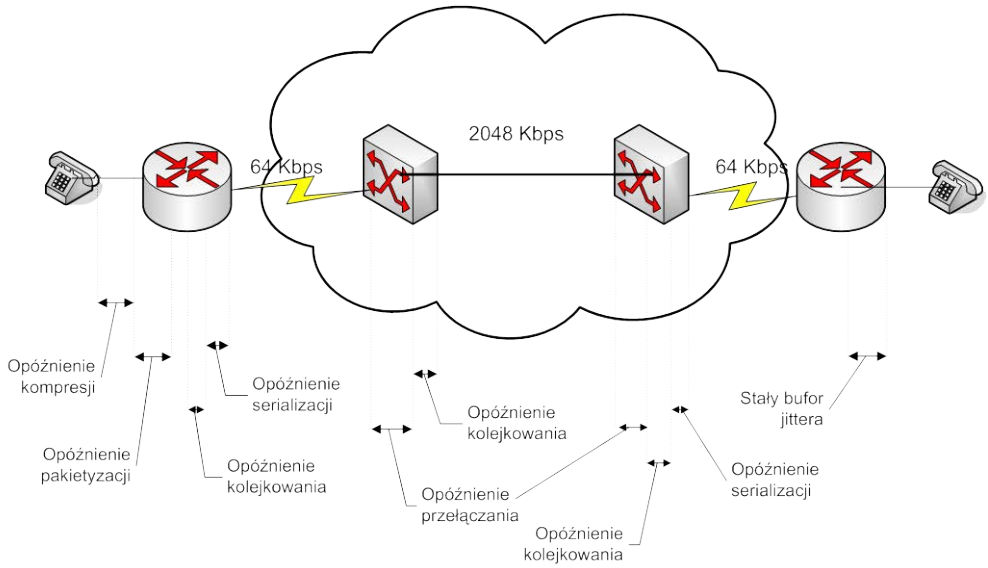
Rozmiar ramki [Byte]	**Przepustowość łącza [Kbps]**										
	19.2	**56**	**64**	**128**	**256**	**384**	**512**	**768**	**1024**	**1544**	**2048**
38	15.83	5.43	4.75	2.38	1.19	0.79	0.59	0.40	0.30	0.20	0.15
48	20.00	6.86	6.00	3.00	1.50	1.00	0.75	0.50	0.38	0.25	0.19
64	26.67	9.14	8.00	4.00	2.00	1.33	1.00	0.67	0.50	0.33	0.25
128	53.33	18.29	16.00	8.00	4.00	2.67	2.00	1.33	1.00	0.66	0.50
256	106.6	36.57	32.00	16.00	8.00	5.33	4.00	2.67	2.00	1.33	1.00
512	213.3	73.14	64.00	32.00	16.00	10.67	8.00	5.33	4.00	2.65	2.00
1024	426.6	149.2	128.0	64.00	32.00	21.33	16.00	10.67	8.00	5.31	4.00
1500	625.0	214.2	187.5	93.75	46.88	31.25	23.44	15.63	11.72	7.77	5.86
2048	853.3	292.5	256.0	128.0	64.00	42.67	32.00	21.33	16.00	10.61	8.00

Praktykę tą stosuje się ponieważ duże pakiety z danymi mogą całkowicie uniemożliwić przesłanie wrażliwych na opóźnienie datagramów z głosem. Na łączach o dużej przepustowości (powyżej 2 Mbps) z reguły czasy te są bardzo małe

i opóźnienie może być pominięte. Charakterystyka opóźnienia serializacji jest przedstawiona w tabeli 3.6, która zawiera opóźnienia dla całego rozmiaru ramki - nie tylko jej ładunku.

Rezultatem przechowywania pakietów w kolejce ze względu na zator na interfejsie zewnętrznym jest opóźnienie szeregowania. Opóźnienie szeregowania ma miejsce wtedy, gdy w danym przedziale czasowym zostaje wysłanych więcej pakietów niż interfejs jest w stanie obsłużyć. Opóźnienie szeregowania ma zmienny charakter. Opóźnienie zależy od stanu kolejki i przepustowości łącza. W źle zarządzanej sieci z zatorami, opóźnienie szeregowania może powodować dodanie nawet dwóch sekund opóźnienia (albo odrzucenia pakietów w zależności od rodzaju kolejki).Opóźnienie przełączania pakietów polega na upływie czasu jaki potrzebuje ruter aby dokonać decyzji o wyborze trasy na podstawie tablicy rutingu dla danego pakietu. Opóźnienie przełączania zależy głównie od procesora, rozmiaru tablicy rutingu oraz obciążenia procesora wybierającego trasy.

Oprócz powyższych opóźnień rozróżniane jest opóźnienie bufora niwelującego zjawisko jittera. Bufor taki jest potrzebny, gdy pakiety przychodzą w różnych odstępach czasu. Aby wyeliminować przerwy w odtwarzanym dźwięku, opróżniany jest bufor, w którym przechowywana jest pewna ilość próbek. Zazwyczaj pojemność bufora nie przekracza 40 ms. Podstawowe opóźnienia na torze przesyłu mowy w sieci pakietowej przedstawione są na rysunku 3.7.



Rys. 3.7. Źródła opóźnień w sieciach pakietowych

Utrata pakietów w sieciach pakietowych jest powszechna i oczekiwana. Należy jednak pamiętać, że sieci VoIP są szczególnie wrażliwe na utratę pakietów.

Im bardziej stratny kodek jest zastosowany w rozmowie głosowej, tym większe pogorszenie i dyskomfort z przeprowadzanej rozmowy jest odczuwany przez słuchaczy w wyniku nawet małej ilości zgubionych pakietów. Wiele protokołów transmisji danych używa utraty pakietów jako swoistego wyznacznika kondycji sieci, i potrafi w razie konieczności redukować liczbę wysłanych pakietów. Straty pakietów mogą być spowodowane przez szereg czynników, takich jak: przeciążenie łącza, „wąskie gardło", kolizje w sieciach LAN, błędy spowodowane przez niepoprawnie działające urządzenia w warstwie fizycznej i wiele innych. Pakiety głosowe, tak samo jak normalne pakiety z danymi, również ulegają utracie. Utrata pakietów jest szczególnie widoczna odkąd protokół czasu rzeczywistego RTP jest używany przez protokół transportowy UDP (ang. User Datagram Protocol), nie zapewniający potwierdzenia przesłanych pakietów.

Istnieje szereg reakcji na periodyczną utratę pakietów. Strategia ukrycia polega na odtworzeniu ostatniego odebranego pakietu. Jeśli pakiet głosowy nie zostanie odebrany wtedy, gdy jest oczekiwany, zakłada się że został utracony i odtworzony zostaje ostatni odebrany pakiet. Ponieważ utracony pakiet zawiera domyślnie 20 ms mowy, przeciętny słuchacz nie zauważy różnicy w jakości głosu. Strategię ukrycia można zastosować tylko wtedy, gdy utracono jeden pakiet. Jeśli zginęło kilka kolejnych pakietów, strategia ukrycia jest stosowana tylko raz. W przypadku dalszego tracenia pakietów, kolejne miejsca są wypełniane ciszą. Metoda ta jest najwcześniej stosowana w przypadku kodeków stratnych [23].

Inną metodą na eliminacji strat pakietów, jest stosowanie przeplotu (ang. Interleaving). Metoda ta polega na interpolowaniu próbek zawartych w sąsiednich datagramach. Protokół RTP może tak zmienić kolejność liczb w próbkach, że najpierw przesłane zostaną ich nieparzyste połowy, a następnie w kolejnym datagramie parzyste. Dzięki temu jeśli pakiet zostanie zgubiony, to poprzedzający go pakiet dysponuje połową informacji. W ten sposób utracona informacja może zostać interpolowana na podstawie datagramu poprzedzającego. Technika ta jest stosowana w przypadku strumieni nieskompresowanych [24].

Oprócz powyższych technik stosuje się zastępowanie utraconych pakietów ciszą. Nie jest to najlepsze rozwiązanie, gdyż w przypadku utracenia większej ilości danych, pogorszenie jakości głosu jest łatwo zauważalne przez użytkownika. Utracone datagramy z głosem można zastępować szumem białym. Amplituda takiego szumu jest równa amplitudzie sygnału z ostatniego odebranego pakietu.

Utrata pakietów zaczyna być poważnym problemem gdy przekracza około 5%. Również ważnym problemem jest okresowość strat. W przypadku pewnej periodyczności strat, mogą one zostać skompensowane. Gdy mamy do czynienia

z dużym blokiem traconych pakietów, nawet najlepsze metody nie mogą zrobić nic poza zastąpieniem zgubionych pakietów ciszą.

Kolejnym elementem jest echo, które podczas rozmowy telefonicznej może przyjmować formę od czynnika lekko denerwującego do tak nieznośnego że odbycie rozmowy jest niemożliwe. Zjawisko nazywane jest „przeciekiem" naszego głosu do ścieżki zwrotnej, którą słyszymy w słuchawce. Odbieranie własnego głosu w sieciach analogowych jest zjawiskiem powszechnym. Nie zdajemy sobie z niego sprawy gdyż opóźnienia są zbyt małe aby ucho ludzkie mogło je wychwycić. Transkontynentalna rozmowa z sieci PSTN do Stanów Zjednoczonych ma zazwyczaj opóźnienie rzędu 10, 20 ms. Analogowa transmisja głosu jest szybka, ograniczona tylko przez szybkość propagacji elektronów lub fotonów w torze transmisyjnym. Aby wystąpiło zjawisko echa odczuwalne przez ucho ludzkie, muszą być spełnione dwa warunki. Echo musi być głośne i opóźnione o co najmniej 25 ms. W normalnej sieci PSTN echo jest głośne ale nie jest tak opóźnione jak w sieciach pakietowych, dlatego nie zdajemy sobie sprawy z jego istnienia. W sieciach zjawisko to występuje bardzo rzadko jeśli cała rozmowa jest prowadzona w obrębie tej samej sieci pakietowej. Echo do sieci VoIP jest wprowadzane na stykach z sieciami PSTN.
Echo jest spowodowane przez następujące czynniki:

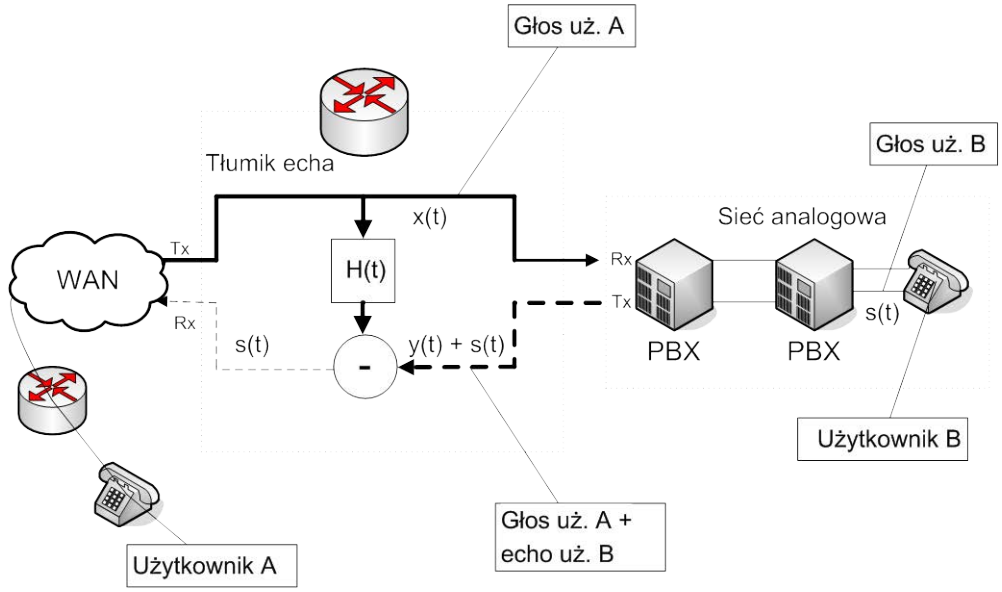
- w tradycyjnych sieciach analogowych przez błąd impedancji na przejściu przełącznika sieci czteroprzewodowej na dwuprzewodową lokalną pętlę (odbicie głosu, innymi czynnikami w sieciach analogowych są :
    - wilgotne lub zniszczone kable w sieci POTS
    - tanie telefony analogowe dołączone do lokalnej pętli
    - rozkręcona skrętka telefoniczna na długim odcinku
- źle zaprojektowana słuchawka, część energii z mikrofonu może zostać powrotem przekazana do słuchawki
- źle wyregulowana głośność po stronie osoby do której dzwonimy, najczęściej jeśli osoba ta używa systemu głośnomówiącego (głos z głośnika jest przekazywany do mikrofonu i z powrotem do osoby dzwoniącej)

W sieciach telefonicznych, w których przesyłany jest głos analogowy, wykorzystują tłumiki echa. Tłumiki te usuwają echo poprzez nałożenie ograniczeń na impedancję w obwodzie. Nie jest to jeden z najbardziej efektywnych mechanizmów usuwania echa. Na linii gdzie znajdują się tłumiki echa nie można używać ISDN ponieważ tłumiki echa zmniejszają zakres częstotliwości używanych przez ISDN.

Aby echo dało się zauważyć muszą być spełnione dwa warunki: opóźnienie oraz odpowiednia głośność powracającego echa. Ponieważ w sieciach pakietowych

istnieje większe opóźnienie niż w sieciach z komutacją łączy, echo występuje przy próbie łączenia obydwu rodzajów sieci.

W sieciach pakietowych można wbudować tłumiki echa w kodeki. Często przeznacza się do tego zadania oddzielny procesor DSP w ruterze. Opis działania tej metody został przedstawiony na Rys 3.8.



Rys. 3.8. Tłumik echa wbudowany w ruter

Użytkownik A, który znajduje się w sieci pakietowej, rozmawia z użytkownikiem B, który jest w sieci analogowej. Sygnał mowy użytkownika A jest oznaczony jako x(t). Kiedy x(t) napotyka na swojej drodze niedopasowanie impedancji lub inne środowisko powodujące echo, wraca do użytkownika A. Aby usunąć echo z linii, głos przechodzący przez ruter podłączony do sieci PSTN przechowuje przez pewien czas odwrócony obraz mowy użytkownika A czyli H(t). Tłumik echa słucha dźwięku płynącego od użytkownika B (czyli s(t) i echo y(t)) i wykorzystuje H(t) do usunięcia echa y(t). Parametrem charakteryzującym tłumik echa jest długość czasu przez jaki oczekuje on na odbitą mowę. Czas ten jest nazywany ogonem echa.

Jakość głosu w sieciach IP nie odbiega, a często znacznie przewyższa znaną z sieci PSTN. Duży wpływ na samą jakość głosu ma użyty kodek. Najlepsza jakość głosu jest osiągana za pomocą kodeka G.711. Jednakże stosowanie są również kodeki G.729. Zapewniają one dobrą jakość głosu oraz co jest ich największą zaletą - zużywają mniej drogocennego pasma w sieciach szkieletowych. Oprócz używanego sposobu kodowania głosu na jego jakość ma wpływ szereg czynników np.

opóźnienie, utrata pakietów, jitter. Są to elementy mające ogromny wpływ na jakość transmisji multimedialnej.

Literatura

1. Hansen, B., The Dictionary of Multimedia, Terms and Acronyms, Fitzroy Dearborn, Chicago, (1998).
2. Kraszewski G., wykład SYSTEMY MULTIMEDIALNE, dostępne na stronie: http://cygnus.et.put.poznan.pl/~tmarcin.
3. Wieczorkowska A., Multimedia. Podstawy teoretyczne i zastosowania praktyczne, PJWSTK (2008).
4. Matysiak S., Sposoby kompresji obrazu w cyfrowych systemach rejestracji i monitoringu, Forsec, 2/2007, s. 16-19.
5. Nilsson F., Intelligent Network Video: Understanding Modern Video Surveillance Systems, Taylor & Francis Group, LLC, London 2009.
6. Przewodnik po sieciowych systemach telewizji dozorowej IP, Axis Communications, 1/2008.
7. DeAngelis P., Bodell P.: Key consideration when selecting a video compression algorithm, październik 2009.
8. Maćkowiak S., Standard MPEG-4 część 10 (AVC), Miesięcznik Twierdza 5(51)/2008, s. 12-15.
9. Grajek T., Karwowski D., Złożoność obliczeniowa i efektywność kodowania entropijnego w standardzie H.264/AVC, Krajowa Konferencja Radiokomunikacji, Radiofonii i Telewizji KKRRiT, Kraków 15-17 czerwca 2005.
10. Walczyk A., Kompresja obrazów ruchomych zgodna ze standardem H.264, Zabezpieczenia.com.pl, lipiec 2008 http://www.zabezpieczenia.com.pl/telewizja-dozorowa/kompresja-obrazow-ruchomych-zgodna-ze-standardem-h-264.
11. Richardson I. E.G., H.264 and MPEG-4 video compression, John Wiley & Sons Ltd, Chichester 2003.
12. http://www.indigovision.com, lipiec 2010.
13. Rudny T., Technik Informatyk. Multimedia i grafika komputerowa, helion 2010.
14. Beach A., Kompresja dźwięku i obrazu wideo, helion 2009.
15. Denes P, Naunton R. F., Masking in Pure-tone Audiometry, Proc R Soc Med. 1952 November; 45(11): 790–794.

16. Bracewell, R. N. The Fourier Transform and Its Applications (3rd ed.), Boston (2000).
17. Zieliński T. P., Cyfrowe przetwarzanie sygnałów: od teorii do zastosowań, Wydawnictwa Komunikacji i Łączności, Wyd. 2 popr, Warszawa 2007.
18. Stallings, William, Digital Signaling Techniques, December 1984, Vol. 22, No. 12, IEEE Communications Magazine.
19. Nasiłowski D., Jakościowe aspektyu kompresji obrazu i dźwięku", Wydawnictwo Mikom, Warszawa 2004.
20. Colomes C., Schmidmer C., Treurniet W. C., Perceptual-quality assessment fordigital audio:Peaq –the proposed itu standard for objective measurement of perceived audio quality, Proceedings of the AES 17th. International Conference, 1999.
21. Haykin S., Systemy telekomunikacyjne, WKŁ Warszawa 1998.
22. Davidson J., Peters J., Voice over IP – Podstawy, MIKOM Warszawa październik 2005.
23. Clark A., Voice Quality Measurement, TMCnet Articles, 28 Feb 2005.
24. Tanenbaum A., Sieci Komputerowe, HELION Gliwice 2004.

4. Multimedia strumieniowe

W ciągu ostatnich dwóch dekad, obserwuje się systematyczne zwiększanie zapotrzebowania na szerokopasmowy dostęp do Internetu, związane z dużą dynamiką rozwoju usług multimedialnych. Ten obserwowalny popyt, rośnie wraz z ilością i wielkością przesyłanych plików. Dawne strony internetowe, praktycznie pozbawione grafiki, stały się z czasem coraz bardziej multimedialne, zawierając animacje, dźwięk i wideo. Ustawiczny rozwój globalnej sieci pozwolił na przesyłanie bardzo dobrej jakości plików muzycznych oraz filmów (jakości telewizyjnej). Przy czym możliwość odtwarzania multimediów obarczona była przymusem zapisu pliku na dysk lokalny, i odczekaniem czasu potrzebnego na jego pobranie, co było szczególnie dotkliwe przy stosunkowo wolnych łączach.

Rozwiązaniem tego problemu stały się transmisje strumieniowe, oparte na idei przesyłania danych od nadawcy do odbiorcy, bez konieczności pobrania materiału na dysk, aby rozpocząć odtwarzanie. Każdy pakiet zawiera ilość informacji, wystarczającą do odtworzenia pewnego fragmentu prezentacji multimedialnej Umożliwiają to technologie kodowania oraz stan wiedzy odnośnie związanych z nimi standardów, będące bardzo ważnym elementem, wszędzie tam, gdzie istotne jest przetwarzanie sygnałów multimedialnych.

Standardy takie jak MPEG-2 [14] czy H.264/AVC [9] dostarczają znaczącego wsparcia dla cyfrowej transmisji wideo, przechowywania tego rodzaju danych oraz rozwoju aplikacji strumieniowych.

Najpoważniejszą wadą pobierania plików, a zarazem największą zaletą strumieniowania jest to, że można za jego pośrednictwem realizować transmisje na żywo. Dzięki strumieniowaniu w czasie rzeczywistym stało się możliwe uruchamianie w sieci różnego rodzaju telewizji i radiostacji internetowych lub przekazów ze zwykłych stacji w Internecie. Stąd, bardzo istotne oraz aktualne zagadnienia stanowią wydajność i elastyczność przepustowości (ang. bandwidth) pomiędzy serwerami wideo a urządzeniami użytkowników końcowych. W odpowiedzi na istniejące i pojawiające się wyzwania w tym zakresie, w pracach [2, 4, 6, 20, 26, 35, 36, 39, 48, 52, 53] zaproponowano szereg technik kodowania oraz różnych typów efektywnych usług strumieniowania wideo. Skalowalne strumieniowanie wideo za pośrednictwem Internetu, zostało wnikliwie przeanalizowane w pracach [4, 20, 21, 35, 48].

4.1. Istota transmisji multimedialnych

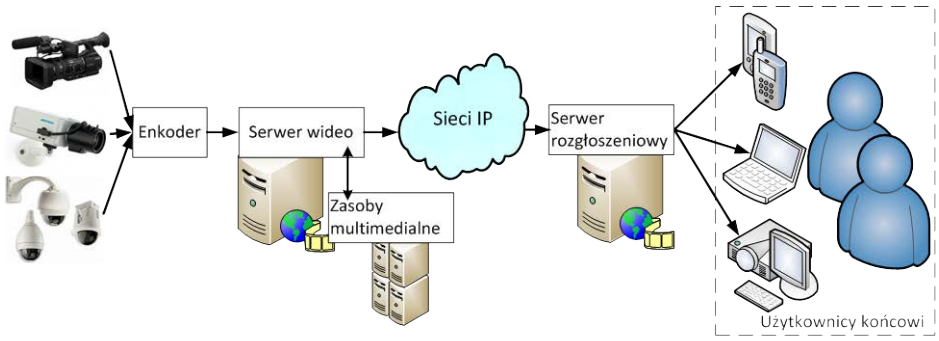
Transmisje strumieniowe (ang. streaming) stanowią sposób przesyłania danych multimedialnych w sieciach pakietowych. Oferują bezpośrednie odtwarzanie (ang. live) materiału audiowizualnego umieszczonego na serwerze w Internecie, lub sieci LAN/WLAN, na urządzeniu (terminalu) użytkownika w czasie rzeczywistym lub „na życzenie" (ang. on demand). Transmitowane dane, dzięki specjalnym serwerom strumieniowym, odtwarzane są przez oprogramowanie klienckie w miarę ich dostarczania, nie pozostawiając swojej kopii na dysku.

W ramach technologii strumieniowania, istnieją dwa podejścia, związane z: przełączaniem pomiędzy wieloma wstępnie zakodowanymi nieskalowalnymi strumieniami bitowymi (ang. switching among multiple pre-encoded non-scalable bitstreams) oraz strumieniowaniem z pojedynczym skalowalnym strumieniem bitowym (ang. streaming with a single scalable bitstream). Praca [1, 2] przedstawia zwięzły przegląd aplikacji strumieniowania wideo i komunikacyjnych.

Różne klasy aplikacji wideo dostarczają odmienne ograniczenia oraz poziomy swobody. Warto zaznaczyć, że trzy podstawowe zagadnienia, które muszą być rozwiązane bezpośrednio w ramach strumieniowania obejmują: nieznaną i zmieniająca się w czasie przepustowość łącza (ang. bandwidth), opóźnienie (ang. delay) fluktuację pakietów (ang. jitter), oraz utratę danych.

W celu dostarczenia pewnych rozwiązań wspomnianych problemów, zaproponowane zostały metody skalowalnego kodowania oraz dekodowania wideo (ang. scalable video coding and transcoding). Techniki tego typu powinny poradzić sobie z dostosowaniem ilości danych do transmisji, nawet przy (nieoczekiwanych) wahaniach przepustowości. W pracach [6, 26, 35, 52, 53] dokonano analizy zagadnienia alokacji bitów (ang. bit allocation) oraz odporności na błędy (ang. error resilience). Źródła [25, 50] podkreślają, że metody kodowania wideo oraz dystrybucja skalowalnego wideo stanowią dwie kluczowe kwestie dla systemów strumieniowania.

Typowy system strumieniowania wideo, przedstawiony na rysunku 4.1, składa się z enkodera, serwera dystrybucji (ang. distribution server) z zasobami wideo (ang. video storage), serwera rozgłoszeniowego (ang. relay server) oraz użytkowników końcowych, którzy otrzymują dane multimedialne.



Rys. 4.1. Schemat blokowy typowego systemu strumieniowania wideo

Serwer dystrybucyjny (ang. distribution server) przechowuje zakodowane dane wideo oraz rozpoczyna ich dystrybucję na żądanie klienta. Stąd, niezależnie od lokalizacji i czasu (wł. strefy czasowej), użytkownicy mogą pobierać i odtwarzać strumienie wideo, po uprzednim nawiązaniu połączenia z serwerem, za pośrednictwem sieci. Kodowanie i transmisja następuje w czasie rzeczywistym w przypadku dystrybucji na bieżąco (określanej także „na żywo"), bądź „na żądanie" (ang. on-demand).

Na potrzeby kodowania tego rodzaju multimediów dostępne są dwa główne rodzaje kompresji sygnałów: nieskalowalne kodowanie wideo (ang. non-scalable video coding) oraz skalowalne kodowanie wideo (ang. scalable video coding). W pierwszym z nich treść wideo jest kodowana, niezależnie od aktualnej charakterystyki kanału. Przy takim podejściu, najistotniejszym czynnikiem jest efektywność kodowania (ang. coding efficiency), natomiast sama kompresja jest optymalizowana względem wstępnie określonej szybkości (ang. pre-specified rate). Głównym problemem tej metody jest trudność adaptacyjnego strumieniowania treści nieskalowalnego wideo do heterogenicznych terminali klienckich za pośrednictwem zmiennych w czasie kanałów komunikacyjnych. Jest to szczególnie istotne, zwłaszcza dla zastosowań bezprzewodowych. W przypadku drugiego podejścia, kodowanie przeprowadzane jest jednorazowo, a następnie uzyskiwane są niższej jakości rozdzielczości przestrzenne czy czasowe przez odwrócenie określonych warstw lub bitów z pojedynczego strumienia.

W rezultacie, jako nadrzędny cel, powstaje w ten sposób skalowalna reprezentacja sygnału wideo, nie posiadająca wpływu na efektywność kodowania. Dla przykładu okrojony, skalowalny strumień (przy niższej szybkości i rozdzielczości przetrzennej/czasowej), powinien umożliwiać uzyskanie zrekonstruowanej jakości, analogicznej jak dla jednowarstwowego strumienia bitowego, w którym zakodowane zostało wideo przy tych samych warunkach i ograniczeniach.

W szczególności przy tej samej szybkości bitowej. Niemniej jednak, praktycznie każde kodery skalowalnego wideo wnoszą straty w efektywność kompresji względem znanych obecnie metod kodowania nieskalowalnego.

Przy rozpowszechnianiu strumieni multimedialnych, zarówno serwer wideo jak i rozgłoszeniowy są dopowiedziane za dopasowanie danych wyjściowych do dostępnych zasobów kanału oraz ostatecznie do możliwości urządzenia klienta. W przypadku danych wideo serwer może przekodować strumień w celu zredukowania szybkości transmisji, ilości klatek lub rozdzielczości [35, 42, 50].

Alternatywnie może zostać wybrany najbardziej odpowiedni strumień bitowy z wielu, wstępnie zakodowanych i posiadających zróżnicowane jakość i rozdzielczość itp.

Rozważając charakterystykę strat pakietów w sieciach komunikacyjnych, serwery mogą również dołączać mechanizm związany z niwelacją błędów (ang. error resilience) do strumienia bitowego na wyjściu [44].

Ogólnie, optymalna rozdzielczość stanowi tą, która zwraca najlepiej zrekonstruowane wideo po stronie odbiorczej (czyli poza wytycznymi, normami i parametrami, najistotniejsza jest percepcja i ocena użytkownika końcowego - widza). Więcej na temat odporności na błędy (ang. error resilience) oraz niwelacji błędów (ang. error concealment) można znaleźć w pozycjach [35, 44, 45].

Ogólna złożoność systemu, z uwzględnieniem serwerów jak i poszczególnych klientów stanowi odrębną istotną cechę tego typu rozwiązań. Należy zauważyć, że w szczególności techniki strumieniowania obejmują również zagadnienia z zakresu sieci komunikacyjnych, dostarczających treści jak np. Akamai CDN. CDN (ang. content delivery networks/content distribution networks) stanowią systemem komputerów zawierających kopie danych rozmieszczonych w różnych węzłach sieci. Właściwie zaprojektowane i wdrożone CDN znacząco wpływa na parametry technik strumieniowania. Może poprawiać dostęp do danych dzięki buforowaniu i zwiększać szybkość dostępu i ograniczać opóźnienia. Na ogół buforowane w CDN dane obejmują obiekty związane z treściami webowymi -pliki multimedialne, oprogramowanie, dokumenty, aplikacje na żywo mediów strumieniowych, i zapytania do baz danych. Więcej na ten temat dostępne jest w [27, 35, 53]

4.2. Przegląd standardów kodowania

Proces kodowania wideo odgrywa znaczącą rolę w wypełnianiu luki pomiędzy dużymi ilościami danych wizualnych oraz ograniczoną przepustowością sieci w dystrybucji wideo. W ciągu ostatnich dwóch dekad opracowano różne metody kodowania na potrzeby komercyjne. Standardy kodowania wideo zostały rozwinięte przez dwie główne grupy organizacji standaryzacyjnych.

Pierwsza z nich to Moving Pictures Expert Group (MPEG) w ramach ISO/IEC, a druga - Video Compression Expert Group (VCEG) z grupy ITU-T [10, 11]. Standardy kodowania wideo, opracowane przez ISO/IEC obejmują: MPEG-1 [13], MPEG-2 [14], i MPEG-4 [15]. Z kolei standardy opracowane prze przedstawicieli Międzynarodowej Unii Telekomunikacyjnej (ITU) to między innymi: H.261 [16], H.262 [14], H.263 [17] oraz H.264/AVC [9]. Należy podkreślić, że H.262 jest analogiczny z MPEG-2, ponieważ jest on wspólnym standardem grup MPEG i ITU [35].

Standard kodowania H.264/AVC został opracowany przez połączone zespoły MPEG and ITU, tworzące tzw. Joint Video Team (JVT) i stąd równocześnie stanowi on schemat MPEG-4 Part 10. Niniejsze standardy znalazły szereg owocnych zastosowań, takich jak DTV, DVD, cyfrowa telefonia czy w zastosowaniach strumieniowania wideo.

Do grona najpopularniejszych standardów kodowania wideo należy zaliczyć [35]:

- H.261, opracowany w 1990, wykorzystywany głównie usługach konferencyjnych technologii ISDN;

- H.263, opracowany w 1996 oraz oparty na H.261, lecz zawiera dodatkowe algorytmy, w celu polepszenia wydajności kodowania;

- MPEG-1, opracowany w 1991. Zasadniczy cel zastosowań MPEG-1 stanowiły cyfrowe nośniki danych jak CD-ROM, przy szybkościach bitowych do 1.5 Mb/s;

- MPEG-2, niekiedy również znany jako H.262, został opracowany w 1994, stanowi rozszerzenie MPEG-1 i pozwala na większą elastyczność formatu wejściowego oraz wyższe prędkości danych zarówno dla HDTV (ang. *High-definition Television*), jak i SDTV (ang. *Standard Definition Television*). Co więcej, europejski standard DTV, jak i amerykański ATSC DTV wykorzystują MPEG-2 jako źródłowy format kodowania. Ponadto, MPEG-2 stosowany jest w cyforwych nośnikach danych, format DVD (ang. *Digital Video Disk*);

- MPEG-4 Part 2, opracowany w 2000 r., stanowi pierwszy obiektowy standard kodowania i jest przeznaczony dla wysoko interaktywnych zastosowań multimedialnych. Zarówno prosty, jak i zaawansowany prosty profil (ang. Simple Profile and Advanced Simple Profile) MPEG-4 Part 2 są wykorzystywane w zastosowaniach mobilnych oraz strumieniowania;

- H.264 jest znany także jako MPEG-4 Part 10 Advanced Video Coding. Stanowi on standard kodowania, opracowany przez JVT z ISO i ITU. H.264 znacząco poprawił wydajność kodowania względem MPEG-2 i MPEG-4 Part 2. Zasadniczym celem zastosowań H.264 są: telewizja broadcastowa (ang. *broadcasting television*), high definition DVD, cyfrowe przechowywanie danych oraz zastosowania mobilne [8].

Należy podkreślić, że poza standardami kodowania, opracowanymi przez grupy MPEG czy VCEG, istnieją również takie schematy kodowania jak VC-1 (Draft SMPTE Standard) opracowany przez Microsoft oraz RealVideo stworzony przez Real Networks. Media w tego rodzaju formatach są znacząco wykorzystywane do strumieniowania wideo w Internecie.

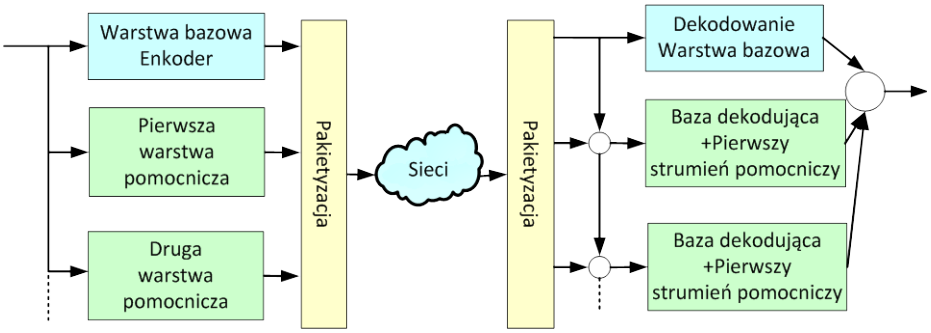
4.3. Skalowalne kodowanie wideo

Nakłady włożone w rozwój schematów skalowalnego kodowania wideo (ang. Scalable Video Coding, SVC) były kontynuowane przez wiele lat w środowisku kodowania wideo w odpowiedzi na pojawiające się na rynku zastosowania transmisji wideo za pośrednictwem heterogenicznych sieci przewodowych lub bezprzewodowych [12, 22, 28, 43].

Zasadniczy cel SVC obejmuje takie zakodowanie sygnału wideo w skalowalny strumień, aby wideo o niższej jakości, rozdzielczości przestrzennej lub czasowej mogło być uzyskiwane przez przycinanie (ang. truncating) skalowalnego strumienia. Tego rodzaju skalowalnie ułatwia również zapewnienie warunków przepustowości oraz dostosowania się do możliwości terminali oraz wymagań co do jakości usług w aplikacjach strumieniowania wideo. Znaczną ilość informacji, poświęconych SVC zamieszczono w [28, 35].

Swego rodzaju protoplastą cech skalowalnego kodowania wideo jest MPEG-2 [14]. W standardzie tym, sygnał wideo jest kodowany do warstwy podstawowej (ang. base layer) oraz kilku warstw pomocniczych (ang. enhancement layers), w ramach których warstwy pomocnicze dołączają poszczególne parametry (w tym SNR) do rekonstruowanej warstwy podstawowej.

Struktura skalowalnego kodowania, opartego na jednej warstwie podstawowej oraz kilku warstwach pomocniczych została przedstawiona na rysunku 4.2. W szczególności, warstwa pomocnicza w ramach skalowania SNR uzupełnia współczynniki dyskretnej transformaty cosinusowej DCT w warstwie bazowej. Dzięki skalowalności przestrzennej, pierwsza warstwa pomocnicza wykorzystuje prognozowanie względem warstwy bazowej bez użycia wektorów ruchu (ang. motion vectors).



Rys. 4.2.Schemat formatu skalowalnego kodowania wideo z jedną warstwą bazową oraz kilkoma warstwami pomocniczymi [35].

W tym przypadku, warstwy mogą mieć różne rozmiary klatek i formatów barwy (ang. chrominance formats). W przeciwieństwie do skalowalności przestrzennej, warstwa pomocnicza w ramach skalowalności czasowej (ang. temporal scalability) wykorzystuje prognozy z warstwy podstawowej za pomocą wektorów ruchu. Niemniej jednak, warstwy te muszą jednocześnie mieć taką samą rozdzielczość przestrzenną i format barw ale mogą posiadać różne częstotliwości odświeżania (ang. frame rates).

Standard MPEG-2 obsługuje każdy z tych skalowalnych trybów, jak również skalowalnie hybrydowe, będące kombinacją dwóch lub więcej rodzajów skalowalności. Należy zauważyć, że strumień warstwy podstawowej i strumienie warstwy pomocniczej mogą być pakietyzowane (ang. packetized) i przesyłane w formie pakietów tego samego kanału lub różnych kanałów, w zależności od struktury sieci.

Z kolei, w standardzie MPEG-4, kodowanie wideo zostało rozszerzone na skalowalność opartą na obiektach, z uwzględnieniem skalowania przestrzennego, czasowego i SNR. Ponadto, jako część standardu wideo MPEG-4 została opracowana nowa forma skalowania, znana jako FGS (ang. Fine Granular Scalability) [22, 23, 35].

W przeciwieństwie do konwencjonalnych skalowalnych schematów kodowania FGS pozwala na znacznie dokładniejsze skalowanie bitów w warstwie pomocniczej. Dokonuje się tego za pośrednictwem metody kodowania płaszczyzny bitowej współczynników DCT w warstwie pomocniczej, co z kolei pozwala by strumień bitowy warstwy pomocniczej mógł być obcięty w dowolnym punkcie. W ten sposób jakość rekonstrukcji klatek jest stopniowo polepszana wraz z liczbą bitów otrzymanych z warstwy pomocniczej. Należy jednak dodać, że FGS ulega znacznej utracie wydajności kompresji przy wyższych szybkościach transmisji, jeżeli tylko

jako odniesienie wykorzystywane są niskiej jakości ramki wideo warstwy bazowej. W związku z tym, do rozwiązania tego problemu zaproponowano ulepszone schematy FGS, w tym: PFGS (ang. Progressive FGS) [49] oraz MC-FGS (ang. Motion-Compensation FGS) [41] itp. W PFGS, warstwy pomocnicze w ramach mogą być prognozowane albo na podstawie warstwy podstawowej, lub ramek referencyjnych warstwy pomocniczej. Ponadto PFGS wprowadza również dryfujący model (ang. drifting model) do estymacji błędów bezpośrednio w enkoderze. W rezultacie, PFGS jest w stanie znacząco poprawić wydajność kodowania, szczególnie przy wyższych szybkościach transmisji. Należy zauważyć, że większość nowych standardów SVC, zawiera omówione wyżej technologie FGS [35].

Mimo, że MPEG-4 FGS posiada pewne zalety w porównaniu z poprzednimi skalowanymi schematami kodowania, to nadal nie znalazł on praktycznego zastosowania. Może być kilka powodów takiej sytuacji. Pierwszym z nich jest wydajność kodowania. System FGS nadal wnosi znaczne sankcje w zakresie wydajności kodowania, co jest okupione tym, że nie chcą go stosować dostawcy treści i usług. Drugim powodem jest wzrost złożoności, a zatem koszt dekoderów.

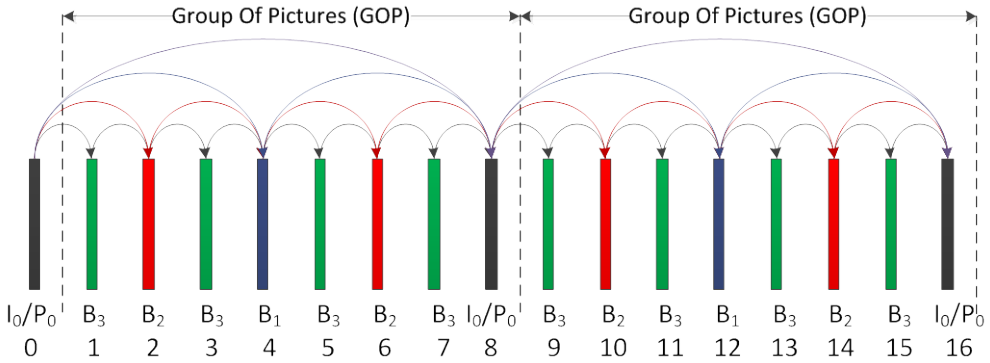
Pomimo tych problemów ze wspominanymi rozwiązaniami, prowadzone są dalsze badania w zakresie opracowania nowych technik skalowalnego kodowania wideo, odkąd skalowalność jest bardzo atrakcyjnym sposobem na zapewnienie uniwersalnego dostępu do multimediów tzw. UMA (ang. Universal Multimedia Access) [43]. Społeczność MPEG nieustannie rozwija standardy kodowania [12].

Standard skalowalnego kodowania wideo (SVC) został opracowany na podstawie narzędzi H.264/AVC i jest rozwijany wspólnie w ramach grupy JDG z MPEG i ITU-T [12]. Oczekuje się, że ten nowy standard będzie w stanie znacząco przezwyciężyć problem strat w wydajności kodowania w porównaniu do istniejących, nieskalowalnych metod kodowania.

Ważnym pojęciem w celu osiągnięcia efektywnego skalowalnego kodowania jest technika MCTF (ang. Motion Compensation Temporal Filtering), oparta schemacie liftingu (ang. lifting scheme) [18, 38]. Schemat ten, stanowi uogólnienie interpolacji predykcyjnej, zaproponowanej przez Wima Sweldensa i pozwala ona na przewidywanie nieparzystych próbek sygnału za pomocą wielomianów stopni znacznie niższych, aniżeli stopień wielomianu interpolacyjnego. Zapewnia doskonałą rekonstrukcję wejścia przy braku kwantyzacji dekomponowanego sygnału, nawet jeśli podczas liftingowania wykonywane są nieliniowe operacje. Korzyści z tego podejścia, w kontekście SVC można znaleźć w [28, 35].

W nowym układzie SVC, wymiary obiektów skalowalnych (ang. scalabilities) obejmują przestrzenne, czasowe i jakościowe (SNR) elementy skalowalne. Z kolei

skalowalność czasowa (ang. temporal scalability) jest możliwa dzięki hierarchicznym obrazom B (ang. B pictures), co pokazano na rysunku 4.3 [12, 35].

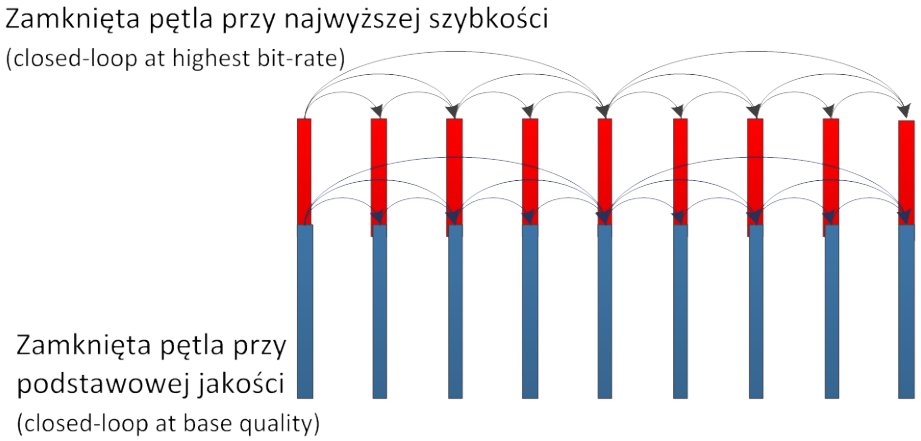


Rys. 4.3. Schemat hierarchicznej struktury kodowania z 4 poziomami czasowymi i GOP w rozmiarze równym 8. Każde zdjęcie B jest prognozowane za pomocą 2 zdjęć referencyjnych na niższym poziomie (poprzedniego i kolejnego) [12, 35]

W niniejszym przykładzie występują cztery poziomy skalowalności czasowej. W pierwszej kolejności kodowane są obrazy ze zgrubną rozdzielczością czasową, a następnie B-zdjęcia są wstawiane przy kolejnym drobniejszym poziomie rozdzielczości czasowej w sposób hierarchiczny. Przestrzenna skalowalność jest osiągana dzięki zastosowaniu podejścia warstwowego, takiego jak w MPEG-2. W celu osiągnięcia skalowalności SNR stosowane są dwa różne podejścia: pierwsze obejmuje wykorzystanie wbudowanych kwantyzacji dla zgrubnej skalowalności (ang. coarse scalability), a drugie stanowi użycie FGS (ang. Fine Grain Scalability), która opiera się na zasadzie kodowania arytmetycznego podpłaszczyzn (ang. sub-bitplane arithmetic coding). W trakcie używania warstw FGS, wykorzystuje się po stronie enkodera dwie zamknięte pętle kompensacji ruchu, w celu poprawy efektywności kodowania. Przy czym, do kodowania warstwy bazowej używana jest pierwsza pętla, a druga - do kodowania warstw pomocniczych.

W celu osiągnięcia lepszej efektywności kodowania, odniesienie do kodowania warstwy pomocniczej odpowiada najwyższej szybkości FGS, co przedstawia rysunek 4.4 [12, 35].

Ostatecznie, należy podkreślić, że technologie związane z odpornością na błędy (ang. error resilience technologies) są bardzo ważne dla strumieni wideo rozpowszechnianych poprzez podatne na błędy sieci bezprzewodowe lub sieci IP. Istnieje kilka bardzo obiecujących technologii, które zostały opracowane dla skalowalnych systemów kodowania wideo [3, 51].

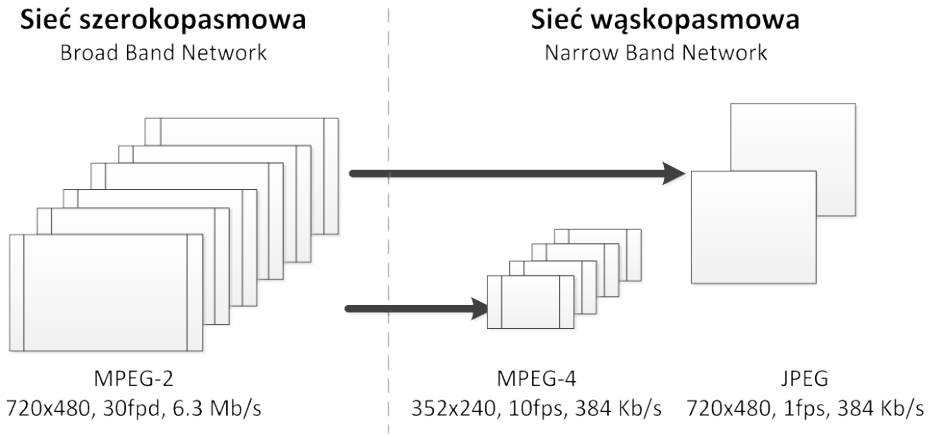


Rys. 4.4. Sterowania enkodera z użyciem dwóch zamkniętych pętli kompensacji ruchu [12,35]

W szczególności, UEP (ang. unequal error protection) stanowi naturalny wybór dla ochrony wideo FGS, ze względu na zróżnicowaną istotność w różnych (poszczególnych) warstwach. Takie techniki okazały się raczej skuteczne [3].

4.4. Transkodowanie sygnału wideo

Teoretycznie, stosunkowo prosta jest obsługa procesu strumieniowania wideo z uwzględnieniem skalowania skompresowanych strumieni bitowych dla serwera, o ile strumień może być łatwo przycięty (ang. truncated) w celu dopasowania do wymagań przepustowości. Jednakże, z powodu szeregu przyczyn, serwery zazwyczaj będą przechowywać nie-skalowalne strumienie bitowe. W tym przypadku, do przenoszenia strumieni do właściwej przepustowości, wymaganej przez sieci lub możliwości urządzenia użytkownika końcowego, może być stosowane transkodowanie. Podstawowe wymagania dla transkodowania wideo obejmują: 1) możliwie najniższą złożoność, w porównaniu do kaskadowej metody pełnego dekodowania i ponownego pełnego kodowania, 2) jakość obrazu nie ulegającą pogorszeniu w porównaniu z podejściem kaskadowego pełnego dekodowania i ponownego, pełnego kodowania. Przykład typowych operacji transkodowania jest pokazany na rysunku 4.5.

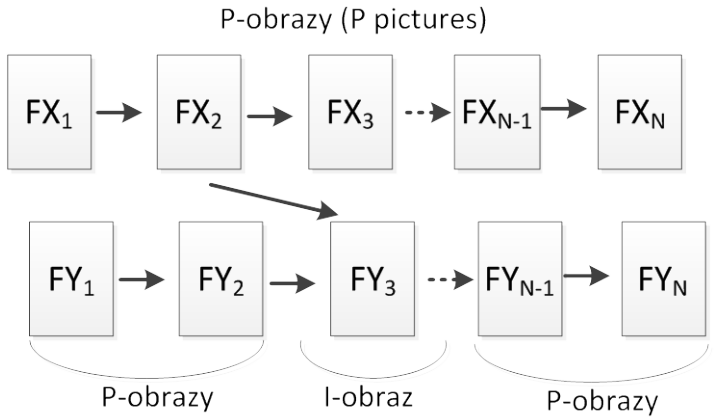


Rys. 4.5. Przykład transkodowanie wideo, w którym strumień MPEG-2 jest konwertowany do MPEG-4 lub JPEG przy mniejszej szybkości transmisji/rozdzielczości

Techniki opracowane do transkodowania mają na celu unikanie pełnego dekodowania i ponownego kodowania strumieni przy jednoczesnym zapewnieniu warunków sieci i możliwości terminali. Przede wszystkim mają one za zadanie znacznie zmniejszyć złożoność konwersji strumieni, przy jednoczesnym zachowaniu wysokiej jakości obrazu. Obszerny zakres dotyczący technik przekodowywania przedstawiono w [42, 50].

4.5. Przełączanie strumieni bitowych

Strumieniowanie wideo stanowi bardzo ważne zastosowanie w sieciach IP i bezprzewodowych sieciach 3G. Jednak, ze względu na zmienne w czasie warunki panujące w sieciach komunikacyjnych, efektywna przepustowość, oferowana użytkownikowi może się znacząco zmieniać. Stąd też, serwer wideo powinien dostosowywać szybkość bitową skompresowanych strumieni wideo lub przełączać się na strumienie bardziej odpowiednie, aby dostosować się do zmian pasma [37]. W przypadku serwerów pracujących z nieskalowalnymi strumieniami, przełączanie zwykle odbywa się w losowo wybranym punkcie dostępu, w sekwencji, np. wewnątrz kodowanej klatki. Zostało to przedstawione na rysunku 4.6.



Rys. 4.6. Dekoder przetwarza strumień FX i dokonuje przełączenia dekodowania

Dla uproszczenia zakłada się, że każda klatka jest prognozowana z pojedynczej referencji. Ponadto, zakłada się, że strumień FX jest kodowany z większą prędkością, a strumień FY jest kodowany przy niższej szybkości transmisji. Po zdekodowania P-obrazów FX1 i FX2 w strumieniu FX, dekoder chce, aby przełączyć się na strumień FY i dokonać dekodowania FY3, FY4 i tak dalej. Główny problem przy takim przełączaniu w strumieni stanowi unikanie dryfu (ang. avoiding drift). Przy czym dryft jest tu rozumiany jako zróżnicowanie (odchylenie) wartości pikseli od oryginalnego obrazu, które stopniowo zwiększa się w kolejnych, prognozowanych ramkach. W kontekście przełączania strumieni, przy próbie prognozowania bieżącej klatki na podstawie innej niż pierwotnie przeznaczona do kodowania również powstanie niezgodność i w rezultacie zjawisko dryfu. Z kolei, w odniesieniu do transkodowania – jest ono zwykle spowodowane utratą wysokiej częstotliwości danych, powodując niedopasowanie pomiędzy aktualną referencyjną klatką dla predykcji w koderze i zdegradowaną klatką referencyjną dla prognozy w transkoderze i dekoderze.

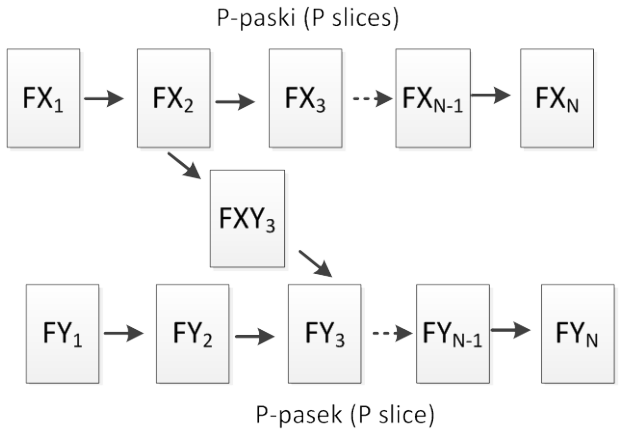
W powyższym przykładzie, ponieważ FY3 stanowi ramkę zakodowaną wewnątrz (ang. intra-coded), można osiągnąć przełączaniem wolnym od dryftu i utraty klatek na skutek przeciążenia sieci. Serwer może dynamicznie przełączać się z wyższych strumieni na niższe, gdy wykryje spadek przepustowości sieci. W ten sposób, przełączanie strumieni jest realizowane poprzez wstawienie I-obrazy (I-Picture) w regularnych odstępach czasu do zakodowanej sekwencji. Pozwala to uzyskać tzw. punkty przełączania (ang. switching points).

Problemem tej metody jest to, że im więcej losowych klatek dostępu (zazwyczaj I-obrazy), dodawanych do nieskalowalnego strumienia, tym większy wpływ na wydajność kodowania. Nadmiar I-klatek zazwyczaj powoduje spadek

wydajności kodowania. Ponadto, ze względu na fakt, że liczba bitów na kod I-klatek jest znacznie większa niż liczba bitów użytych do kodowania P-ramek lub B-ramek, szybkość bitowa ma tendencję do znacznej zmienności w każdym punkcie przełączania. Ta zmiana z kolei wymaga większego bufora i oznacza większe opóźnienia, które mogą nie być do zaakceptowania dla określonych aplikacji czasu rzeczywistego.

W celu rozwiązania powyższych problemów, zaproponowano użycie P-pasków przełączania (ang. Switching P slices, SP) i przełączania I (ang. Switching I, SI), które zostały zaproponowane w standardzie kodowania wideo H.264/AVC [19]. Głównym celem SP i SI jest umożliwienie efektywnego przełączania strumieni wideo, jak również efektywny swobodny dostęp do dekoderów wideo. Dzięki mechanizmowi SP, możliwe jest przejście od jednego strumienia zakodowanego z określoną prędkością do alternatywnego, zakodowanego przy innej prędkości bez powodowania dryf oraz oferując bardziej stabilną szybkość bitową na wyjściu.

Dla uproszczenia, zakłada się dostępne dwa strumienie X i Y, jak pokazano to na rys. 4.7.



Rys.4.7. Przełączanie strumieni z użyciem SP pasków [35]

Po zdekodowaniu P-pasków FX1 i FX2 w strumieniu FX, dekoder chce przełączyć się na strumień Y i rozpocząć dekodowanie FY3, FY4 i tak dalej. Paski SP są umieszczone w punktach przełączania (ang. switching points). Rysunek 7 ilustruje kodowanie z prognozowanie SP-paska FXY3 w odniesieniu do FX2, w celu rekonstrukcji FY3. W ten sposób SP-pasek nie spowoduje szczytu (spiętrzenia) w strumieniu, gdyż jest kodowany za pomocą metody MCP (ang. motion compensated prediction), która jest bardziej wydajna niż intra coding. Ponadto, przełączanie strumieni nie spowoduje dryftu.

W przypadku serwerów pracujących ze strumieniami skalowalnymi, wykorzystanie przełączania strumieni w celu dostosowania do zmian w przepustowości sieci jest stosunkowo łatwe. W większości skalowanych schematów kodowania, sekwencja wideo jest zwykle kodowana do warstwy podstawowej i kilku warstw pomocniczych. Warstwa podstawowa jest zakodowana do nieskalowalnego strumienia i wykonywane jest przycinanie strumienia w warstwach pomocniczych, np. strumień zakodowany w formacie MPEG-4 FGS teoretycznie może zostać ucięty w każdym miejscu w strumieniu warstwy pomocniczej. Jest to właściwe przy dostosowywaniu do zmian przepustowości sieci. Niemniej jednak, jego wydajność kodowania jest znacznie niższa niż w przypadku nie-skalowalnego kodowania wideo, ponieważ kompensacji ruchu opiera się na najniższej jakości warstwie bazowej. Nowoczesne techniki skalowalnego kodowania wideo starają się rozwiązać to zagadnienie przez prognozowanie na podstawie wysokiej jakości obrazu odniesienia z warstwy pomocniczej (ang. high quality reference picture of the enhancement layer) [35].

Niemniej jednak, i w tym podejściu pojawiają się również problemy. Wysokiej jakości strumień referencyjny, na podstawie którego wykonywane jest prognozowanie, może zostać uszkodzony podczas przycinania w warstwie pomocniczej. W celu rozwiązania tych problemów, w pracy [37] zaproponowano układ adaptacyjnego przełączania warstw dwóch skalowalnych strumieni. W tym podejściu, dwa skalowalne strumienie są kodowane z warstwami bazowymi o różnej jakości, a przełączenia wykonywane są tylko na warstwach bazowych skalowalnych strumieni. Do zalet tego rozwiązania można zaliczyć: wysoką wydajność kodowania i brak dryftu przy przełączaniu.

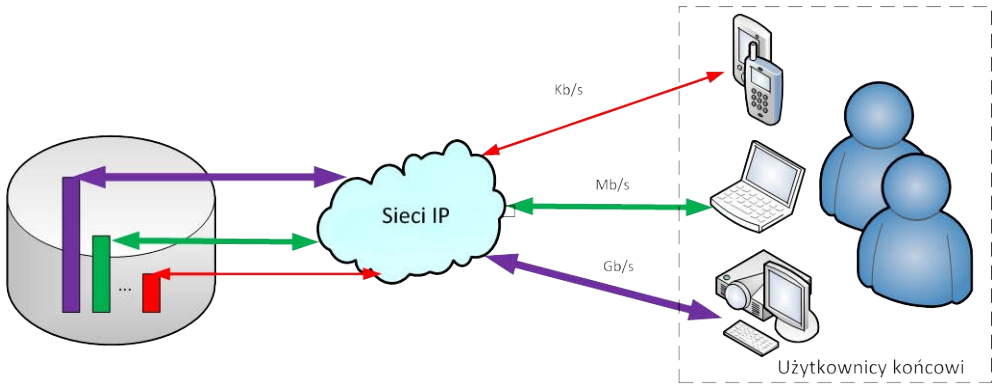
4.6. Przegląd metod strumieniowania wideo

Od strony dekodera, istnieją dwa podejścia do wyświetlania filmów z sieci, które zostały dokładnie zbadane w ostatnich latach. Pierwszym z nich jest podejście oparte na pobieraniu, w którym przed odtwarzaniem musi zostać pobrany cały plik wideo do lokalnej pamięci (na ogół na dysk twardy). Przy takim podejściu czas pobierania pliku wideo wzrasta wraz z ilością danych, która jest proporcjonalna do jakości i czasu trwania wideo. Również w tym podejściu przepustowość sieci również odgrywa znaczącą rolę na etapie pobierania. Alternatywnym sposobem na wyświetlenie treści multimedialnych jest strumieniowanie (ang. video streaming), w ramach, którego, film jest odtwarzany na bieżąco, w trakcie jego przesyłania.

W ramach transmisji strumieniowej, użytkownik może rozpocząć odtwarzanie nagrania wideo, praktycznie bezpośrednio (wł. z niewielkim opóźnieniem) po tym jak rozpocznie się pobieranie. W celu uzyskania płynnego odtwarzanie, dane muszą

być odbierane z szybkością, która pozwala urządzeniu klienta na zdekodowanie i wyświetlenie wszystkich klatek sekwencji video, zgodnie z harmonogramem odtwarzania. Serwer wideo ma dwa sposoby na zapewnienie skompresowanych strumieni wideo. Pierwszym z nich jest wybór jednego spośród wielu nieskalowalnych strumieni, a drugi jest wykorzystanie pojedynczego strumienia bitów, który jest bazuje na skalowalnym kodowaniu wideo lub może być przekodowany (ang. transcoded) w trakcie strumieniowania [5, 7, 35].

W ramach pierwszego sposobu, kilka strumieni tego samego filmu z różnymi szybkościami bitowymi (ang. bitrate), oraz o różnych rozdzielczościach czasowej lub przestrzennej jest uprzednio zapisane na serwerze wideo. Końcowy użytkownik może wybrać strumień w zależności od możliwości jego urządzenia odtwarzającego oraz dostępnej przepustowości sieci komputerowej.
Metoda ta została zilustrowania rysunku 4.8.

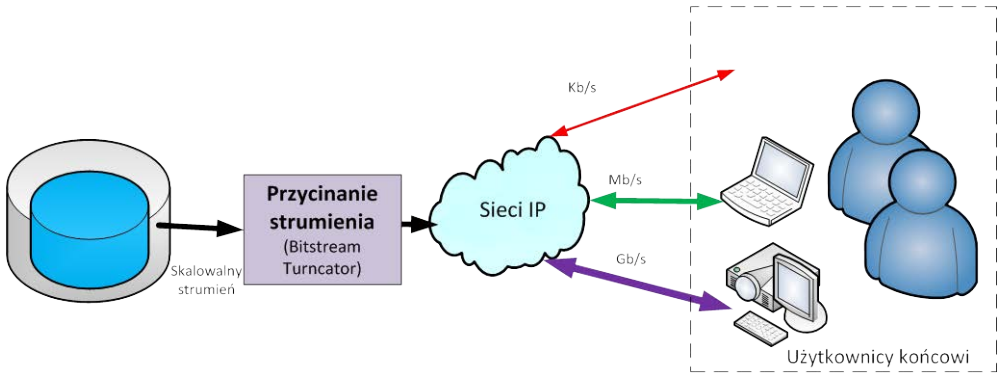


Rys.4.8. Strumieniowanie z użyciem wielu przekodowanych strumieni o różnej prędkości transferu [35]

Do niewątpliwych zalet tej metody należy fakt, że skompresowany strumień jest optymalizowany pod kątem określonego użytkownika, a dekoder ma mniejszą złożoność, ponieważ odbiera i dekoduje on wyłącznie jedną warstwę. Z kolei podstawową wadą omawianego rozwiązania jest to, że serwer wideo musi przechowywać redundantnie wiele strumieni tego samego wideo, a to z kolei może znacznie wpływać na ograniczenia dostępnej pamięci, zwłaszcza przy bardzo dużych repozytoriach. Ponadto, różne wersje wideo muszą być wstępnie zakodowane, co utrudnia jego zastosowania w czasie rzeczywistym (ang. real-time applications). Ponadto, podejście to posiada ułomności w zakresie z ziarnistości (ang. granularity), którą wydaje się powinno zapewniać.

Druga metoda, wykorzystująca skalowalne strumieniowanie wideo (ang. scalable video streaming) jest realizowana z użyciem skalowalnego kodowania strumieni multimedialnych [12, 22, 35]. W ten sposób film jest zakodowany jednorazowo i przechowywany na serwerze wideo.

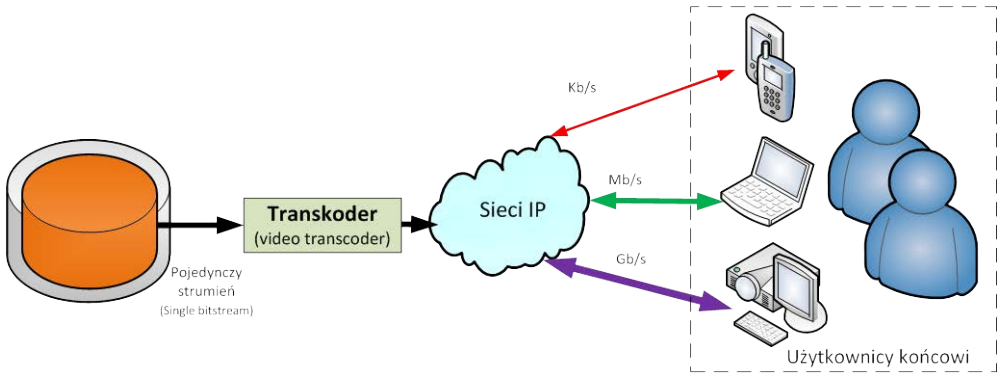
Zakodowany strumień wideo może zostać przycięty na różne sposoby, takie jak SNR, czy skalowanie przestrzenne i czasowe, zależnie od wymagań urządzenia użytkownika końcowego oraz warunków sieci. Zostało to przedstawione na rysunku 4.9.



Rys.4.9. Strumieniowanie ze skalowalnym kodowaniem strumienia [35]

Metoda ta jest atrakcyjna, ponieważ zapewnia większą elastyczność w uzyskaniu pożądanego kompromisu pomiędzy granulacją skalowalności (ang. granular scalability) i wydajnością kodowania. Metoda ta nie powinna wnosić znaczących strat wydajności kodowania w porównaniu do rozwiązań z pojedynczą warstwą kodowania. W przeciwnym razie dostawcy treści multimedialnych nie przyjmą takiego formatu kodowania. Odrębną kwestię stanowi tu złożoność dekodera. W przypadku, gdy wytworzenie takiego dekodera (w tym sprzętowego) będzie zbyt kosztowne, to wdrażanie i rozpowszechnianie go w urządzeniach, pozwalających na odbiór skalowalnych strumieni multimedialnych będzie ograniczone lub wręcz niemożliwe (nie rozwijane).

Trzecim sposób stanowi wykorzystanie pojedynczego zakodowanego strumienia o wyższej jakości, co pokazano na rysunku 4.10. Podczas transmisji strumieniowej, strumienie są dopasowywane do urządzenia użytkownika końcowego i warunków panujących w sieci dzięki konwersji w transkoderze. Główną zaletą tej metody jest to, że techniki transkodowania mogą być stosunkowo łatwo instalowane na serwerach w odpowiedzi na potrzeby bardzo zróżnicowanych sieci i ograniczeń terminali klienckich.



Rys. 4.10. Strumieniowanie z transkodowaniem pojedynczego strumienia [35]

Rozwiązanie to wprowadza swego rodzaju pośrednią warstwę elastyczności pomiędzy dostawcami treści multimedialnych, którzy kodują dane i konsumentami, którzy chcą otrzymywać te dane. Podstawową wadę tego podejścia, w porównaniu do kodowania skalowalnego stanowi fakt, że konwersja w transkoderach zazwyczaj wymaga więcej obliczeń w porównaniu do prostego przycięcia strumienia. Jednak postępy w dziedzinie transkodowania znacząco obniżyły tą złożoność w stosunku do pełnego ponownego kodowania wideo bez znaczących strat jakości [35].

4.7. Porównanie metod strumieniowania

Skalowalne kodowanie określa format danych na etapie kodowania, niezależnie od wymagań transmisji, podczas gdy transkodowanie dokonuje konwersji istniejącego formatu danych w celu dostosowania strumienia do jej bieżących wymagań. Dzięki skalowalnemu kodowaniu, wideo jest kodowane jednorazowo, a następnie ekstrahowane są z niego różne cechy, rozdzielczość przestrzenna i/lub czasowa itp. Pożądanym, idealnym rozwiązaniem byłoby uzyskiwanie tej skalowalnej reprezentacji wideo, bez wpływu na wydajność kodowania.

O ile wydajność kodowania jest niewątpliwe bardzo istotna, należy również wziąć pod uwagę przestrzeń aplikacji (ang. application space). Na przykład, dostawcy treści wysokiej jakości zastosowań mainstream, takich jak DTV i DVD, wykorzystują jednowarstwowe kodowania wideo jak MPEG-2 jako formatu domyślnego. Jest to przyczyną istnienia dużej ilości treści multimedialnych, zakodowanych w formacie MPEG-2, natomiast przemysł zmierza w kierunku formatu kodowania H.264/AVC. Niemniej, w celu uzyskania dostępu do takich treści z różnych urządzeń oraz przy zróżnicowanych warunkach sieci niezbędne jest transkodowanie (przekodowanie tych treści).

Zestawienie wybranych zalet i wad różnych metod strumieniowania wideo zostało podane jest podana w tabeli 4.1.

Tabela 4.1. Cechy metod strumieniowania [35]

Metody strumieniowania	Zalety	Wady
Wiele strumieni	-Wysoka jakość, -Prosty dekoder	-Limitowana liczba strumieni, -Potrzeby znacznych zasobów przechowywania treści
Pojedynczy strumień *(Single scalable bitstream)*	-Małe potrzeby zasobów pamięci do przechowywania treści, -Zastosowania multicastowe, -Proste przełączanie strumieni	-Złożony dekoder -Straty wydajności kodowania
Pojedynczy, nieskalowalny strumień z transkodowaniem *(Single non-scalable bitstreams with transcoding)*	-Prosty dekoder, -Małe potrzeby zasobów pamięci do przechowywania treści, -Zdolność do wstawiania nowych informacji na odporność na błędy	-Możliwość wystąpienia dryftu -Wysoka złożoność -Dodatkowe opóźniena

Mimo, że pojedynczy skalowalny strumień posiada niewielkie zapotrzebowanie na pamięć masową oraz umożliwia proste przełączanie strumieni, to pojawiają się potencjalne straty wydajności kodowania i może być wymagany bardziej skomplikowany dekoder. Z drugiej zaś strony, transkodowanie wymaga dodatkowej złożoności oraz wnosi możliwość potencjalnego wystąpienia opóźnień po stronie transmisji, ale w rezultacie, uzyskany strumień może być odbierany za pomocą standardowych jednowarstwowych dekoderów.

W najbliższym czasie skalowalne kodowanie może zaspokoić szeroki zakres zastosowań sygnałów multimedialnych, takich jak nadzór i strumieniowanie w Internecie, podczas gdy transkodowanie będzie nadal czynnikiem integrującym formaty starszych treści multimedialnych i nowe urządzenia. Może się wydawać, że różne metody transmisji strumieniowej oparte na skalowalnym kodowaniu, transkodowaniu wideo i przełączaniu strumieni mogą być postrzegane, jako technologie przeciwstawne lub konkurencyjne. Niemniej jednak, należy je

rozpatrywać raczej, jako technologie spełniające zróżnicowane potrzeby w danej przestrzeni zastosowań, które powinny ze sobą współistnieć [18, 35].

4.8. Protokoły sieciowe wykorzystywane w skalowalnym strumieniowaniu wideo

TCP (ang. Transfer Control Protocol) stanowi dominujący protokół do transmisji (przesyłania) danych w Internecie. W ogólności może być on również używany do strumieniowej transmisji wideo w sieci Internet [47]. Niemniej jednak, w celu zapewnienia wiarygodnego oraz dobrej jakości strumieniowania z protokołem transportowym TCP, należy rozwiązać kilka problemów. Pierwszy dotyczy trudności obsługi zmienności (szybkości) transmisji danych. W Internecie, szybkość transmisji danych może mieć charakterystykę piłokształtną, czyli addytywnie narasta oraz i opada multiplikatywnie. Drugie zagadnienie stanowi opóźnienie dwupunktowe (ang. end-to-end delay) ze względu na retransmisję w tym samym czasie (połączenia pełnodupleksowe TCP). Problemy te można jednak rozwiązać za pomocą buforowania danych. Stąd, właściwy rozmiar bufora powinien być określany z uwzględnieniem wpływu różnych wskaźników wydajności, takich jak opóźnienia (ang. delay), płynność odtwarzania (ang. smoothness of playback) oraz straty danych (ang. data loss). Ogólnie rzecz biorąc, mały rozmiar bufora oznacza mniejsze opóźnienia pomiędzy czasem od rozpoczęcia transmisji oraz odtwarzaniem pierwszego wyświetlanego obrazu. W zakresie płynności odtwarzania, większy rozmiar bufora zazwyczaj zapewnia płynne odtwarzanie, ponieważ tolerowane są większe różnice w częstotliwości i czasie transmisji. Większe rozmiary bufora oznaczają również mniejszą liczbę odrzucanych pakietów w urządzeniu odbiorczym, ze względu na przepełnienie bufora. Uwzględniając powyższe, istnieje konieczność opracowania analitycznego modelu strumieniowania wideo w ramach TCP. W pracach [4] i [20], przebadano minimalne wymagania odnośnie rozmiaru bufora dla trzech scenariuszy: 1) gdy przepustowość TCP odpowiada szybkości kodowania wideo, 2) gdy przepustowość protokołu TCP jest mniejsza od szybkości kodowania, i 3), gdy przepustowość protokołu TCP jest ograniczona maksymalnym rozmiar okna w ramach mechanizmu okienkowania (ang. sliding window). Kolejny problem ze strumieniowaniem wideo z wykorzystaniem TCP stanowi kwestia obsługi strat pakietów na poziomie warstwy sieciowej. W przypadku, gdy pakiety są opóźnione lub uszkodzone, TCP skutecznie blokuje ruch dopóki nie napłynął oryginały lub kopie tych pakietów, zgodnie z zaimplementowanym w TCP mechanizmem wiarygodnego przesyłania danych (ang. reliable data transfer). W tych kategoriach, TCP słabo nadaje się do

strumieniowania wideo, ponieważ protokół ten obsługuje straty pakietów z metodą retransmisji, powodując tzw. jitter, czyli krótkookresowe odchylenie od ustalonych, okresowych charakterystyk sygnału oraz zjawisko wyścigu (ang. skew) [35].

UDP (ang. User Datagram Protocol) to protokół transportowy dla strumieniowych transmisji wideo [30]. Obsługuje on straty pakietów lub opóźnienie w inny sposób. Umożliwia odrzucanie pakietów przeterminowanych (ang. timeout) lub uszkodzonych (ang. damaged). Dzięki tej funkcji, zaistniała utrata pakietów jest zauważalna przez użytkownika, ale strumień może być w dalszym ciągu odtwarzany. W związku z tym, przy transmisji z udziałem UDP może być niezbędna w dekoderach wideo funkcja ukrywania/niwelacji błędów (ang. error concealment). Odrębny problem z protokołem UDP stanowi fakt, że wiele zapór sieciowych (ang. firewalls) blokuje pakiety UDP i wówczas strumieniowanie z użyciem TCP stanowi jedyną alternatywę (ponieważ może obejść zapory przy użyciu popularnych, znanych numerów portów jak np. HTTP lub RTSP).

RTP (ang. Real-time Transport Protocol) stanowi alternatywy wybór dla strumieniowania wideo [32]. RTP to standardowy protokół internetowy do transportu danych czasu rzeczywistego, w tym audio i wideo. Składa się on z dwóch części: część danych oraz część kontrolnej (sterującej), nazywanej RTCP. Część danych RTP obsługuje transmisję w czasie rzeczywistym do ciągłych mediów, takich jak wideo i audio. Zapewnia rekonstrukcji czasową (ang. timing reconstruction), wykrywanie strat (ang. loss detection), bezpieczeństwa (ang. security) i identyfikację treści (ang. content identification). Z kolei, RTCP (ang. RTP Control Protocol) zapewnia identyfikację źródła i wsparcie dla bram takich jak audio i mostów wideo oraz translacji dystrybucji multicast-to-unicast. Oferuje on informację zwrotną QoS (ang. Quality of Service feedback) od odbiorców do grupy multicastowej, jak również wsparcie dla synchronizacji z różnych strumieni mediów. RTP / RTCP jest powszechnie nadbudowany na protokole UDP i zawiera dodatkowe funkcjonalności na potrzeby transportu mediów. Należy jednak podkreślić, że RTP nie gwarantuje jakości usług (QoS), rezerwacji adresu czy negocjowania formatu mediów [33, 35].

RSVP (ang. Resource reSerVation Protocol) jest protokołem specjalnie zaprojektowanym dla zastosowań strumieniowania [29]. Protokół ten umożliwia aplikacji na przesyłanie danych przez routowalną sieć na żądanie zasobów w każdym węźle sieciowym (ang. node) i próbuje dokonać rezerwacji zasobów dla konkretnego strumienia. Ta cecha protokołu może być stosowana w celu zapewnienia pożądanej jakości usług (QoS), w ramach wiarygodnego połączenia

(ang. reliable connection). Kolejną zaletą protokołu jest jego skalowalność, ponieważ RSVP można skalować nawet do bardzo dużych grup multicastowych. Za wadę można uznać, że węzły sieciowe muszą obsługiwać skomplikowany mechanizm żądań (ang. request mechanism). Ponadto, jeżeli routery nie mogą odpowiednio filtrować tych rezerwacji, to w odbiornikach mogą wystąpić losowe straty pakietów w przypadku małych rezerwacji. Więcej informacji odnośnie sieci IP oraz powyżej przedstawionych protokołów można znaleźć w [18, 33, 35,].

4.9. Skalowalne strumieniowanie z ROI

W poprzednich sekcjach został przedstawiony przegląd różnych metod i narzędzi skalowalnego strumienia wideo. Poniżej omówiono koncepcję skalowalnego strumieniowania wideo zwaną „obszarem zainteresowań" (ang. region-of-interest, ROI). Należy zauważyć, że sama idea ROI nie jest nowa, bo pojawiła się już jako selektywne wzmocnienie (ang. selective enhancement) MPEG-4 FGS [40] oraz w użyciu jest JPEG 2000 [34], będący skalowalnym formatem kodowania obrazu, w celu zilustrowania kluczowych punktów. JPEG 2000 różni się od klasycznego JPEGa i innych istniejących skalowalnych schematów kodowania wideo, które używają nie-skalowalną warstwę bazową, oparte na kodowaniu DCT (ang. discrete cosine transform, DCT – dyskretnej transformaty cosinusowej). JPEG2000 jest skalowalnym formatem kodowania, opartym na dyskretnej transformacie falkowej (ang. discrete wavelet transform, DWT), która dzieli obraz na wysokie i niskie częstotliwości. Część odpowiadająca niskim częstotliwościom może być dzielona dalej w ten sam sposób. Tak przygotowaną tablicę próbek dzieli się na bloki, a następnie kwantuje i koduje niezależnie od siebie. Stopień kompresji jest regulowany poprzez wysłanie tylko niektórych bloków, jak również przez zmienną kwantyzację próbek.

Zaletą JPEG 2000 jest nieco lepsza jakość obrazu przy tym samym stopniu kompresji. W odróżnieniu od JPEG, obraz może być również skompresowany bezstratnie, co czyni nowy standard konkurencyjnym dla formatu PNG. Inna zaleta JPEG 2000 to możliwość przeplotu danych – w miarę odbierania (np. przez sieć) kolejnych próbek obrazu jego jakość stopniowo się poprawia (podobny tryb, choć bardzo uproszczony, oferuje JPEG). Wadą algorytmu JPEG 2000 jest duża złożoność obliczeniowa, w związku z tym nie przewiduje się zastąpienia nim standardu JPEG.

Schemat, użyty w JPEG 2000 jest często określany jako wbudowany system kodowania (ang. embedded coding scheme), ponieważ bity, które odpowiadają różnym jakościom i rozdzielczościom przestrzennym mogą być zorganizowane w strumień bitów, w sposób, który pozwala na stopniową odbudowę zdjęć/obrazów i arbitralne przycinanie w dowolnym punkcie strumienia. W celu

skutecznego dostępu do falkowych współczynników, odpowiadających poszczególnym przestrzennym obszarom obrazu, JPEG2000 wprowadza pojęcie obszaru/okręgu [percinct], który grupuje bloki kodu w większe prostokątne regiony w poziomie rozdzielczości. Każdy taki obszar generuje jeden pakiet. Zbiór pakietów, po jednym z każdego obszaru, na każdym poziomie rozdzielczości, składa się z warstwy jakości (ang. quality layer). Strumień jest zatem zorganizowany w kolejnych warstwach jakości, które tworzą hierarchicznie uporządkowany i osadzony (ang. embedded) strumień [35].

W ciągu ostatnich kilku lat proponowane były różne techniki kodowania ROI dla zdjęć JPEG 2000. Celem tych metod jest zapewnienie wyższego jakości ROI z niższą jakością obszaru tła. Metody te można podzielić na dwie kategorie: statyczne i dynamiczne kodowanie ROI. W ramach statycznego kodowania, ROI jest wybierany i określany w trakcie procedury kodowania. Obejmuje ona metody: max-shift method [34], general wavelet coefficient scale up scheme [25], a bitplane-by-bitplane shift method [46], oraz partial significant bitplane shift method [24]. Główną wadą tych metod jest to, że po zakodowaniu ROI, nie może on już ulec zmianie, co może mieć ograniczenia dla interaktywnych skalowalnych zastosowań strumieniowych, które wymagają większej elastyczności. W celu przezwyciężenia tych niedogodności, opracowano dynamiczne metody ROI opisane w [21, 31]. Pozwalają one na określenie definicji ROI w interaktywnym środowisku, przez dynamicznie wstawianie i rozmieszczanie warstw jakości. Przy tak dynamicznych metodach ROI, jest możliwość przycinania strumienia i zmiany kolejności pakietów w strumieniu w celu spełnienia ograniczeń szybkości i różnic w przepustowości sieci. Stanowi to alternatywny sposób na osiągnięcie skalowalnego strumieniowania wideo, który jest skuteczny w zastosowaniach, wymagających wysokiej jakości ROI, jednocześnie funkcjonujących w sieciach o ograniczonej przepustowości.

Media strumieniowe to technika dostarczania informacji multimedialnej na życzenie, która opiera się na idei strumienia pakietów, interpretowanych po kolei w momencie ich odbioru. Każdy pakiet zawiera ilość informacji wystarczającą do otworzenia pewnego fragmentu prezentacji multimedialnej. Dzięki strumieniowaniu, użytkownik może rozpocząć odtwarzanie danych wideo, zanim zostanie w pełni pobrany z sieci cały plik multimedialny.

W celu dostarczenia medium strumieniowego należy razem złożyć wiele komponentów złożonego systemu, między innymi: infrastruktura komunikacyjna, protokoły sieciowe, oprogramowanie kodujące materiał, serwery strumieniowe oraz odtwarzacze klientów końcowych.

Najczęściej w formie mediów strumieniowych przesyłane są: dźwięk (radio internetowe), obraz (telewizja internetowa) oraz dodatkowe dane opisowe, np. napisy do filmu albo nazwy piosenek. Przy czym charakter i parametry mediów transmisyjnych w znacznej mierze rzutują na sposób realizacji rozwiązań mediów strumieniowych. Internet nie został stworzony do transmisji w czasie rzeczywistym, a pakiety często przesyłane są różnymi drogami, co powoduje zamianę ich kolejności, albo docieranie pakietów do użytkownika w dużych porcjach. W efekcie konieczne jest buforowanie komunikacji - nowe pakiety nie są od razu wyświetlane, tylko umieszczane są w specjalnym buforze, gdzie czekają na swoją kolej. Kluczem do skutecznego stosowania mediów strumieniowych są algorytmy kodowania, zawierające wyrafinowane algorytmy matematyczne.

Problem mediów strumieniowych jest bardzo złożony i nie został w pełni skutecznie rozwiązany. Powoduje to, że istnieje wiele różnych protokołów i formatów transmisji tego rodzaju mediów.

Transmisja strumieniowa odgrywa coraz bardziej znaczącą rolę w ogólnym ruchu sieciowym. Ważne informacje czy ciekawe zdarzenia coraz częściej transmitowane są na żywo w sieci. Tendencje ta nasila się również dzięki możliwościom odtwarzania strumieni nie tylko na komputerach, ale także na najnowszych telewizorach czy telefonach.

Literatura

1. Antosik B., Transmisja internetowa danych multimedialnych w czasie rzeczywistym, WKŁ, 2010r.
2. Apostolopoulos J.G. , Tan W., Wee S.J., Video Streaming: Concepts, Algorithms, and Systems, Mobile and Media Systems Laboratory HP Laboratories Palo Alto, HPL-2002-260, Sept. 2002.
3. Cai H., Zeng B., Shen G., Li S., Error-Resilient Unequal Error Protection of Fine Granularity Scalable Video Bitstreams, accepted by EURASIP Journal on Applied Signal Processing, special issue on Advanced Video Technologies and Applications for H.264/AVC and Beyond.
4. Conklin G., Greenbaum G., Lillevold K., Lippman A. , Reznik Y., Video coding for streaming media delivery on the Internet, IEEE Trans. Circuits Syst. Video Technol., vol. 11, pp. 269-281, Mar. 2001.
5. Farber N., Girod B., Robust H.263 Compatible video transmission for mobile access to video servers, Proc. IEEE Int?l Conf. Image Processing, vol. 2, pp.73-76, Oct. 1997.

6. Gallant M., Kossentini F., Rate-distortion optimized layered coding with unequal error protection for robust internet video, IEEE Trans. Circuits Syst. Video Technol., vol. 11, pp. 357-372, Mar. 2001
7. Girod B., Farber N., Horn U., Scalable codec architecture for internet video on demand, Proc. Asilomar Conf. Signals and Systems, vol. 1, pp.357-361, Nov. 1997.
8. Hanzo L., Cherriman P. J., Streit J., Video Compression and Communications From Basics to H.261, H.263, H.264, MPEG4 for DVB and HSDPA-Style Adaptive Turbo-Transceivers Second Edition
9. ISO/IEC 14496-10 AVC or ITU-T Rec. H.264, September 2003.
10. ISO/IEC IS 10918-1:1994 | ITU-T Recommendation T.81 (JPEG), Digital compression and coding of continuous-tone still images, 1994.
11. ISO/IEC IS 15444-1:2004 | ITU-T Recommendation T.800, JPEG 2000 image coding system: Core coding system, 2004.
12. ISO/IEC JTC 1/SC 29/WG 11 N7555, Working Draft 4 of ISO/IEC 14496-10:2005/AMD3 Scalable Video Coding, October 2005, Nice France.
13. ISO/IEC JTC1 IS 11172 (MPEG-1), Coding of moving picture and coding of continuous audio for digital storage media up to 1.5 Mbps, 1992.
14. ISO/IEC JTC1 IS 13818 (MPEG-2), Generic coding of moving pictures and associated audio, 1994.
15. ISO/IEC JTC1 IS 14386 (MPEG-4), Generic Coding of Moving Pictures and Associated Audio, 2000.
16. ITU-T Recommendation H.261, Video Codec for Audiovisual Services at px64 Kbit/s, March 1993.
17. ITU-T Recommendation H.263, Video Coding for Low Bit Rate Communication, Draft H.263, May 2, 1996
18. Jaromin M., Aproksymacja sygnaˆu na podstawie schematu lifting, Studia Informatica, rok: 2009, Vol. 30, nr 3A, s. 5--23
19. Karczewisz M., Kurceren R., The SP- and SI-Frames Design for H.264/AVC, IEEE Trans. Circuits Syst. Video Technol., vol. 13, no. 7, pp. 637-644, July 2003.
20. Kim T. , Scalable video streaming over internet, Ph.D. Thesis, School of Electrical and Computer Engineering, Georgia Institute of Technology, Jan. 2005.
21. Kong H.S., Vetro A., Hata T., Kuwahara N., Fast Region-of-Interest Transcoding for JPEG 2000 Images, Proc IEEE Intl Symp. Circuits and Systems, Kobe, Japan, May 2005.
22. Li W. , Overview of Fine Granularity Scalability in MPEG-4 Video Standard, IEEE Trans. Circuits Syst. Video Technol., vol. 11, pp. 301?317, 2001.

23. Li W. , Streaming video profile in MPEG-4, IEEE Trans. Circuits Syst. Video Technol., vol. 11, pp. 301?317, Mar. 2001
24. Liu L., Fan G., A new JPEG 2000 region-of-interest image coding method: partial significant bitplanes shift, IEEE Signal Processing Letters, vol. 10, no. 2, pp. 35-38, Feb. 2003.
25. Liu L., Fan G., A new JPEG 2000 region-of-interest image coding method: partial significant bitplanes shift, IEEE Signal Processing Letters, vol. 10, no. 2, pp. 35-38, Feb. 2003.
26. Liu Y., Salama P., Li Z., Delp E.J., Error Resilient Scalable Video Streaming: Combining Nested Scalability and Parallel Scalability, VIPER Laboratory Report, School of Electrical and Computer Engineering, Purdue University, January 2003. [online at: (accessed on August 25, 2006), http://cobweb.ecn.purdue.edu/~yuxin/resources/Zoe_MDC_Leaky.pdf]
27. Lu J., Reactive and proactive approaches to media streaming: from scalable coding to content delivery networks, Proc. Int?l Conf. Information Technology: Coding and Computing, April 2-4, 2001, Las Vegas, pp. 5-9.
28. Ohm J.-R., Advances in Scalable Video Coding, Proc. IEEE, vol. 93, no. 1, Special Issue on Advances in Video Coding and Delivery, pp. 42-56, Jan. 2005.
29. Online at: http://www.isi.edu/rsvp/overview.html
30. Postel J., RFC 768 - User Datagram Protocol, August 1980.
31. Rosenbaum R., Schumann H., Flexible, dynamic and compliant region of interest coding in JPEG 2000, Proc. IEEE Int?l Conference on Image Processing, pp. 101-104, Rochester, New York, Sept. 2002.
32. Schulzrinne H. et al, RFC 1889 - RTP: A Transport Protocol for Real-Time Applications, January 1996.
33. Schulzrinne H., IP Networks, in Compressed Video Over Networks, , Amy Reibman and Ming-Ting Sun (eds.) , Marcel Dekker , 2001.
34. Skodras A., Christopoulos C., Ebrabimi T., The JPEG2000 Still Image Compression Stadard, IEEE Signal Processing Magazine, vol. 18, no. 5, pp.36-58, Sept. 2001.
35. Sun H., Vetro A., Xin J., An Overview of Scalable Video Streaming
36. Sun X. , Wu F., Li S., Gao W., Zhang Y.Q., Seamless Switching of Scalable Video Bitstreams for Efficient Streaming, IEEE Transaction on Multimedia, vol. 6, no. 2, pp. 291-303, Apr. 2004.
37. Sun X., Wu F., Li S., Gao W., Zhang Q.Y., Seamless Switching of Scalable Video Bitstreams for Efficient Streaming, IEEE Multimedia, vol. 6, no. 2, pp. 291-303, April 2004.

38. Sweldens W., A custom-design construction of biorthogonal wavelets, J. Appl. Comp. Harm. Anal., vol. 3, no. 2, pp. 186-200, 1996.
39. Tan W., Zakhor A. , Video multicast using layered FEC and scalable compression, IEEE Trans Circuits Syst. Video Technol., vol. 11, pp. 373?386, Mar. 2001.
40. Van der Schaar M., Lin Y.-T., Content-basd selective enhancement for streaming video, Proc. Int?l Conf. Image Processing, pp. 977-980, Oct. 2001.
41. Van der Schaar M., Radha H., Motion-compensation Fine-GranularScalability (MC-FGS) for wireless multimedia, Proceedings of IEEE Symposium on Multimedia Signal Processing (Special Session on Mobile Multimedia Communications), October 2001.
42. Vetro A., Christopoulos C., Sun H., An overview of video transcoding architectures and techniques, IEEE Signal Processing Magazine, vol. 20, no. 2, pp. 18-29, Mar 2003.
43. Vetro A., Timmerer C., Digital Item Adaptation: Overview of Standardization and Research Activities, IEEE Trans. Multimedia, vol. 7, no. 3, pp. 418-426, June 2005.
44. Vetro A., Xin J., Sun H., Error-Resilience Video Transcoding for Wireless Communications , IEEE Wireless Communications, vol. 12, no. 4, Aug. 2005.
45. Wang Y., Zhu Q.F., Error control and concealment for video communications: A reviews, Proc. IEEE, vol.86, no. 5, pp. 974-997, May 1998.
46. Wang Z., Bovik A.C., Bitplane-by-bitplane shift (BbBShift) ? A suggestion for JPEG 2000 region of interest coding, IEEE Signal Processing Letters, vol. 9, pp. 160-162, May 2002.
47. Widmer J., Denda R., Mauve M., A survey on TCP-Friendly congestion control, IEEE Network Magazine, vol. 15, pp. 28-37, May/June 2001.
48. Wu D. , Hou Y. T., Zhu W. (et all), Streaming video over the internet: Approaches and directions, IEEE Trans. Circuits Syst. Video Technol., vol. 11, pp. 282?300, Mar. 2001.
49. Wu F., Li S., Zhang Q.Y., A framework for efficient progressive fine granularity scalable video coding, IEEE Trans. Circuits Syst. Video Technol., vol. 11, no. 3, pp. 332-344, Mar 2001.
50. Xin J., Lin C.W., Sun M.T., Digital Video Transcoding, Proc. IEEE, vol. 93, no. 1, Special Issue on Advances in Video Coding and Delivery, pp. 84-97, Jan. 2005.
51. Yan R., Wu F., Li S., Tao R., Error Resilience Methods for FGS Video Enhancement Bitstream , The First IEEE Pacific-Rim Conference on Multimedia (IEEE-PCM 2000), Dec. 13-15, 2000 Sydney, Australia.

52. Yang F., Zhang Q., Zhu W., Zhang Q. Y., Bit Allocation for Scalable Video Streaming over Mobile Wireless Internet, Infocom, 2004.
53. Ziliani F., Michelou J-C. , Scalable Video Coding in Digital Video Security, White paper, VisioWave, 2005.

5. Algorytmy kompresji i rozpoznawania obrazów

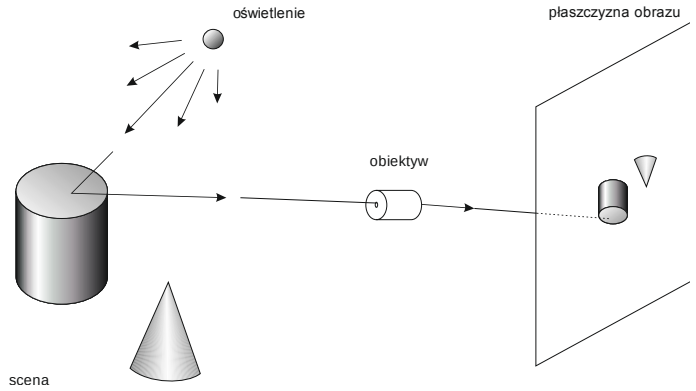
5.1. Wstęp

Obrazy widzialne stanowią blisko 90% informacji odbieranej przez człowieka [1]. Są jej specyficzną formą, charakteryzując się przy tym nadmiarowością i niejednoznacznością. Ta sama scena może być postrzegana przez różne osoby w niejednakowy sposób. Przetwarzanie informacji obrazowej jest jednym z ważniejszych obszarów zastosowań informatyki. Stało się to możliwe dzięki postępowi technologicznemu tak w przypadku przetworników obrazowych (CCD, CMOS) i urządzeń wyświetlających obraz jak również mocy obliczeniowej współczesnych systemów komputerowych.

Obraz w rozumieniu technicznym stanowi przestrzenny rozkład informacji, niezależnie od rodzaju zastosowanego nośnika, którym może być promieniowanie elektromagnetyczne, wiązka elektronów, ale również potencjał elektrostatyczny, chemiczny, miara ukształtowania powierzchni, itp. [2]. Pojęcie informacji przestrzennej, powszechnie utożsamianej z informacją optyczną, jest więc pojęciem szerszym i niezależnym od nośnika informacji. W przeważającej mierze, rolę optycznego nośnika informacji pełni promieniowanie elektromagnetyczne w zakresie światła widzialnego.

Proces przetwarzania informacji odbywa się w tym przypadku w układzie przetwarzania obrazu. Pierwszym etapem tego procesu jest obrazowanie, które polega na przeniesieniu i zlokalizowaniu, przy zachowaniu relacji sąsiedztwa, dowolnego przedmiotu punktowego w polu widzenia w inny, niewielki obszar, umownej przestrzeni obrazowej. Wspomniane przeniesienie i lokalizacja, zachodzą wokół wyróżnionego w przestrzeni obrazowej obszarze geometrycznego obrazu tego przedmiotu [3].

Proces ten dokonywany są najczęściej za pomocą klasycznych układów optycznych wytwarzających obraz (obiektywy) i przedstawiony jest na rys. 5.1. Z punktu widzenia teorii informacji, obrazowanie można traktować jako linię łączności, w której poszczególnym kanałom, mającym charakter przestrzenny, odpowiadają rozłożone w płaszczyźnie obrazu wyjściowego obszary, stanowiące najmniejsze elementy obrazu [3]. Wspomniane elementy nazywane są pikselami, w praktyce których rozmiar determinowany jest budową przetwornika obrazu (CMOS, CCD). W danej chwili czasowej, kanał może zawierać informację o jednym z rozróżnialnych poziomów intensywności, przy czym ma on ograniczone pasmo przenoszenia w dziedzinie czasu, wynikające z bezwładności układu obrazowania.



Rys.5.1.Oświetlany obiekt i jego obraz

Całkowita ilość informacji przenoszona przez układ obrazowania zależy od:

- ilości informacji strukturalnej, która jest związana z liczbą pikseli;
- ilości informacji metrycznej, uzależnionej od liczby rozróżnialnych poziomów intensywności;
- ilość informacji czasowej, która wynikają z liczby przenoszonych kadrów w jednostce czasu.

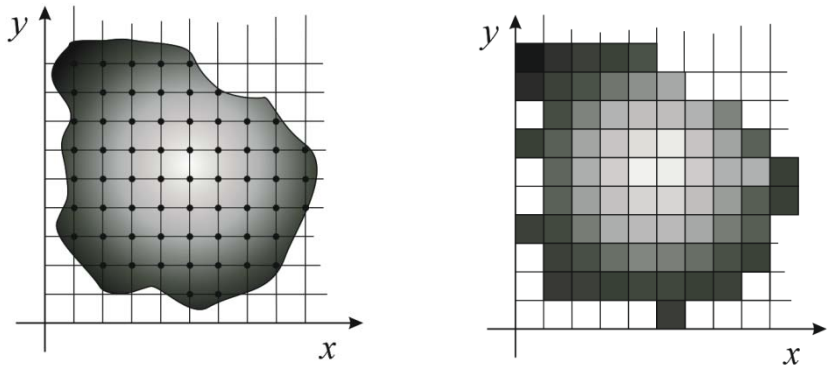
5.2. Reprezentacja obrazu

Obraz można traktować jako dwuwymiarowy sygnał reprezentowany jest przez funkcję:

$$I = \mathrm{f}(x, y), \tag{1}$$

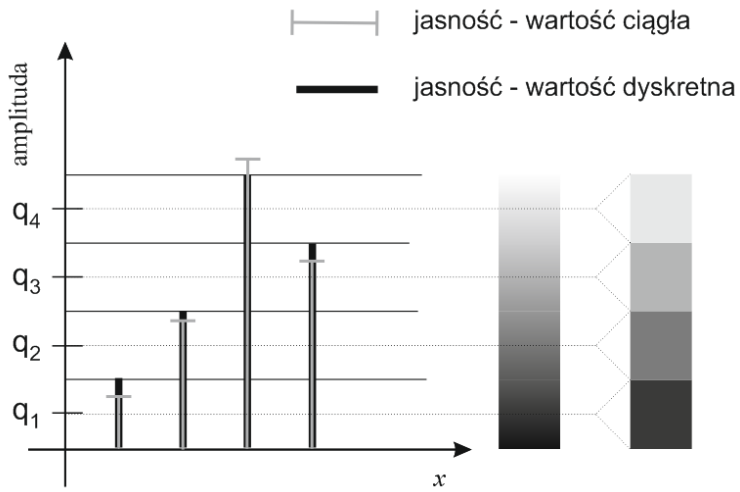
Przy czym *I* oznacza irradiancję [W/m^2], natomiast *x*, *y* są współrzędnymi przestrzennymi. Powyższa postać funkcji obrazu reprezentuje jedynie rozkład jasności na płaszczyźnie i dotyczy wyłącznie obrazów monochromatycznych. Ze względu na upowszechnienie cyfrowych metod przetwarzania informacji, zarówno współrzędne przestrzenne, jak i wartości funkcji muszą przyjmować wartości dyskretne. Dokonuje się tego w wyniku dwóch operacji: próbkowania dwuwymiarowego oraz kwantyzacji.

Próbkowanie dwuwymiarowe polega na pobraniu wartości jasności w ściśle określonych punktach obrazu, jak pokazano na rys. 5.2. Najczęściej spotykanym rozwiązaniem są węzły siatki prostokątnej, chociaż wykorzystuje się także siatkę sześciokątną. W wyniku tej operacji otrzymuje się ciągłą funkcję jasności określoną dla dyskretnych wartości współrzędnych.



Rys. 5.2. Próbkowanie przestrzenne obrazu analogowego

Aby jednak obraz mógł być przetworzony przez system cyfrowy, uzyskane wartości ciągłe muszą zostać skwantowane a następnie zmienione na postać cyfrową. Operacja kwantyzacji polega na podzieleniu całego dopuszczalnego zakresu jasności na przedziały zmienności, najczęściej o jednakowej długości i przypisaniu tym przedziałom jednej wartości, jak pokazane to zostało na rys. 5.3.



Rys. 5.3. Operacja kwantyzacji ciągłych wartości jasności

Warto w tym miejscu nadmienić, że zwiększenie liczby przedziałów, w których dokonywana jest kwantyzacja skutkuje mniejszym błędem związanym z dyskretyzacją jasności (błąd kwantyzacji). Zazwyczaj, liczba tych przedziałów wynosi 28 = 256, co oznacza, że jasność można zapisać za pomocą 1 bajta (8 bitów), przy czym 0 odpowiada poziomowi czerni, natomiast 255 – poziomowi bieli. Taka liczba rozróżnianych poziomów jest w praktyce wystarczająca, ponieważ oko ludzkie postrzega je jako ciągły w sensie jasności [3].

Liczba przedziałów kwantyzacji bezpośrednio wpływa na wielkość pamięci niezbędnej do przechowywania obrazu. W przypadku obrazów binarnych do zapisania informacji o jasności pojedynczego piksela wystarcza 1 bit, co oznacza, że dozwolone są tylko dwa poziomy jasności – czerń i biel. Przechowanie obrazu monochromatycznego o rozdzielczości 32x32 pikseli będzie wymagało 1024 bitów. Ilość potrzebnych bitów szybko wzrasta wraz ze zwiększeniem rozmiarów obrazu, rozumianej jako liczba reprezentujących go pikseli i liczby poziomów szarości. Przykładowo, przechowanie obrazu o 256 poziomach szarości w rozdzielczości HD (1920x1080) wymaga 1920x1080x256 = 530 841 600 bitów (ponad 63MB).

Ważna klasą obrazów stanowią obrazy barwne. Wrażenie barwy związane jest z długością fali świetlnej. Stosowane w kolorymetrii I prawo Grassmanna stanowi, że każda dowolna barwa może być odwzorowana za pomocą trzech barw składowych (podstawowych), czyli takich, które są niezależne pod względem kolorymetrycznym (nie możliwe jest uzyskanie dowolnej barwy składowej z dwóch pozostałych) [4].

W celu zapamiętywania obrazów kolorowych stosowane są rozmaite modele (przestrzenie) barw. W zależności od obszaru zastosowań, wyróżnia się dwa rodzaje przestrzeni barw:

- addytywne (np. RGB, YUV, HSB, XYZ),
- subtraktywne (np. CMY, CMYK).

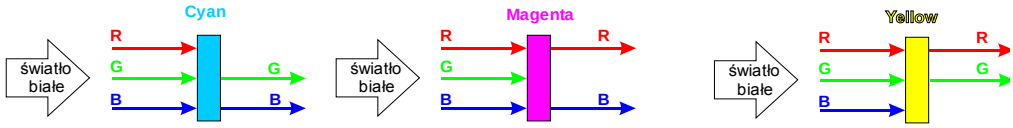
W pierwszym przypadku których barwy pośrednie uzyskuje się poprzez wymieszanie promieniowania pochodzącego od trzech źródeł odpowiadającym trzem barwom podstawowym. Najczęściej stosowaną w praktyce jest przestrzeń RGB (ang. Red, Green, Blue), przy czym w urządzeniach wejściowych (kamery, skanery) i wyjściowych (monitory, projektory) stosowane są różne wersje tej przestrzeni różniące się definicjami barw podstawowych. Jeżeli każda barwa składowa zostanie odwzorowana za pomocą 8 bitów, wówczas całkowita liczba odcieni barw będzie wynosić $28 \times 28 \times 28 \cong 16{,}7$ mln (kolor 24 bitowy, tzw. True Color). Liczba ta znacznie przekracza możliwości percepcyjne oka ludzkiego, jak również urządzeń wyjściowych [4].

Oprócz różnych odmian przestrzeni RGB (np. sRGB, e-sRGB, Adobe RGB, ProPhotoRGB) stosowane są również inne przestrzenie, wśród których szczególne znaczenie w przypadku przesyłania obrazów na odległość (telewizja) i kompresji obrazów mają przestrzenie zawierające składową luminancji i składowe chrominancji, np. YUV, stosowany w standardzie telewizyjnym PAL, YIQ (NTSC), czy też YCBCR, stosowane w algorytmach kompresji obrazów. Luminancja (Y) zawiera informację o jasności, natomiast składowe chrominancji zawierają pełną informacji

o kolorze. YCBCR nie jest bezwzględną przestrzenią kolorów, a sposobem na opisanie informacji na podstawie danych o kolorze w postaci RGB. Wartość YCBCR można określić tylko na podstawie tych danych według zależności:

$$\begin{pmatrix} Y \\ C_B \\ C_V \end{pmatrix} = \begin{pmatrix} 16 \\ 128 \\ 128 \end{pmatrix} + \frac{1}{256} \begin{pmatrix} 65{,}378 & 129{,}057 & 25{,}064 \\ -37{,}945 & -74{,}494 & 112{,}439 \\ 112{,}439 & -94{,}154 & -18{,}285 \end{pmatrix} \cdot \begin{pmatrix} R_N \\ G_N \\ B_N \end{pmatrix} \qquad (2)$$

Modele subtraktywne stosowane są w poligrafii i polegają na wymieszanie barw odbitych od powierzchni, np. powierzchni papieru. Jeśli obiekt oświetlony światłem białym postrzegany jest np. jako zielony, oznacza to że jego powierzchnia odbija barwę zieloną, a pochłania barwy, będące dopełnieniem barwy zielonej. Barwy podstawowe w modelu CMY (Cyan, Magenta, Yellow) są dopełnieniami barw składających się na model RGB, jak pokazano na rys. 5.4.



Rys. 5.4. Barwy składowe modelu CMY

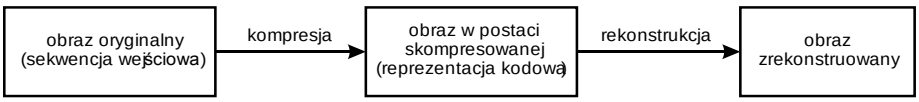
Kolor czarny w modelu CMY można uzyskać mieszając w równych proporcjach poszczególne barwy składowe. Jednak w praktyce tak uzyskana kolor czarny będzie charakteryzował się dominantą jednej z barw składowych. Z tego względu, oraz mając na uwagę fakt, że kolor czarny stosunkowo często używany jest w wydrukach (np. w dokumentach), częściej stosowany jest model CMY uzupełniony dodatkowo barwą czarną – CMYK (Cyan. Magenta, Yellow, blacK).

5.3. Kompresja obrazów

Upowszechnienie cyfrowych metod pozyskiwania, przechowywania oraz przesyłania obrazów nie byłoby możliwe bez algorytmów kompresji, pomimo nieustannego spadku cen pamięci w przeliczeniu na jednostkę pojemności. Kompresja obrazów, zarówno nieruchomych jak i ruchomych wykorzystuje niedoskonałości ludzkiego wzroku, co pozwala na znaczne zmniejszenie wymaganej pamięci przy stosunkowo niewielkiej utracie jakości obrazu.

Przez algorytm kompresji rozumieć należy algorytm kodowania oraz algorytm rekonstrukcji (dekompresji). Algorytm kodowania dla pewnych danych wejściowych X generuje pewną reprezentację XC, która wymaga mniejszej ilości bitów. Algorytm rekonstrukcji na podstawie skompresowanej reprezentacji danych XC, tworzy

rekonstrukcję Y. Rysunek 5.5 przedstawia ogólny schemat działania tych algorytmów [5].



Rys. 5.5. Ogólny schemat procesu kompresji i rekonstrukcji

Algorytmy kompresji dzieli się na dwie klasy [4]:

- bezstratne,
- stratne.

W schemacie kompresji bezstratnej wymagane jest, aby obraz źródłowy X był identyczny z rekonstruowanym w odróżnieniu do schematu kompresji stratnej, w której dopuszcza się, aby X i Y były różne od siebie.

Inna klasyfikacja algorytmów uwzględnia sposób ich działania, wśród których wyróżnia się algorytmy: statyczne, póładaptacyjne i adaptacyjne.

Algorytmy statyczne działają w jednakowy sposób dla całej sekwencji kompresowanych danych i nie istnieje konieczność dopisywania parametrów z jakimi działa algorytm w danych wyjściowej. Działanie algorytmów póładaptacyjnych jest identyczne jak algorytmów statycznych, ale tylko dla części danych. Parametry z jakimi ona działają wyznaczane są na podstawie wstępnej analizy przetwarzanej części danych, stąd istnieje konieczność dopisywania parametrów algorytmu w nagłówku sekwencji wyjściowej. Algorytmy adaptacyjne dostosowują swoje działanie dynamicznie do lokalnej statystyki danych źródłowych na podstawie analizy wejściowej sekwencji danych. Nie istnieje w takim przypadku konieczność dopisywania parametrów działania algorytmów adaptacyjnych w sekwencji wyjściowe. Cechują się ona największą elastycznością, ale okupione jest to stosunkowo dużą złożonością obliczeniową.

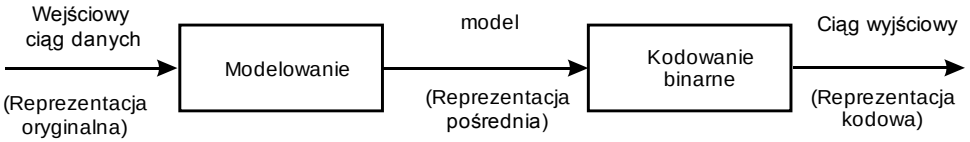
Kompresja bezstratna (kodownie)

Kompresję bezstratną stosuje się wszędzie tam, gdzie nie dopuszcza się utraty (przekłamania) kompresowanej informacji. Typowym przykładem jest archiwizacja programów lub danych komputerowych, gdzie zmiana jakiegokolwiek bitu jest równoważna ze zmianą kodu źródłowego programu, przez co program może w ogóle nie dać się uruchomić. W przypadku obrazów, bezstratna kompresja stosowana bywa m.in. w diagnostyce medycznej oraz niektórych zastosowaniach wojskowych, szczególnie gdy informacja obrazowa zawiera dane o szczególnym znaczeniu.

Na kompresję bezstratną, niezależnie od rodzaju danych (obraz, dźwięk) składają się dwie następujące po sobie etapy:

- modelowania,
- kodowania binarnego.

Faza modelowania polega na użyciu odpowiednio skutecznych modeli statystycznych i predykcyjnych w celu oszczędnego opisu lokalnych zależności danych. Wynikiem jej działania jest utworzenie reprezentacji pośredniej, do której dobiera się odpowiednią metodę kodowania sekwencji źródłowej, poddawanej w następnej fazie kodowaniu binarnemu, jak zostało to pokazane na rys. 5.6. Kodowanie binarne polega ono przypisaniu ciągu bitów (słów kodowych) poszczególnym danym (poszczególnym symbolom alfabetu źródła informacji), aby w możliwie oszczędny sposób, utworzyć wyjściową sekwencję bitową reprezentującą dane lokalne[5,6].



Rys. 5.6. Etapy kompresji bezstratnej

Utworzenie reprezentacji pośredniej w fazie modelowania możliwe jest poprzez:

- przekształcenie danych z przestrzeni oryginalnej w inną z wykorzystaniem metrycznych zależności danych, określonego sposobu porządkowania danych lub zmiany wymiarowości oryginalnej dziedziny danych,
- utworzenie modelu deterministycznego, opisującego bezpośrednie (chwilowe, lokalne lub globalne) właściwości danych, np. ciagi jednakowych danych,
- utworzenie modelu probabilistycznego źródła informacji (przy założeniu stacjonarności i ergodyczności źródła) na podstawie określonego kontekstu wystąpienia danych.

Faza modelowania obejmuje kilka rodzajów modeli, wśród których można wymienić:

- uproszczony model statystyczny, stosowany np. w kodzie RLE (ang. Run-Length Encoding) kodowaniu długości serii dla serii bitów (koder Z) lub bajtów (format PCX), kodzie Golomba (stosowanym w standardzie JPEG-LS), oraz Huffmana (wykorzystywanym w standardzie JPEG),
- rozbudowany dynamiczny model kontekstowy wyższych rzędów,

- metody słownikowe (np. format GIF, PNG),
- wstępną dekompozycję danych gdy tworzona jest reprezentacja pośrednia, stosowana w przypadkach:
    - metod predykcyjnych i interpolacyjnych,
    - kodowaniu map bitowych,
    - skanowaniu danych wg określonego porządku,
    - całkowitoliczbowym kodowaniu transformacyjnym, (np. AVC)
    - całkowitoliczbowym dekompozycjom pasmowym i falkowym (np. JPEG2000).

W przypadku kodowania binarnego stosowane są następujące rozwiązania:

- Przypisywanie słów kodowych pojedynczym symbolom alfabetu źródła – Metody Huffmana, Shannona-Fano, Rice'a – kody o zmiennej długości słów,
- Przypisywanie słów kodowych (zazwyczaj o stałej długości) ciągom symboli wejściowych o zmiennej długości (RLE, kodowanie słownikowe). Modyfikacja tych metodo polega na użyciu adaptacyjnych koderów słownikowych o zmiennym rozmiarze słownika (a więc i indeksów),
- Utworzenie sekwencji kodowej w postaci jednego binarnego słowa kodowego wyznaczanego sukcesywnie dla całego ciągu wejściowego.

Do najważniejszych metod bezstratnych należą metody entropijne [6]. Należą do nich kodowanie słownikowe, Huffmana oraz kodowanie arytmetyczne. W tych metodach, współczynnik kompresji jest bliski współczynnikowi optymalnemu, wyznaczonymi na podstawie entropii danego sygnału.

5.3.1. Kodowanie Huffmana

Kodowanie metodą Huffmana opiera się na założeniach metody Shannona. Podobnie jak w metodzie Shannona-Fano, tworzone są bitowe słowa kodowe o zmiennej długości przyporządkowane poszczególnym symbolom alfabetu. Kody generowane przy użyciu tej metody są nazwane kodami Huffmana, które są kodami prefiksowymi. Są one optymalne dla przyjętego modelu probabilistycznego.

Algorytm Huffmana odwołuje się do dwóch obserwacji, które dotyczą optymalnych kodów prefiksowych. Pierwsza z nich wskazuje, że symbolom, które mają większe prawdopodobieństwo wystąpienia, odpowiadają w kodzie optymalnym krótsze słowa kodowe niż symbolom, które występują rzadziej. Jeżeli symbole występujące częściej mają dłuższe słowa kodowe niż symbole, które występują rzadziej, to średnia liczba bitów na symbol będzie większa niż w sytuacji odwrotnej. Dlatego kod, który przyporządkowuje dłuższe słowa kodowe symbolom występującym częściej, nie spełnia warunków kodu optymalnego.

Na podstawie drugiej obserwacji można stwierdzić, że dwa najrzadziej występujące symbole maja w kodzie optymalnym słowa kodowe o tej samej długości. Aby się przekonać o tym, że obserwacja jest właściwa, należy rozważyć następującą obserwacje. Istnieje kod optymalny , w którym słowa kodowe odpowiadające dwóm najrzadziej występującym symbolom różnią się między sobą długością. Dłuższe słowo kodowe jest o k bitów dłuższe od krótszego słowa kodowego. Kod ten jest kodem prefiksowym, stąd krótsze z tych słów kodowych nie może być prefiksem dłuższego słowa kodowego. Oznacza to, że pomijając ostatnie k bitów dłuższego słowa kodowego, obydwa słowa kodowe nadal będą różne. Ponieważ te słowa odpowiadają symbolom o najmniejszych prawdopodobieństwach, więc żadne inne słowo nie może być krótsze od tych dwóch słów. A więc nie ma niebezpieczeństwa, że skrócone słowo kodowe stanie się prefiksem innego słowa kodowego. Pomijając także te ostanie k bitów, otrzymamy nowy kod, który ma krótszą średnią długość niż kod . Lecz to przeczy założeniu nie wprost, że jest kodem optymalnym. Zatem kod optymalny spełnia warunek podany w drugiej obserwacji [5].

W metodzie słownikowej fragmenty, strumienia bitów zastępuje się odpowiednim indeksem do specjalnie stworzonej książki kodowej (słownika), która może zmieniać się adaptacyjnie, tzn. może dostosowywać się do kompresowanego sygnału. Jeśli słownik jest dobrze zaprojektowany, to wyjściowa sekwencja bitów indeksów pozycji w słowniku jest mniejsza niż wejściowa sekwencja bitów kompresowanych, i na jej podstawie można dokładnie odtworzyć oryginalny sygnał [5].

W przypadku entropijnego kodera Huffmana wysyłanym symbolom (liczbom, bitom) przyporządkowuje się binarne słowa kodowe o różnej liczbie bitów tym mniejszej, im większe jest prawdopodobieństwo wystąpienia danego symbolu w kompresowanych danych [7].

Metodę Huffmana otrzymuje się poprzez dodanie wymogu do powyższych obserwacji, który mówi, że słowa kodowe dwóch symboli o najmniejszym prawdopodobieństwie wystąpienia różnią się jedynie ostatnim bitem. Czyli dla ς i δ, symboli o najmniejszych prawdopodobieństwach, słowa kodowe mają postać [3]:

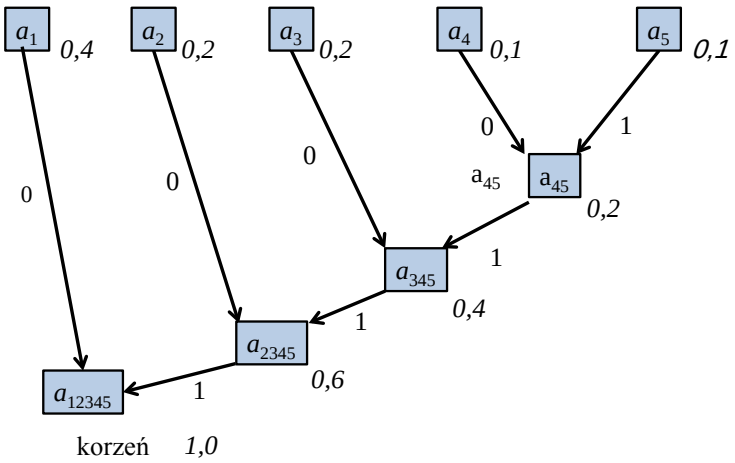
$$\varsigma = m * 0 \quad , \quad \delta = m * 1, \tag{3}$$

gdzie: *m* – ciąg złożony z zer i jedynek, * zero lub jeden binarne. W zależności od tego, czy ciąg skończył się jedynką czy zerem. W przypadku zakończenia ciągu m zerem, * będzie oznaczać zero, a w przypadku zakończenia jedynką, * oznacza jedynkę binarną.

Sposób konstruowania kodu Huffmana można przedstawić za pomocą drzewa binarnego. Symbolom w takim drzewie odpowiadają wierzchołki zewnętrzne, czyli liście. Przechodząc od korzenia do liści odpowiadających tym symbolom oraz dodając ‘0’ do słowa kodowego wtedy, gdy wybieramy niższą gałąź, a ‘1’ gdy wyższą, można uzyskać kod Huffmana dla poszczególnych symboli.

Drzewo binarne kodu Huffmana zaczyna się tworzyć od liści. Słowa kodowe dwóch najmniej prawdopodobnych symboli mają identyczne słowa kodowe. Różnią się one jedyne ostatnim bitem. A zatem ścieżki prowadzące od korzenia do liści, które odpowiadają tym symbolom są takie same z wyjątkiem ostatniej gałęzi. A to oznacza, że symbole te są potomkami tego samego wierzchołka- rodzic. Od tej pory wierzchołek – rodzic staje się reprezentantem nowego zredukowanego alfabetu. Prawdopodobieństwo tego symbolu jest sumą prawdopodobieństw jego potomków. Następnie należy posortować wierzchołki odpowiadające symbolom zredukowanego alfabetu według prawdopodobieństw ich wystąpień i stosując tą samą regułę utworzyć wierzchołek – rodzica. Postępując w ten sposób dochodzi się do etapu, gdy zostanie jeden wierzchołek – korzeń drzewa. Aby uzyskać kod konkretnego symbolu, trawersuje się drzewo od korzenia do liścia odpowiadającego temu symbolowi, przy czym ‘0’ – odpowiada gałęzi niższej, a ‘1’ – wyższej.

Schemat tworzenia drzewa Huffmana przedstawia rys. 5.7. Zaznaczono prawdopodobieństwa wystąpienia poszczególnych symboli.



Rys. 5.7. Binarne drzewo Huffmana

Postać kodu Huffmana dla ciągu przedstawia tabela 5.1.

Tabela 5.1.

Prawdopodobieństwo	Symbol	Kod symbolu
0,4	a_1	0
0,2	a_2	10
0,2	a_3	110
0,1	a_4	1110
0,1	a_5	1111

Sekwencje kodowe przyporządkowane poszczególnym symbolom ciągu, potwierdzają słuszność obserwacji pierwszej, dotyczącej optymalnych kodów prefiksowych. Tak, więc najczęściej występującemu symbolowi ciągu przyporządkowane zostało jednobitowe słowo kodowe, podczas gdy najrzadziej występującym czterobitowe słowo kodowe.

Średnia długość tego kodu wynosi:

$$l = 0{,}4 \times 1 + 0{,}2 \times 2 + 0{,}2 \times 3 + 0{,}1 \times 4 + 0{,}1 \times 4 = 2{,}2 \text{bita/symbol}$$

W tym przypadku redundancja wynosi 0,078 bita/symbol.

Proces dekodowania polega na pobraniu ze zbioru danych zakodowanych informacji na temat prawdopodobieństwa występowania w zbiorze poszczególnych symboli ciągu, zbudowaniu drzewa binarnego analogicznie jak w procesie kodowania, a następnie przejść do pobierania kolejnych bitów, organizowania ich w słowa kodowe i przeszukiwania drzewa w celu znalezienie kolejnych liści- dekodowanych symboli ciągu.

Proces dekodowania przebiega w następujący sposób. Należy pobrać ze zbioru danych zakodowanych informacji na temat prawdopodobieństwa występowania w zbiorze poszczególnych symboli ciągu, zbudować drzewo binarne analogicznie jak w procesie kodowania, a następnie przejść do pobierania kolejnych bitów, organizowania ich w słowa kodowe i przeszukiwania drzewa w celu znalezienie kolejnych liści- dekodowanych symboli ciągu. [5,6].

5.3.2. Kodowanie słownikowe

Metoda słownikowa to technika kompresji, która w celu zwiększenia stopnia kompresji wykorzystuje cechy strukturalne danych. Algorytmy słownikowe polegają na czytaniu kolejnych sekwencji danych wejściowych i przeglądaniu słownika w celu znalezienia takiej samej sekwencji danych. Gdy przeszukiwanie zakończy się

sukcesem, indeks właściwej pozycji słownika staje się reprezentacją wyjściową. Im dłuższa sekwencja i krótszy indeks, tym kompresja jest efektywniejsza.

W kodowaniu słownikowym można stosować statyczny model słownika budowanego w trakcie etapu analizy danych. Jest on skuteczny, ponieważ zna on z wyprzedzeniem pojawiające się ciągi danych. Praktyczną przydatność modelu statycznego ogranicza problem związany z koniecznością dopisania do zbioru danych skompresowanych słownika użytego przy kodowaniu. Algorytmy przeznaczone do ogólnego użytku są w większości adaptacyjne.

W wielu zastosowaniach ciągi generowane przez źródło danych zawierają bardzo często powtarzające się wzorce. Prostym przykładem niech będą dane tekstowe, gdzie niektóre wzorce lub też całe słowa ciągle się powtarzają. Logicznym sposobem kodowania tego typu źródeł jest przechowywanie listy często występujących wzorców w słowniku. Kodowanie wystąpień wzorców w kodowanym ciągu odbywa się za pomocą odwołań do słownika. A gdy wzorzec nie istnieje w słowniku, można go zakodować inną mniej efektywną metodą.

Dane dzielone są na dwie klasy: wzorce często występujące i wzorce występujące rzadko. O skuteczności techniki tej świadczy znacznie mniejszy słownik od liczby wszystkich możliwych wzorców.

Kompresja metodą słownikową ma dać dużo lepszy efekt niż kompresja oparta na założeniu, że wszystkie wzorce są jednakowo prawdopodobne. Aby dokonać dobrego wyboru wzorców, konieczna jest duża wiedza o strukturze danych generowanych przez źródło. Jeśli takiej wiedzy nie ma przed rozpoczęciem kodowania, należy ją uzyskać podczas kodowania. Gdy posiadana wiedza o danych generowanych jest wystarczająca, można zastosować podejście statyczne, w przeciwnym razie należy wykorzystać podejście dynamiczne (adaptacyjne).

W przypadku podejścia statycznego, pierwszy etap odpowiadający fazie modelowania polega na zbudowaniu słownika, którego poszczególne frazy wypełniają pojedyncze symbole alfabetu lub też sekwencje kolejnych symboli w oparciu o wstępną analizę zbioru danych kompresowanych. Następnie słownik ten musi być zapisany lub przesłany do dekodera. Kodowanie strumienia danych polega na znajdowaniu odpowiednich fraz w słowniku, które odpowiadają sekwencjom danych wejściowych oraz zapamiętaniu indeksów tych fraz.

Metoda, która wykorzystuje słownik statyczny jest nazywana kodowaniem digramowym. Jest to metoda mało zależna od konkretnego zastosowania. Kodowanie digramowe jest często stosowaną metodą kodowania ze słownikiem statycznym. W takim przypadku, słownik składa się ze wszystkich liter alfabetu wejściowego i z tylu par liter, które nazwane są digramami ile można w nim zmieścić.

Wśród najbardziej znane metod słownikowe należy wymienić metody wykorzystujące słowniki dynamiczne: LZ77, LZ78, LZW oraz ich modyfikacje. W algorytmie LZ77 (nazwanego tak od nazwisk jego twórców – Abrahama Lempela i Jacoba Ziv-a). Słownikiem w tej metodzie jest zbiór danych poprzedzających bezpośrednio w strumieniu wejściowym kodowany symbol lub sekwencję symboli. Ciąg wejściowy jest przeglądany poprzez przesuwające się okno nałożone na strumień danych kodowanych okno podzielone jest na dwie części:

- bufor słownikowy, czyli słownik właściwy, który zawiera ostatnią zakodowaną część ciągu, bezpośrednio poprzedzający przeznaczoną do zakodowania sekwencję danych,
- bufor kodowania, (ang. *look-ahead buffer*) zawierający fragment ciągu danych przeznaczonych do zakodowania.

Algorytm kodowanie w metodzie LZ77 można przedstawić następująco:

1. Strukturę okna należy ustawić na początku strumienia danych wejściowych.
2. W buforze słownikowym szukany jest najdłuższy łańcuch danych (zaczynając od danej znajdującej się na ostatniej pozycji w buforze), który ma swój dokładny odpowiednik w buforze „patrz w przód".
3. Tworzona sekwencja kodowa tego łańcucha zawiera wskaźnik jego położenia w słowniku (przesunięcie), długość oraz pierwszy symbol w buforze występujący bezpośrednio po kodowanej frazie.
4. Okno przesuwane jest wzdłuż wejściowego strumienia danych o długości zakodowanej frazy.
5. Czynności z pp. 2, 3 i 4 powtarzane są do momentu, gdy zostanie zakodowany ostatni symbol ze zbioru danych wejściowych

Metoda LZ-78 to metoda, która nie używa żadnego bufora szukającego. Zamiast okna przesuwnego wzdłuż strumienia danych wejściowych jest słownik, który budowany jest jako nieprzesuwna struktura. Początkowo słownik jest pusty, a jego rozmiar jest jedynie ograniczony przez ilość dostępnej pamięci. Dane wejściowe kodowane są za pomocą dwóch liczb, tworzących parę (i, c), przy czym:

i – jest indeksem odpowiadającym elementowi słownika, dający najdłuższe dopasowanie najdłuższego ciągu,

c – jest kodem symbolu w danych wejściowych, znajdującego się bezpośrednio za dopasowanym fragmentem [5].

Podobnie jak w algorytmie LZ77 indeks o wartości 0 odpowiada przypadkowi, w którym nie ma żadnego dopasowania. W metodzie tej ze słownika nic nigdy nie

będzie usunięte. Słownik dąży do szybkiego rozrostu i do wypełnienia całej pamięci która jest mu przydzielona.

W przypadku algorytm LZ78 słownik jest potencjalnie nieograniczonym zestawem fraz już kodowanych, uzupełnionych zbiorem fraz dodatkowych i dobranych apriorycznie. Kod łańcucha symboli składa się z pojedynczego symbolu, jaki występuje bezpośrednio po tym łańcuchu oraz z pozycji takiej samej frazy w słowniku. Modyfikacja słownika przebiega na bieżąco; frazy ze słownika, które kolejno odnajdywane są w trakcie kodowania, jako odpowiadające sekwencjom wejściowym, wchodzą w skład nowych fraz uzupełnione o symbol, wchodzący w skład kodu łańcucha. Nowe frazy umieszczane na kolejnych pozycjach słownika przyczyniając się do dynamicznej rozbudowy słownika.

Kodowanie długości sekwencji

Kodowanie długości sekwencji (ang. RLE – Run-Length Encoding) jest metoda kompresji bezstratnej, polegająca na zamianie łańcuchów złożonych z tego samego symbolu przez parę (licznik powtórzeń, symbol). Koncepcja tej metody dla danych obrazowych opiera się na zauważeniu, że jeśli dane nie są całkiem chaotyczne, to istnieje spora szansa, że piksel danego koloru będzie się powtarzał kilkakrotnie pod rząd. Wystarczy, zatem w ciąg pikseli dwa rodzaje znaczników, które oznaczają odpowiednio: ilość powtórzeń najbliższego piksela oraz informacje, że kolejną ilość pikseli należy traktować dosłownie.

Znacznikami są liczby, na ogół z tego samego zakresu co kolory, czyli np. (0,255). Należy założyć, że liczby, które są z zakresu $0 \le z \le 127$ oznaczają ilość powtórzeń z+1 razy piksela występującego po znaczniku, natomiast liczby z zakresu $128 \le z \le 256$ oznaczają, że kolejne z–127 pikseli nie zostało zakodowanych. Należy także założyć, że pierwszym elementem ciągu wynikowego jest znacznik, co wystarcza do jednoznacznej interpretacji zakodowanego ciągu.

Dla przykładu niech będzie ciąg pikseli obrazu o kolorach w skali szarości, który ma postać:

$$0, 0, 0, 0, 255, 255, 255, 0, 63, 0, 127, 127, 127, 127, 127, 127, 127$$

Można go zakodować w następujący sposób:

$$\overline{3}, 0, \overline{2}, 255, \overline{130}, 0, 63, 0, \overline{6}, 127\ ,$$

przy czym liczby z poziomą kreską u góry są znacznikami. Pierwsza wartość wskaźnika oznacza, że wartość 0 jest powtórzona czterokrotnie, (tzn. 3+1), druga, że wartość 255 będzie powtórzona trzy razy, itd. W ten prosty sposób długość ciągu została skrócona o 7 bajtów.

Przedstawiony sposób kodowania jest bardzo efektywny szczególnie dla obrazów zawierających zeskanowany tekst w których dominuje białe tło. Stosowany jest w formatach zapisu obrazów jak np. PCX, BMP, TGA. Jednak w przypadku, kiedy w kodowanej obrazie nie występują obok siebie piksele o jednakowych wartościach, długość zakodowanej sekwencji będzie dwa razy większa niż sekwencji oryginalnej. W takim przypadku konieczne jest sprawdzanie kodowanego ciągu na okoliczność wystąpienia sytuacji, kiedy nie występuje w nim ciąg jednakowych wartości. W zmodyfikowanej wersji opisywanego algorytmu, tylko ciągi dłuższe niż wartość progowa (zazwyczaj = 2) są kodowane i zamieniane na pary (licznik powtórzeń, symbol), której wystąpienie w sekwencji wyjściowej sygnalizowane jest specjalnym, zarezerwowanym znakiem.
Dla przykładu, jeżeli kodowana sekwencja będzie miała postać:

10 10 13 100 100 100 100 100 111 211 211 211 255 255,

wówczas ciąg wyjściowy będzie miał postać:

10 10 13 $ 5 100 111 $ 3 211 255 255,

przy czym znak $ sygnalizuje wystąpienie pary (licznik powtórzeń, symbol).

Kodowanie RLE dane obrazowe traktuje jako dane jednowymiarowe. Kodowanie może odbywać się wzdłuż linii poziomych (wierszy) lub wzdłuż linii pionowych (kolumn), zaczynając od piksela znajdującego się w lewym górnym rogu obrazu oraz w sposób zygzakowaty.Istnieje ponadto dwuwymiarowa wersja algorytmu RLE, która uwzględnia wzajemne relacje sąsiedztwa pikseli na płaszczyźnie obrazu. Oprócz sąsiedztwa symboli kodowanych kolejno wzdłuż wierszy uwzględnia się sąsiedztwo z góry, oraz góra-skos.

Niech będzie dany fragment obrazu w odcieniach szarości, w którym jasności poszczególnych pikseli są następujące (rys. 5.8):

85	90	100	100	100
85	100	100	100	100
85	85	100	100	100
85	120	85	100	100
85	120	85	100	100

Rys. 5.8. Fragment obrazu z danymi jasnościami pikseli

Reprezentacja pośrednia (wynik modelowania), przy analizowaniu obrazu poczynając od lewego górnego rogu, będzie miała postać:

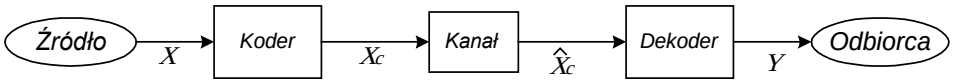
1w85 1w90 3w100 1g 4w100 2w85 3g 1g 1w120 3gl 5g,

przy czym „w" oznacza sąsiedztwo wzdłuż wiersza, „g" oznacza sąsiedztwo góra-dół, natomiast „gl" oznacza sąsiedztwo góra-skos.

5.4. Kompresja stratna

Kompresję stratną (nieodwracalną) stosuje się w przypadku rzeczywistych sygnałów multimedialnych, czyli mowy, muzyki, obrazów oraz sekwencji cyfrowego sygnału telewizyjnego. W procesie dekompresji nie otrzymuje się tej samej sekwencji bitów, co w sygnale oryginalnym. Jego dolnoczęstotliwościowa aproksymacja jest pozbawiona wysokoczęstotliwościowych szczegółów. Im sygnał zostaje pozbawiony mniejszej ilości detali w procesie kompresji, tym uzyskany zostaje mniejszy stopień kompresji, lecz zrekonstruowany sygnał bliższy będzie oryginałowi.

Na rysunku 8 zawarty jest schemat kompresji stratnej [5]. Informacje generowane przez źródło opisuje zmienna losowa X. Koder przekształcane dane wejściowe na ich wersję skompresowaną XC. Blok Kanał symbolizuje wszelkie przekształcenia, którym poddawane są dane skompresowane przed rekonstrukcją. Kanał oznacza przekształcenia wzajemnie równoważne $X_C = \hat{X}_C$. Na podstawie wersji skompresowanej $\hat{X}_C$ dokonuje się rekonstrukcji danych po stronie odbiorcy [5,6].



Rys. 5.9. Ogólny schemat kompresji stratnej

Proces eliminowania wysokoczęstotliwościowych składników obrazu może być realizowany w różny sposób [5,7]:

- drogą śledzenia i upraszczania zmian dynamicznych np. metoda ADPCM (ang.Adaptive Differential Pulse Code Modulation),
- drogą przekształcania sygnału do dziedziny częstotliwościowej i usunięciu współczynników widmowych niskoczęstotliwościowych np. standard

kompresji nieruchomych obrazów JPEG, gdzie stosuje się dwuwymiarową dyskretną transformatę kosinusową 2D-DCT,

- drogą modelowania w sposób uproszczony źródeł generacji sygnałów i wyznaczania parametrów tych modeli na podstawie analizowanych sygnałów
- drogą eliminacji z sygnału składowych, których brak jest nie zauważalny np. w MPEG (ang. Moving Pictures Exprets Group) audio gdzie kompresowany sygnał rozkładany jest na wiele sygnałów podpasmowych (wysokoczęstotliwościowych).

5.5. Efektywność kompresji

Efektywność kompresji jest rozumiana na wiele sposobów w zależności od rodzaju kompresowanych danych, zastosowania lub sprzętowych możliwości implementacji. Najbardziej powszechnym rozumieniem tego pojęcia jest zdolność do maksymalnego zredukowania rozmiaru nowej reprezentacji kompresowanego sygnału. Do ilościowych miar tak rozumianej efektywności kompresji należą [5,6]:

- stopień kompresji *CR*,
- procent kompresji *CP*,
- średnia bitowa *BR*,
- energia i entropia sygnału.

Stopień kompresji CR

Stopień kompresji *CR* (ang. *Compression Ratio*) wyraża się jako stosunek liczby bitów reprezentacji sygnału oryginalnego do liczby bitów reprezentacji sygnału skompresowanego.

$$CR = \frac{a}{b}, \tag{4}$$

gdzie:

a – liczba bitów reprezentacji sygnału oryginalnego,

b – liczba bitów reprezentacji sygnału skompresowanego.

Procent kompresji CP

Procent kompresji CP (ang. Compression Percentage) określa wyrażenie:

$$CP = (1 - \frac{1}{CR}) \cdot 100\% \tag{5}$$

Średnia bitowa BR

Średnia bitowa BR (ang. Bite Rate) definiuje się jako średnia ilość bitów skompresowanej reprezentacji sygnału przypadającej na element oryginalnego sygnału.

Efektywność (skuteczność) metod kompresji oznacza najczęściej dążenie do osiągnięcia możliwie dużych wartości CR i CP, oraz możliwie małej średniej bitowej BR [6].

Energia sygnału

Energia sygnału jest wyrażana wzorem

$$E_x = \sum_n x^2(n), \tag{6}$$

gdzie:

n – ilość próbek sygnału,

x – dany sygnał.

Entropia sygnału

Jeżeli założy się, że *X* jest zmienna losową przyjmującą wartości z ciągu źródłowego

$$X = \{x_0, x_1, ..., X_{N-1}\},$$

Y jest zmienną losową przyjmującą wartości z ciągu rekonstrukcji

$$Y = \{y_0, y_1, ... , y_{M-1}\}$$

Entropię źródła i rekonstrukcji wyraża się zależnościami odpowiednio [1]:

$$H(X) = -\sum_{i=0}^{N-1} P(x_i) \log_2 P(x_i) \tag{7}$$

$$H(Y) = -\sum_{j=0}^{M-1} P(y_j) \log_2 P(y_j) \tag{8}$$

gdzie:

$P(x_i)$ – prawdopodobieństwo wystąpienia danej wartości próbki w sygnale wejściowym,

$P(y_i)$ – prawdopodobieństwo wystąpienia danej wartości próbki w sygnale wyjściowym.

Miary kompresji wykorzystujące różnice pomiędzy wartościami oryginalnymi i zrekonstruowanymi tzn. zniekształceń powstałych w trakcie kompresji:

- kwadratowa miara błędu SEM oraz bezwzględna miara błędu ADM,

- błąd średniokwadratowy MSE,
- stosunek szumu do sygnału SNR,

Kwadratowa miara błędu SEM oraz bezwzględna miara błędu ADM

Najbardziej rozpowszechnioną miarą zniekształceń jest kwadratowa miara błędu SEM(ang. Squared Error Measure) oraz bezwzględna miara błędu ADM (ang. Absolute Difference Measure). Te dwie miary nazywane są inaczej różnicowymi miarami zniekształceń DDM (ang. Difference Distortion Measure) [5].

Jeżeli założy się, że XN oznacza ciąg oryginalny, a YN to ciąg danych powstałych w wyniku procesu dekompresji, wtedy kwadratowa miara błędu dana jest następującą zależnością:

$$d(x_i, y_j) = (x_i - y_j)^2, \tag{9}$$

gdzie

x_i – dowolnym elementem ciągu X_N,

y_j – dowolnym elementem ciągu Y_N,

natomiast bezwzględna miara błędu:

$$d(x_i, y_j) = \left| x_i - y_j \right| \tag{10}$$

Błąd średniokwadratowy MSE (ang. Mean Square Error)

W wielu przypadkach trudno jest wyciągnąć wnioski na temat różnic między sygnałem oryginalnym i rekonstrukcji na podstawie ciągu różnic poszczególnych bitów. W związku z tym do reprezentacji zniekształceń wykorzystuje się najczęściej sumy różnic pomiędzy elementami ciągów.

Najpopularniejsza miara tego typu jest średnia arytmetyczna z kwadratów błędów poszczególnych elementów ciągu – błąd średniokwadratowy – σ^2,

$$\sigma^2 = \frac{1}{N}\sum_{n=1}^{N}\left(x_n - y_n\right)^2 \tag{11}$$

Stosunek sygnału do szumu SNR

SNR (ang. Signal to Noise Ratio) odnosi się do wartości oryginalnych sygnału, jest to iloraz średniej wartości kwadratów danych i MSE.

$$SNR = \frac{\sigma_x^2}{\sigma^2}, \tag{12}$$

gdzie:

σ_x^2 - średnia wartość kwadratów danych oryginalnych,

σ^2 - *MSE*.

Wartość SNR jest często wyrażana w skali logarytmicznej, i dlatego używaną wówczas jednostką jest decybel (dB).

$$SNR(dB) = 10 \cdot \log_{10} \frac{\sigma_x^2}{\sigma^2} \tag{13}$$

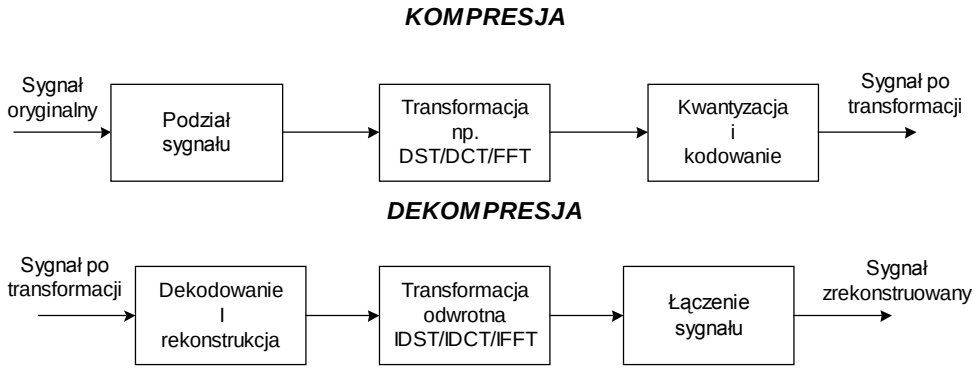
Kodowanie transformacyjne

Metoda kompresji stratnej zwana kodowaniem transformacyjnym TC (ang. Transform coding) jest najpopularniejszym sposobem kompresji danych. Znalazła zastosowanie w kompresji stratnej obrazów JPEG, MPEG oraz dźwięków (mp3 – MPEG I Layer III). Stosowanie metod transformacyjnych ma duże walory praktyczne ze względu na niewielkie koszty obliczeniowe oraz łatwość realizacji koderów, zarówno programowych jak i sprzętowych.

Ogólny schemat kodowania transformacyjnego jest techniką, w której sygnał dzielony jest na pewną liczbę składowych, gdzie każda składowa jest kodowana niezależnie. Celem transformacji jest usunięcie korelacji występującymi pomiędzy próbkami sygnału. Dekorelację osiąga się zasadniczo poprzez fakt, iż po transformacji energia sygnału zawarta jest w niewielkiej liczbie współczynników transformaty. Umożliwia to efektywną kompresję wskutek usunięcia wielu wartości pozostałych współczynników procesie kwantyzacji. Odpowiedni dobór współczynników kwantyzacji pozwala przy zastosowaniu technik transformacyjnych na osiągnięcie metod kodowania stratnego i bezstratnego. Na rysunku 5.10 został zamieszczony ogólny schemat blokowy kodowania transformacyjnego.

W technikach kodowania transformacyjnego, obok efektywnego przeprowadzenia transformacji na sygnale oryginalnym istotny jest również proces kwantyzacji współczynników transformaty, które są kodowane. W większości przypadków stosowane są schematy kwantyzacji skalarnej. Nierzadko używane są techniki przydziału bitów (ang. Bit allocation), które ustalają liczbę poziomów kwantyzacji dla poszczególnych współczynników. Z ilości poziomów kwantyzacji wynika liczba bitów do zapisu tychże poziomów. . Idea polega na wierniejszym opisie współczynników przenoszących znaczącą informację, a eliminacja danych

małoznaczących poprzez zerowy przydział bitów. Istnieją dwa zasadnicze sposoby kwantyzacji skalarnej współczynników transformat.



Rys. 5.10. Schemat kodowania transformacyjnego

Pierwszy z nich wykorzystuje selekcję próbek, zachowując wartości współczynników tylko w określonym przedziale transformowanego sygnału, poza którym wszystkie pozostałe współczynniki będą zerowane, bez względu na ich wartość. W przypadku sygnału mowy oraz audio, taki model pozwala na usunięcie współczynników o wybranej częstotliwości, nie naruszając pozostałych.

Drugim schematem kwantyzacji skalarnej jest progowa selekcja próbek. Możliwe jest wyeliminowanie najbardziej wyróżniającego się fragmentu sygnału. Możliwe jest pominięcie wartości współczynnika zawierającego znaczną energię, co może powodować duży błąd rekonstrukcji. Dobierany jest poziom progu, powyżej którego następuje kwantyzacja i kodowanie współczynników. W tym momencie pojawia się problem pamiętania pozycji współczynnika, którego wartość przekracza wartość progu. Do zakodowania pozycji używa się technik kodowania długości sekwencji. Struktura kwantyzatora jest skonstruowana jako funkcja położenia poszczególnych współczynników w sygnale

5.6. Dyskretne przekształcenie kosinusowe DCT

Dyskretne przekształcenie kosinusowe DCT (ang. Discrete Cosine Transform) jest typem przekształcenia częstotliwościowego. W tego typu metodzie liczba bitów używana do kodowania poszczególnych składowych może być niezależna od liczby bitów dla pozostałych składowych. Dokładność kodowania może być dostosowywana do aktualnych potrzeb dla danej składowej częstotliwościowej. W poszczególnych przypadkach składowa częstotliwościowa o małej lub zerowej energii może nie być wcale kodowana.

DCT jest dyskretnym ciągiem $x(n) = 0,1,2,....,(N-1)$ i jest wyrażane następującą zależnością [5]:

$$X_{\cos}(k) = \sqrt{\frac{2}{N}} C(k) \sum_{k=0}^{N-1} x(n) \cdot \cos \frac{(2n+1)k\pi}{2N} \tag{14}$$

$k = 0,1,2,...,(N-1)$,

$C(k) = \frac{1}{\sqrt{2}}$ dla *k=0*,

$C(k) = 1$ dla $k = 0,1,2,...,(N-1)$.

gdzie:

n – numer próbki,

k – numer bazy.

Podobnie transformata odwrotna do DCT, IDCT (ang. Inverse Discrete Cosine Transform) wyraża się równaniem [5]:

$$x(n) = \sqrt{\frac{2}{N}} \sum_{k=0}^{N-1} C(k) \cdot X(k) \cdot \cos \frac{(2n+1)k\pi}{2N} \tag{15}$$

Za pomocą IDCT można zrekonstruować sygnał, wcześniej przetransformowany metodą DCT. Zostaje on odtworzony na podstawie zakodowanych współczynników.

W przypadku, gdy przetwarzane są dane obrazowe, należy zastosować dwuwymiarową transformata kosinusowa (2D DCT). Jeżeli, założy się, że macierz **X**, składająca się z *n* wierszy i kolumn reprezentuje obraz, wówczas macierz zawierająca współczynniki transformaty można zapisać jako:

$$\mathbf{Y} = \mathbf{A}\mathbf{X}\mathbf{A}^{\mathbf{T}}, \tag{16}$$

przy czym A jest macierzą, której współczynniki określone są jako:

$$A_{i,j} = C_i \cos \frac{(2j+1)i\pi}{2N}. \tag{17}$$

W praktyce obliczenie 2D DCT polega na obliczeniu najpierw jednowymiarowej DCT dla każdego rzędu w danym bloku (w przypadku JPEG blok ma rozmiar 8×8

pikseli), a następnie dla każdej kolumny w bloku, co jest równoważne pierwszej operacji na bloku przetransponowanym.

W technikach transformacyjnych, współczynnikom posiadającym bardziej istotną informację można przydzielić większą ilość bitów. Najczęściej są to składowe niskoczęstotliwościowe. W ten sposób, można zredukować liczbą bitów wymaganych do przesyłania sygnału i tym samym ograniczyć pasmo sygnału. Redukcja szerokości pasma jest podstawowym celem większości technik kompresji.

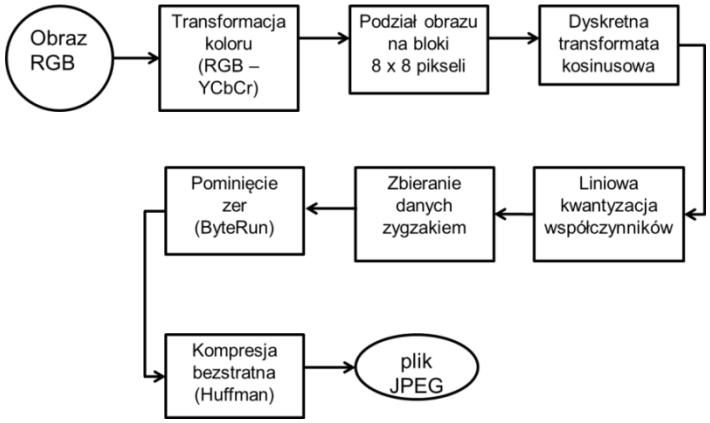
5.7. JPEG

Standard JPEG jest jednym z najbardziej znanych standardów stratnej kompresji obrazów. Nazwa pochodzi się od Joint Photographic Experts Group, która w późnych latach osiemdziesiątych opracowała wspomniany standard kompresji. JPEG z założenia miał być stosowany w odniesieniu do obrazów naturalnych, które charakteryzują się łagodnymi przejściami tonalnymi. Standard opisuje metodą stratną, wykorzystującą kodowanie transformatowe, a konkretnie dyskretną transformatę kosinusową (ang. DCT – Discrete Cosine Transform) jak również bezstratną, która ze względu na niską wydajność nie jest obecnie stosowana. Standard JPEG obsługuje obrazy w odcieniach szarości (8 bitów/piksel) oraz kolorowe (24 bity/piksel). Wykorzystuje niedoskonałości zmysłu wzroku związaną [5]:

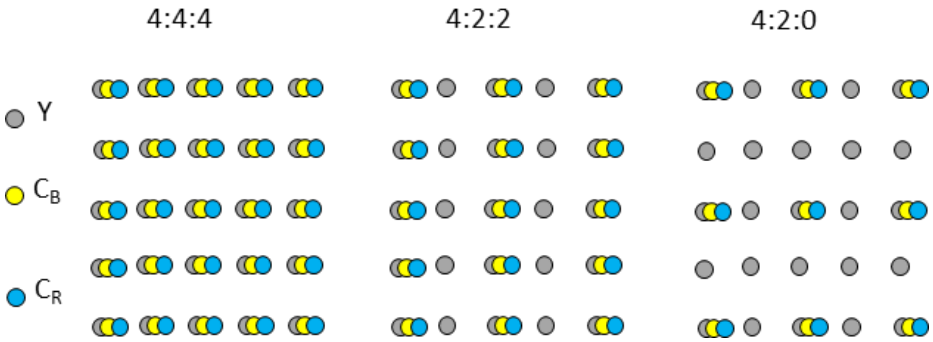
- z większą wrażliwością na zmianę poziomu jasności,
- z mniejszą wrażliwością na przestrzenną zmianę kolorów niż zmianę jasności.

Standard jest dostosowany do transmisji danych przez sieć umożliwiając wyświetlanie progresywne. Jego przewaga nad wyświetlaniem sekwencyjnym polega na tym, że wyświetlany jest od razu cały obraz, a w miarę odbierania kolejnych porcji danych polepsza się jakość wyświetlanego obrazu. Ponadto w pliku jpg przewidziano możliwość zapisu stosowanego profilu kolorów, co pozwala na wierniejszą reprodukcję barw w zależności od stosowanego urządzenia.
Schemat kompresji JPEG przebiega według następującego porządku, przedstawionego na rys. 5.11.

Konwersja z przestrzeni RGB na przestrzeń YCBCR ze względu na wspomnianą wcześniej różną wrażliwość na zmiany jasności niż barwy. Składowe chrominancji mogą być dodatkowo próbkowane przestrzennie. Polega to na pominięciu składowych chrominancji w co drugiej kolumnie (format 4:2:2) lub w co drugiej kolumnie i co drugim wierszu przetwarzanego obrazu (format 4:2:0), jak pokazano na rys. 5.12.



Rys. 5.11. Etapy kompresji JPEG

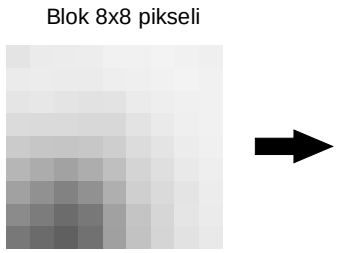


Rys. 5.12. Warianty próbkowania przestrzennego składowych chrominancji

Każda ze składowych Y, CB,CR jest następnie dzielona na bloki o rozmiarach 8×8 pikseli. Jeżeli składowe chrominancji były przeskalowywane (formaty 4:2:2, 4:2:0), wówczas bloki tych składowych obejmują obszar odpowiednio: 16×8 lub 16×16 pikseli oryginalnego obrazu. Gwarantuje to stały rozmiar 64-pikselowego bloku, który jest w dalszej kolejności przetwarzany jest za pomocą dyskretnej transformaty kosinusowej, osobno dla każdej ze składowych luminancji i chrominancji. Problem powstaje, gdy wysokość obrazu nie jest całkowitą wielokrotnością 8. Wówczas skrajne bloki należy dopełnić do pełnego rozmiaru, najlepiej pikselami o takich wartościach, aby odwrotna transformata kosinusowa odtworzyła piksele o wartościach najbardziej zbliżonych do oryginalnych.

Przekształcenie każdego bloku o rozmiarze 8×8 pikseli zawierającego 8-bitową informację o każdej ze składowych daje w efekcie macierz o takim samym rozmiarze zawierającą współczynniki transformaty z 16-bitową rozdzielczością, rys. 5.13. Kolejny etap działania algorytmu polega na liniowej kwantyzacji tych

współczynników. Rozmiary kroków kwantyzacji zapamiętane w specjalnej tablicy i zwiększają się dla współczynników transformaty odpowiadającym wyższym częstotliwościom przestrzennym. Dzieje się tak dlatego, że błędy kwantyzacji dla składowych niskoczęstotliwościowych są łatwiejsze do dostrzeżenia niż dla składowych wysokoczęstotliwościowych. Ze względu na fakt, że wszystkie kwantyzatory są kwantyzatorami o zerowym poziomie wyjścia, działanie procesu kwantyzacji jest jak w przypadku funkcji progowej. Współczynnikom, których wartość bezwzględna jest mniejsza od wartości progowej, zostanie przypisane 0.



macierz amplitud pikseli

228	234	236	237	242	242	243	242	240
234	236	234	234	238	240	242	243	242
229	230	228	226	227	234	238	241	242
218	217	217	216	216	227	233	239	242
203	198	196	199	205	219	228	236	241
181	170	162	171	191	212	222	232	238
161	144	129	144	174	206	217	228	235
139	124	108	119	161	195	214	229	234
119	107	97	113	159	193	213	226	232

DCT

macierz współczynników DCT

1859.	-134.75	31.8733	30.5996	8.19617	-4.09477	-2.	0.980264	-1.32288
176.602	74.3742	-16.7065	-20.3562	-8.51581	1.4291	1.11894	-0.606004	0.147753
-29.8388	-17.8486	-3.68666	5.81439	1.48384	-1.27217	-2.99665	-0.267235	1.70411
-1.73205	-0.603602	-0.808908	2.41667	2.62326	0.174606	-0.144338	-0.278208	-0.54542
0.527027	-0.866268	-3.48008	-3.548	-1.28582	0.417619	0.0799967	0.0449687	0.893174
-1.18519	0.300778	1.4709	0.0676461	-0.847544	-0.216912	0.194138	0.256792	-0.31455
-2.33333	-1.12606	-2.05917	-0.144338	0.329619	-0.173269	0.583333	-0.0867636	-0.38879
-1.19117	-0.255638	1.02543	-0.173847	-0.732452	-0.221893	-0.662912	0.176062	0.604974
0.30084	0.396179	-0.507337	-0.0195486	0.140702	0.0500919	0.506684	-0.262048	-0.36085

Rys. 5.13. Przykładowy blok pikseli i odpowiadająca mu macierz luminancji i współczynników DCT

Na rys. 5.13 można zauważyć, że współczynniki transformaty DCT o największych wartościach bezwzględnych znajdują się w lewym górnym rogu macierzy. Po przeprowadzonej kwantyzacji, wartości bezwzględne współczynników odczytywane zygzakiem, stają się coraz mniejsze. Ponieważ rozmiar kwantyzacji jest coraz większy, prawdopodobieństwo wystąpienia zer staje się coraz większe. Cały ciąg zer w może być w końcu zakodowany za pomocą odpowiedniego znacznika, który wskazuje na brak występowania niezerowych współczynników. Daje to istotne zmniejszenie objętości danych, a tym samym znaczną kompresję. Następnie, skwantowane współczynniki transformaty kosinusowej są dodatkowo kompresowane przy użyciu kodowania Huffmana z wykorzystaniem słownika dynamicznego lub predefiniowanego.

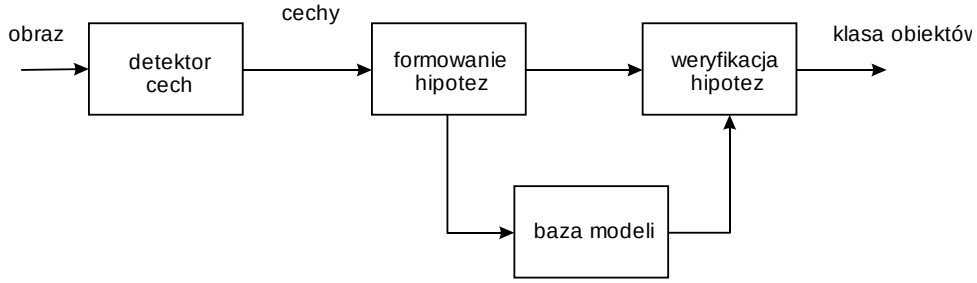
Jak to już zostało wspomniane, kompresja obrazu o łagodnych przejściach tonalnych za pomocą opisanego wcześniej algorytmu daje bardzo dobre rezultaty, w odróżnieniu do obrazów zawierające jednobarwne, kontrastowe obiekty. Dotyczy

to w szczególności obrazów zawierającego tekst. Inną niedoskonałością standardu JPEG jest wrażliwość na przekłamania danych, powstające np. podczas transmisji oraz brak obsługi obrazów wielkoformatowych (zdjęcia lotnicze, satelitarne), Wspomniane niedogodności skłoniły do opracowania nowszej wersji standardu – JPEG 2000, w którym zamiast DCT stosowana jest transformata falkowa.

5.8. Rozpoznawanie obrazu

Rozpoznawanie obrazu jest szczególnym przypadkiem bardziej ogólnego problemu jakim jest rozpoznawanie wzorców (ang. pattern recognition). W obecnej chwili jest szeroko stosowane w wielu dziedzinach. Typowymi przykładami zastosowania rozpoznawania obrazów jest rozpoznawanie pisma (OCR – ang. optical character recognition), systemy biometryczne rozpoznawania osób na podstawie twarzy, odcisków palców, systemy diagnostyki medycznej i przemysłowej, czy też zastosowania militarne, z których warto wymienić np. systemy automatycznego rozpoznawania celów (ATR – ang. Automated Target Recognition).

Rozpoznawanie obrazów ogólnie polega on na przyporządkowaniu etykiet obiektom lub zbiorom obiektów w obrazie. Rozpoznanie obiektu w obrazie polega na zaklasyfikowaniu go do pojedynczej i unikalnej klasy. Problem stanowi tutaj brak aprioryczne informacji na temat reguł przynależności do tych klas [8]. Innym problemem, szczególnie w przypadku rozpoznawania obrazów, jest przetwarzanie bardzo dużej ilości informacji, co wymaga użycia dużego nakładu mocy obliczeniowej i znacznych zasobów pamięci. Klasyfikacja jest więc formą uczenia się, przy czym ciąg uczący złożony jest z obiektów, dla których właściwa klasyfikacja jest znana. Rozpoznawanie natomiast można postrzegać jako pewną regułę wnioskowania [9]. System rozpoznawania wzorców przedstawiona rys. 5.14.



Rys. 5.14. Ogólny schemat procesu rozpoznawania obrazu

Zadanie rozpoznania wymaga następujących czynności [9]:

- określania liczby klas,
- określenia rozkładu a priori klas,

- określenia cech klas,
- określenia rozkładu cech w poszczególnych klasach,
- określenia stopnia gradacji cech i dyskretyzacji obrazu,
- minimalizacji liczby cech polegającej na wyborze takich, aby dokładność rozpoznawania nie uległa zmniejszeniu,
- końcowego ustalenia liczby cech.

Ze względu na sposób przetwarzania informacji zwartej w poszczególnych cechach systemy rozpoznawania można podzielić na:

- systemy rozpoznawania strukturalnego,
- systemy rozpoznawania metodami statystycznymi.

Wśród najczęściej stosowanych metod rozpoznawania obrazów można wyróżnić metody:

- minimalnoodległościowe,
- aproksymacyjne,
- statystyczne.

Dokładniejszy klasyfikacja metod rozpoznawania obrazu można znaleźć w [8]

Zadanie rozpoznania obrazu w prostym przypadku może zostać rozwiązane z użyciem metod deterministycznych, co wymaga podzielenia przestrzeni cech na części, które odpowiadają poszczególnym klasom. W przypadku, kiedy np. istnieje potrzeba przyporządkowania obiektów występujących do dwóch klasach, wówczas problem ten można rozwiązać poprzez indeksując wszystkie obiekty (X) i znajdując funkcję dyskryminacyjną d(X).

Niech $\Omega = \{\omega_1, \omega_2, \ldots \omega_m, \}$ oznacza zbiór klas, którego liczność będzie oznaczona jako *m*. Klasyfikacja może być rozpatrywana jako podział przestrzeni cech na *m* obszarów, odpowiadających klasie, takich, że [9]:

$$\bigwedge_{i \neq j} d_i(X) = d_j(X) \tag{18}$$

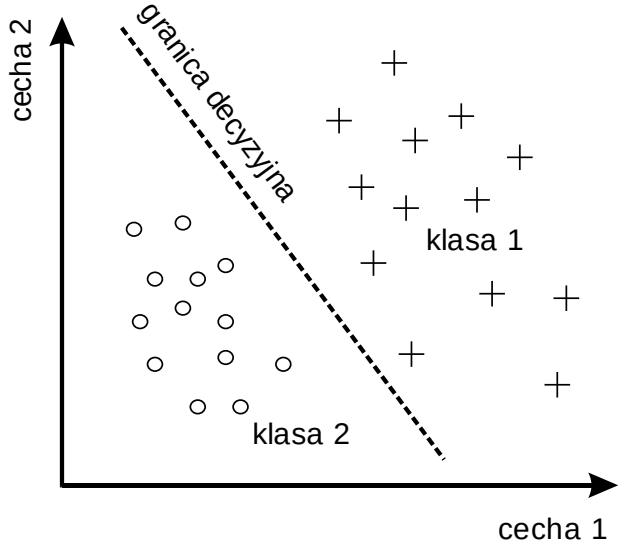
Gdy obiekt *X* należy do *i*-tej klasy, wówczas:

$$d_i(X) > d_j(X), \qquad i, j = 1,2 \ldots m \quad i \neq j \tag{19}$$

Funkcja dyskryminacyjna pomiędzy dwiema dowolnymi różnymi klasami i, j spełnia warunek:

$$d_{ij}(X) = d_i(X) - d_j(X) = 0, \tag{20}$$

Obiekt *X* należy do *i*-tej klasy, jeżeli $d_{ij}(X) > 0$, natomiast do j-tej klasy jeżeli $d_{ij}(X) > 0$. Funkcja dyskryminacyjna, określająca granicę decyzyjną może być liniowa jak na rys. 5.15, bądź też nieliniowa.



Rys. 5.15. Liniowa funkcji dyskryminacyjnej dla dwóch przykładowych klas

Metody minimalnoodległościowe opierają się na przesłankach związanych z geometrią przestrzeni cech. W przypadku, gdy punkty odpowiadające obiektom przybierają postać skupisk, można posłużyć się pojęciem odległości (metryki) przy podejmowaniu decyzji – rys. 14 . Metody te są często stosowane ze względu na prostotę, jak również dlatego, że charakteryzują się stosunkowo małym wskaźnikiem błędnych odwzorowań. Jednak ich stosowanie wymaga archiwizowania całego ciągu uczącego, oraz obliczania metryki rozpoznawanego obiekty od każdego elementu zbioru uczącego. Jest to zatem zadanie wymagające znacznych nakładów obliczeniowe, co implikuje jego czasochłonność.
Metryką nazywa się odwzorowanie określone na *q*-wymiarowej przestrzeni cech $\boldsymbol{R}^q$:

$$\rho : X \times X \to R^q \text{,(21)}$$

które spełnia następujące warunki:

$$\bigwedge_{i \neq j} \rho\left(X_i, X_j\right) = 0 \Leftrightarrow X_i \equiv X_j \tag{22}$$

$$\bigwedge_{i \neq j} \rho\left(X_i, X_j\right) = \rho\left(X_j, X_i\right) \tag{23}$$

$$\bigwedge_{i\neq j\neq k} \rho(X_i, X_j) < \rho(X_j, X_k) + \rho(X_i, X_k) \tag{24}$$

Metryka jest miarą podobieństwa obiektów. Im większa jest wartość metryki (większa odległość), tym są one do siebie mniej podobne

Najbardziej znaną metryką jest metryka euklidesowa, dana w postaci zależności:

$$\rho_e(X,Y) = \sqrt{\sum_{\nu=1}^{q} (X_\nu - Y_\nu)^2}\,, \tag{25}$$

Jest ona szczególnym przypadkiem uogólnionej metryki euklidesowe, którą definiuje się jako:

$$\rho_{eu}(X,Y) = \sqrt{\sum_{\nu=1}^{q} [\lambda_\nu (X_\nu - Y_\nu)]^2}\,, \tag{26}$$

gdzie λ oznacza stała normalizującą. Stosowanie metryki euklidesowej wymaga operacji podnoszenia do kwadratu i pierwiastkowania, stąd w przypadku dużej ilości obiektów, które należy zaklasyfikować może to oznaczać duży czas obliczeń, szczególnie gdy klasyfikowane obiekty charakteryzują duża liczbą cech (wymiarów). Rozwiązaniem jest stosowanie metryki miejskiej (taksówkowej, Manhattan), danej jako:

$$\rho_m(X,Y) = \sum_{\nu=1}^{q} |X_\nu - Y_\nu|\,. \tag{27}$$

Oprócz wymienionej wcześniej metryk stosowane są również metryki:

Czebyszewa

$$\rho_{czeb}(X,Y) = \max_{1\leq\nu\leq n} |X_\nu - Y_\nu|\,, \tag{28}$$

Minkowskiego:

$$\rho_{mink}(X,Y) = \sqrt[t]{\sum_{\nu=1}^{q} |X_\nu - Y_\nu|^t}\,, \tag{29}$$

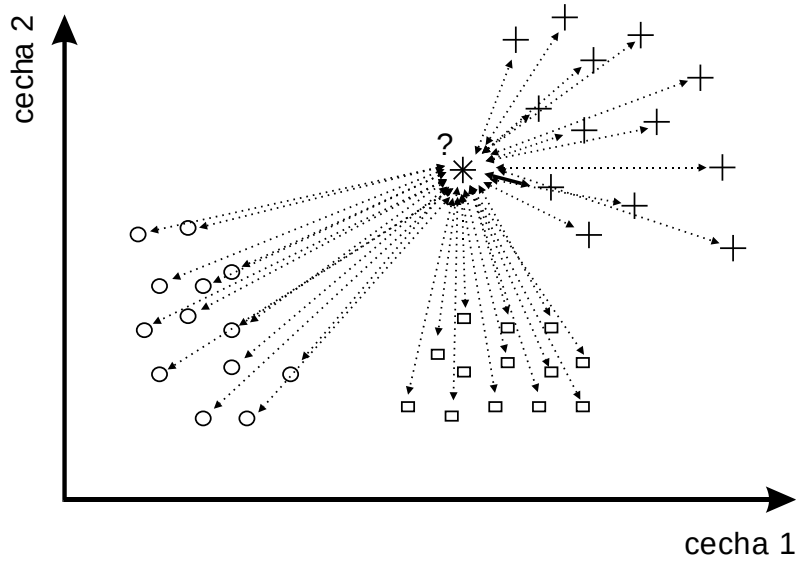
Mahalanobisa:

$$\rho_m(X,Y) = \sqrt{(X-Y)^T \mathbf{T}^{-1} (X-Y)}\,, \tag{30}$$

gdzie **T** oznacza macierz kowariancji *X* i *Y*.

Metoda k-najbliższych sąsiadów (k-NN – ang. k–*Nearest Neighbour*)

Metoda k-najbliższych sąsiadów polega na tym, że dany obiekt zostanie zaklasyfikowany do tej klasy, dla której należy obiekt ciągu uczącego położny najbliżej według założonej metryki w przestrzeni cech, jak przykładowo pokazano na rys. 5.16. W prezentowanym przykładzie istnieją 3 klasy obiektów, z których każdy charakteryzowany jest przez 2 cechy. Dla obiektu, który ma zostać zaklasyfikowany, obliczane są odległości, według założonej metryki, od wszystkich obiektów. Na rys. najmniejsza odległość zaznaczona została linią ciągła. Stąd też obiekt zostanie zaklasyfikowany do klasy, której elementy zostały oznaczone symbolem „+".



Rys. 5.16 Ilustracja metody k-NN

Jednak taki sposób postępowanie niesie za sobą pewne niebezpieczeństwo, kiedy jeden z obiektów zostanie błędnie zaklasyfikowany, np. wskutek zakłócenia. Pociągnie to za sobą kolejne błędne klasyfikacje. Z powyższego powodu stosowane są zmodyfikowane wersje prezentowanej metody: αNN oraz jkNN.

Metoda αNN wykorzystuje dodatkowy parametr α, którego wartość określa się z góry tak aby:

$$\alpha << \min_{i \in I} N^i , \tag{31}$$

przy czym *N* oznacza liczbę elementów ciągu uczącego, i – indeks klasy, I – zbiór indeksów klas. W praktyce wartość α jest w praktyce małą liczbą całkowitą [8].

Podobnie jak w przypadku podstawowej metody NN, obliczane są odległości pomiędzy obiektem, który ma zostać sklasyfikowany a wszystkimi elementami ciągu uczącego. Następnie, ciąg uczący zostaje uporządkowany według rosnących wartości odległości. Kolejnym krokiem jest wybranie α początkowych obiektów tak utworzonego ciągu, tworząc podzbiór, który rozbijany jest na kolejne podzbiory (mogą one być puste), które odpowiadają poszczególnym klasom. Funkcja przynależności wyznaczana jest na podstawie liczebności podzbiorów dla rozpatrywanych klas.

Literatura

1. Malina W., Smiatacz M.: Metody cyfrowego przetwarzania sygnałów, Akademicka Oficyna Wydawnicza EXIT, Warszawa 2008,
2. Pirga M., Kujawińska M., Modified procedure for automatic surface tomography,
3. Measurement, vol. 13, pp. 191-197, 1994
4. Woźnicki J., Podstawowe techniki przetwarzania obrazu, Warszawa, WKiŁ, 1996
5. Pastuszak W, Barwa w grafice komputerowej, PWN, Warszawa, 2000
6. Sayood K.: Kompresja danych - wprowadzenie, Wydawnictwo RM, Warszawa, 2002
7. Przelaskowski A.: Kompresja danych obrazowych, zarys zagadnień, Oficyna Wydawnicza Politechniki Warszawsiej, Warszawa, 2002
8. Cutler C.C.: Differential Quantizantion for Television Signals, U.S. Patent 2,605,361, July, 1952
9. Tadeusiewicz R., Flasiński M.: Rozpoznawanie obrazów, PWN, Warszawa, 1991
10. Choraś R.: Komputerowa wizja – metody interpretacji i identyfikacji obiektów, Akademicka Oficyna Wydawnicza EXIT, Warszawa, 2006

6. Administrowanie sieciami komputerowymi

6.1. Wstęp

Administracja siecią komputerową nie jest pojęciem jednoznacznym. W zależności od przedsiębiorstwa i rodzaju zasobów IT jakie są niezbędne do jego funkcjonowania może to oznaczać zupełnie inne czynności. Zarządzanie siecią musi być dopasowane do infrastruktury przedsiębiorstwa i do jego potrzeb biznesowych. W skład infrastruktury mogą wchodzić zasoby takie jak: urządzenia przełączające, serwery, urządzenia pamięci masowych, urządzenia dostępowe, okablowanie oraz stacje robocze. Oprócz administrowania samymi urządzeniami sieciowymi zakres pracy administratora może dotyczyć również systemów oraz aplikacji. Główną rolą sieci komputerowej jest świadczenie odpowiednich usług. W najwęższym znaczeniu administracja siecią oznacza administrację jedynie infrastrukturą niezbędną do przesyłu danych. Oznacza to w zasadzie urządzenia takie jak: routery, przełączniki oraz punkty dostępowe. Nieco szersze podejście dołącza do tego usługi sieciowe co oznacza administrację serwerami usług sieciowych. Kolejnym rozszerzeniem jest administracja aplikacjami. W takim przypadku najczęściej mamy do czynienia z systemami rozproszonymi. Najszerzej rozumiana administracja siecią do wymienionych wcześniej składników dołącza administrację hostami.

W związku z dynamicznym rozwojem technologii sieciowych dynamicznie zmienia się również zakres pojęciowy związany z zarządzaniem i administracją sieciami. Cechą charakterystyczną pracy administratora sieci jest ciągłe zdobywanie wiedzy oraz umiejętności.

Organizacja OSI (ang. Open Systems Interconnection) zdefiniowała pięć następujących obszarów zarządzania sieciami komputerowymi [1]:

- zarządzanie błędami i awariami,
- zarządzanie konfiguracją,
- zarządzanie i rejestrowanie wykorzystania zasobów,
- zarządzanie wydajnością,
- zarządzanie bezpieczeństwem.

Konfiguracja oznacza najczęściej proces zmiany określonych parametrów sieci lub poszczególny ch jej elementów. Często konfiguracja zaczyna się od instalacji określonego software lub też nawet fizycznej instalacji lub konfiguracji określonych połączeń pomiędzy elementami sieci. W niektórych przypadkach przed przystąpieniem do konfiguracji urządzenia należy się z nim połączyć a wcześniej wykryć. Konfiguracja może dotyczyć wybranego elementu urządzenia, całego

urządzenia, kilku urządzeń lub wręcz całej sieci. Wszystkie zmiany konfiguracyjne powinny być dokumentowane. Dane konfiguracyjne mogą być przesyłane i składowane w różny sposób.

Zarządzanie awariami może odbywać się w sposób aktywne lub w sposób wyprzedzający awarię. Typowo w przypadku pojawienia się awarii należy przede wszystkim ją zlokalizować a następnie odizolować od reszty sieci (systemu). Następnym krokiem jest powrót do normalnego stanu. Wykrycie i poprawienie błędnego działania wymaga właściwych technik monitorowania.

Zarządzanie wydajnością polega na działaniu mającym na celu osiągnięcie wymaganego i możliwego w danych warunkach poziomu usług. Określenie właściwego poziomu usług jest elementem umowy biznesowej pomiędzy usługodawcą a klientem lub też wymaganiem biznesowym w stosunku do infrastruktury dla usług sieciowych świadczonych w firmie przez specjalistów IT. Należy określić sposób pomiaru QoS oraz akceptowalny poziom poszczególnych parametrów. Sieć jest zbiorem elementów, z których każdy może w określony sposób wpływać na jakość usług. Słaba wydajność poszczególnych elementów może znacząco wpływać na wydajność sieci widzianej z perspektywy klienta. Z pomocą administratorowi przychodzą tutaj narzędzia monitorowania i analizy sieci, protokołów oraz urządzeń.

Rejestrowanie polega na śledzeniu wykorzystywania zasobów przez użytkowników. Do tego celu budowana jest infrastruktura, która oprócz rejestrowania posiada funkcję uwierzytelniania oraz autoryzacji. Typowymi rozwiązaniami są tutaj usługi takie jak RADIUS lub usługi katalogowe. Usługi takie służą do zarządzania użytkownikami, którzy korzystają z usług oraz zasobów.

Zarządzanie bezpieczeństwem opiera się o analizę zagrożeń całej sieci oraz poszczególnych jej zasobów. W wyniku analizy powinny powstać zasady bezpieczeństwa oraz procedury stosowane dla zwiększenia bezpieczeństwa.

6.2. Metody i technologie zarządzania siecią

System zarządzania może być scentralizowany i hierarchiczny lub też rozproszony. System rozproszony jest przeważnie szybszy od scentralizowanego. Dla dużych sieci centralne rozwiązanie dla całości zasobów może okazać się mało skuteczne. Może jednak sprawdzać się dla niektórych obszarów sieci.

Zarządzanie siecią, szczególnie w dużej organizacji, może podzielone na warstwy zgodnie z zakresem oddziaływania oraz biznesową hierarchią zarządzania. Kolejne warstwy od najwyższej związanej z potrzebami biznesowymi oraz biznesową polityką firmy do najniższej związanej z konfiguracją konkretnego urządzenia można rozumieć jako zakres oddziaływania warstwy na poszczególne elementy sieci.

Kolejne warstwy przedstawić można następująco:

- zarządzanie biznesowe – w tej warstwie zapadają decyzje o tym jakie usługi w jaki sposób będą świadczone, z wykorzystaniem jakich zasobów, jakie będą koszty oraz zyski, w jaki sposób zaplanować działanie, zmiany, strategie,
- zarządzanie usługami – najbliżej klienta, związek z końcowym użytkownikiem, któremu dostarczane są usługi świadczone przez firmę,
- zarządzanie siecią – dotyczy wszystkich elementów całej sieci,
- zarządzanie zbiorem elementów – dotyczy zbioru elementów pełniących podobną funkcję np. serwerami DNS,
- zarządzanie poszczególnymi elementami sieci – dotyczy pojedynczego urządzenia np. przełącznika sieciowego.

Każde urządzenie w sieci potrzebuje przynajmniej od czasu do czasu określonych zadań administracyjnych. Jeśli nawet urządzenie po podłączeniu nie wymaga nawet wstępnej konfiguracji to przydatne okazać się może monitorowanie jego pracy. Nawet jeśli polegało by to na sprawdzeniu czy urządzenie działa lub nie działa. Dla bardziej skomplikowanych urządzeń o wielu parametrach możliwych do konfiguracji określenie jego stanu nie jest już tak oczywiste. Dla złożonych urządzeń sieciowych dane konfiguracyjne mogą dotyczyć poszczególnych jego modułów fizycznych lub logicznych.

Konfiguracja urządzenia może być wykonana z wykorzystaniem kilku metod, z których najczęściej używane to:

- protokoły konfiguracji automatycznej takie jak DHCP,
- pliki konfiguracyjne,
- linia poleceń do konfiguracji,
- graficzny interfejs konfiguracji,
- właściwe techniki skryptów.

Każda z metod konfiguracji ma swoje wady i zalety. Protokoły konfiguracji automatycznej pozwalają w łatwy sposób skonfigurować wiele urządzeń w sposób automatyczny z jednego centralnego miejsca. Pliki konfiguracyjne pozwalają na ich kopiowanie, modyfikację i wysyłanie do nadzorowanego urządzenia. Linia poleceń w większości przypadków pozwala na najdokładniejszą i szczegółowa konfigurację na każdym poziomie. Największą jej wadą jest konieczność nauczenia się poleceń, które znacznie różnią się w zależności od producenta urządzenia. Niektóre urządzenia wymagają fizycznego połączenia do zarządzanego urządzenia specjalnym interfejsem służącym tylko do zarządzania. Inne pozwalają na połączenie za pomocą zdalnej konsoli. Interfejs graficzny jest najbardziej przyjazny dla administratora.

Każdy parametr, który podlega konfiguracji może być zmieniony oraz odczytany z urządzenia. Dodatkowo poszczególne zakładki (ekrany) konfiguracji mogą być prezentowane w sposób ułatwiający przedstawienie obrazu elementu urządzenia, całego urządzenia bądź też sieci. W wielu przypadkach konfiguracja może być przeprowadzona kilkoma metodami, z których skrypt może zastąpić kilka poleceń, plik konfiguracyjny większą liczbę poleceń a interfejs graficzny pozwalać na wykonanie tego samego za pomocą kilku kliknięć myszy.

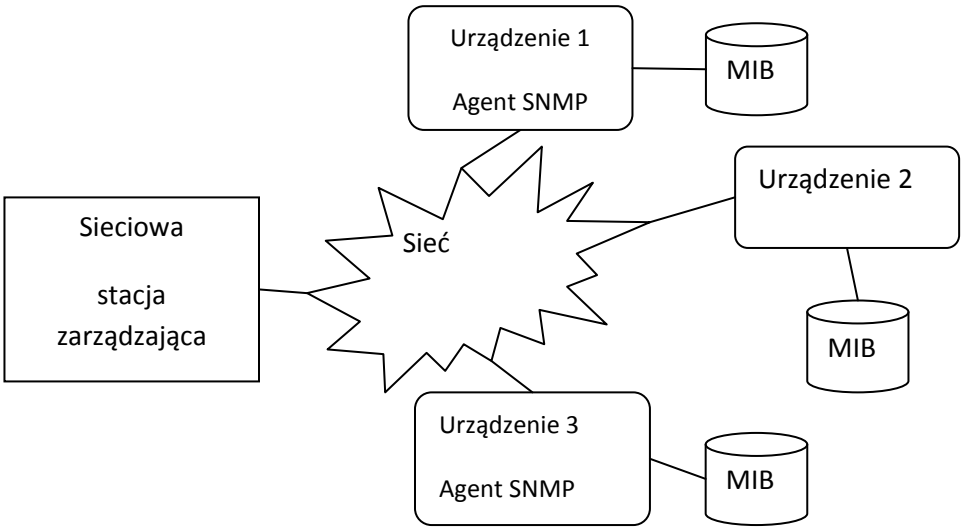
Do zarządzania większą ilością elementów sieci tworzone są odpowiednie aplikacje, które pozwalają na pobieranie informacji oraz przedstawianie statystyk za pomocą pojedynczego interfejsu. Aby aplikacja taka mogła pracować elementy sieci muszą przedstawiać dane w sposób jednolity ze standardowym formatem oraz z wykorzystaniem jednolitego protokołu komunikacyjnego. Trzy najczęściej używane standardy służące do zarządzania siecią to:

- SNMP Simple Network Management Protocol,
- COBRA,
- XML eXtensible Markup Language.

Dla każdego protokołu zdefiniowane zostały operacje konfiguracyjne jakie mogą być wykonane oraz informacje jakie mogą być pobierane z urządzenia. Każdy protokół posiada własny zbiór operacji. Informacje te są dostępne w postaci bazy MIB, w której każdy element sieci jest identyfikowany przez unikalny identyfikator OID (object identifier). Do wprowadzania zmian w MIB używane są wymienione wcześniej metody.

Najczęściej używaną metodą zarządzania jest SNMP [2], protokół działający na poziomie aplikacji. Jest on protokołem typu klient-serwer. Za jego pomocą można odczytywać oraz zapisywać poszczególne obiekty w bazie MIB wykorzystując do tego ich identyfikatory OID. Program kliencki nazywany jest sieciowym systemem zarządzającym. Łączy się on z programami serwerowymi pracującymi na zarządzanych elementach sieci (rys. 6.1). Programy serwerowy są to agenty SNMP wbudowane w urządzenia sieciowe. Agent łączy się z bazą MIB pozwalając na pobieranie danych oraz zmianę określonych parametrów. Każde z urządzeń posiada standardowy zbór odczytywanych i zapisywanych wartości. Możliwe jest również zastosowanie specyficznych parametrów poprzez specjalnie bazy MIB. Najprostsze dwie komendy do Get oraz Set służące odpowiednio do pobierania wartości oraz do jej ustawiania. Jeśli zachodzi potrzeba pobrania kilku wartości wykorzystywana jest komenda Get-Bulk. Do pobrania kolejnego obiektu (wartość OID) wykorzystać można Get-Next. Urządzenie może również w pewnych przypadkach samodzielnie wysłać zawiadomienie do stacji zarządzającej. Używana jest wtedy wiadomość o nazwie Trap. Istnieją trzy wersje protokołu SNMP. Pierwsze dwie powstały

w początkach rozwoju Internetu i nie uwzględniały aspektów bezpieczeństwa. Wersja SNMPv3 używa uwierzytelniania z wykorzystaniem kryptografii.



Rys. 6.1. Architektura SNMP

XML definiuje obiekty danych zwane dokumentami razem z zasadami dostępu do tych obiektów. Zasady te mogą być użyte do przesyłania i zmian danych dotyczących obiektów. XML jest uproszczoną wersją SGML. W zarządzaniu siecią XML służy do kodowania danych przysyłanych następnie przez HTTP. Do wymiany dokumentów XML może być używany protokół SOAP Simple Object Access Protocol. Służy on do zdalnego wywołania dostępu do obiektów XML.

CORBA jest technologią pozwalającą na zarządzanie poprzez wykorzystanie wymiany informacji pomiędzy aplikacjami zarządzającymi oraz zarządzanymi obiektami. Interfejs dostępu do obiektów jest definiowany w pliku IDL Interface Definition Language.

Łączenie się z zarządzanymi urządzeniami może odbywać się z wykorzystaniem łącz po których przesyłany jest zwykły ruch sieciowy lub za pomocą łącz dedykowanych do zarządzania. W pierwszym przypadku mamy do czynienia z zarządzaniem w paśmie (in-band), którego wadą jest możliwość wpływu zwykłego ruchu sieciowego na możliwości zarządzania. Drugie rozwiązanie nazywane zarządzaniem poza pasmem (out-of-band) wykorzystuje specjalne połączenia a nawet sieć pomiędzy urządzeniami. Rozwiązanie to wymaga dodatkowych nakładów, daje jednak separację ruchu związanego z normalną pracą sieci oraz ruchu związanego z zarządzaniem.

Wydaje się, że administracja siecią komputerową jest pojęciem bliższym sprzętu i jego konfiguracji natomiast zarządzanie siecią bliższą spojrzeniu całościowemu na sieć oraz świadczone przez nią usługi. Wyróżnić można następujące zadania wykonywane przez typowego administratora:

- ocena wymagań oraz planowanie sieci,
- instalacja urządzeń oraz software,
- konfiguracja urządzeń oraz aplikacji,
- monitorowanie sieci,
- diagnozowanie, wykrywanie i usuwanie usterek.

6.3. Ocena wymagań oraz planowanie

Etap planowania budowy i konfiguracji sieci pozwala osiągnąć poziom usług związany z wymaganiami biznesowymi. Wymagania funkcjonale określają jakie funkcje powinien spełniać system lub sieć. Wymagania nie-funkcjonalne mówią w jaki sposób dana sieć powinna działać uwzględniając koszty jej budowy, dostępność usług, niezawodność, bezpieczeństwo, skalowalność, łatwość w utrzymaniu, koszt działania oraz wydajność. Innym aspektem, na który należy zwrócić uwagę są możliwości efektywnego zarządzania planowaną siecią. Jakie funkcje zarządzania będzie można wykorzystać oraz czy pozwolą one na efektywne administrowanie w przyszłości.

6.4. Monitorowanie sieci

Monitorowanie wykonywane być może w dwóch głównych celach:

- wykrywanie anomalii, zawiadomienie o wystąpieniu określonego zdarzenia. Anomalie mogą być przejściowe lub permanentne. Te pierwsze nie zawsze wymagają dodatkowego działania administracyjnego. Same zdarzenia mogą być tylko informacjami o określonym działaniu potwierdzającymi sukces lub porażkę chociażby w zmianach konfiguracyjnych lub też wydarzeniami niezależnymi od działań administracyjnych. W pewnych przypadkach czas i szybkość reakcji jest kluczową ze względu na wymagania biznesowe.
- monitorowanie w celu analizy trendu i planowania. Wykonywane w dłuższym czasie oraz zazwyczaj z dużo mniejszą częstotliwością niż w poprzednim przypadku. Każde tego typu monitorowanie wymaga wstępnego określenia poziomu bazowego, pozwalającego zauważyć czasowe zmiany analizowanych parametrów.

Monitorowanie wymaga właściwego planowania. Zbyt duża ilość generowanych danych może powodować zmniejszenie wydajności urządzeń jak również problemy

z analizą i wykryciem rzeczywiście istotnych zdarzeń. Zbyt mała również może doprowadzić do pomijania zdarzeń istotnych z punktu widzenia sieci. Monitorowanie może być na dwa sposoby:

- Pasywnie – co oznacza okresowe odpytywanie urządzenia o wartości interesujących nas parametrów oraz przesyłanie ich wartości do systemu zarządzania siecią. Mierzony tutaj parametry przedstawiają rzeczywiste używanie zasobów. Podejście takie nie zwiększa dodatkowo obciążenia zasobów sieciowych. Pasywne monitorowanie może być używane w celach rozliczeniowych.
- Aktywne oznacza generowanie dodatkowego obciążenia poprzez wysyłanie ruchu przez urządzenia przetłaczające lub wysyłanie dodatkowych żądań do serwerów i aplikacji. Aktywne monitorowanie jest szczególnie przydatne do diagnozowania sieci, wykrywania problemów oraz określania jakości i poziomu świadczonych usług SLA (Service Level Agreement).

Standardem rozszerzający SNMP w zakresie monitorowanie sieci jest RMON Remote Network Monitoring. Protokół ten pozwala na badanie przepływów w sieci. Zamiast systemu agentów stosowane są sondy. Sam RMON nie wnosi dodatkowych zmian w stosunku do SNMP, choć do jego pracy stosowane być muszą bazy MIB-II zdalnego nadzoru. Pierwsza wersja protokołu służyła do badania ruchu na poziomie warstwy drugiej modelu OSI. Wersja druga pozwala na monitorowanie na poziomie dowolnej warstwy. Można więc np. analizować dane konkretnego protokołu dla wybranego hosta sieciowego. Sondy RMON mogą być instalowane na specjalnych urządzeniach służących tylko i wyłącznie do monitorowania sieci, choć coraz częściej są one dostępne w urządzeniach sieciowych a nawet w kartach sieciowych hostów. Stosując SNMP administrator może określać ruch dla poszczególnych urządzeń, RMON pozwala patrzeć na ruch bardziej całościowo jako całkowity przepływ sieciowy. Sonda RMON czyli nadzorca sieci (analizator, monitor sieci) analizuje ruch z wykorzystaniem filtrów. Dane są przechowywane do późniejszej analizy, możliwe jest wykonywanie i prezentowanie różnorodnych statystyk w stosunku do analizowanego przepływu. Typowa architektura monitorowania składa się z sond (zdalnie nadzorcy) umieszczonych w każdej podsieci oraz stacji zarządzania, z którą się komunikują. Cechą charakterystyczną RMON jest specyficzna definicja bazy MIB oraz niewielki wpływ na obciążenie systemów z nadzorcami oraz stacji zarządzających. RMON MIB definiuje 10 grup informacji takich jak: statystyki czasu rzeczywistego, dane archiwalne odnośnie statystyk, statystyki dla poszczególnych adresów MAC, najbardziej aktywne hosty w sieci, charakterystyki połączeń pomiędzy węzłami, elementy specyficzne dla monitorowania Token Ring. Kolejne

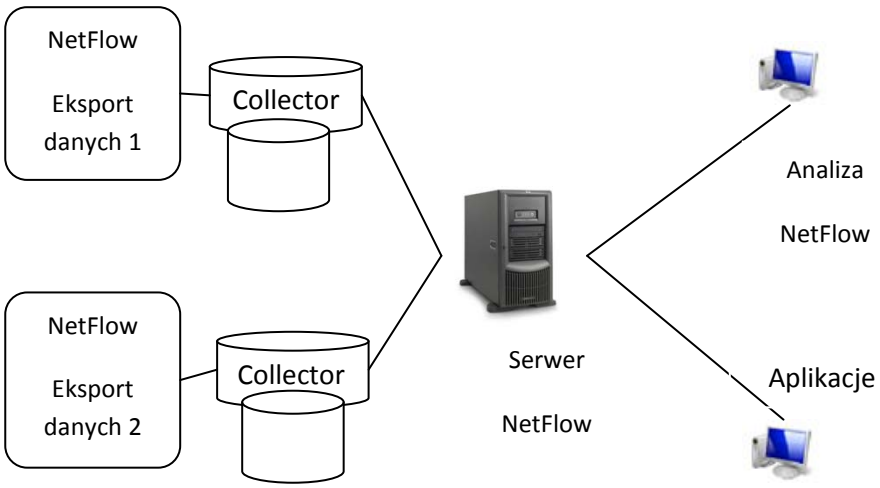
4 grupy informacji umożliwiają precyzyjne określenie monitorowanych parametrów. Należą do nich [2]:

- określane przez administratora poziomy parametrów przepływu generujące w alarmy w kierunku stacji zarządzającej,
- filtry określające jakiego rodzaju ruch będzie podlegał monitorowaniu,
- dane przechowujące ruch określony przez zdefiniowane wcześniej filtry,
- pułapki protokołu SNMP wysyłane przy wystąpieniu alarmu po przekroczeniu zdefiniowanego poziomu.

Baza RMONv2 MIB dodała 10 nowych grup pozwalających na monitorowanie w wyższych warstwach modelu OSI. Sam system zarządzający nie musi ciągle odpytywać poszczególnych nadzorców, którzy nadal będą zbierać dane i prześlą je jeśli zajdzie taka potrzeba. Dane mogą być analizowane przez nadzorcę i wysyłane do stacji zarządzającej w postaci statystyk. Jeden nadzorca jest w stanie, przy odpowiedniej konfiguracji, wysyłać informacje do wielu stacji zarządzających. Konfiguracja nadzorcy jest wykonywana przez zarządcę. Do komunikacji pomiędzy zarządcą oraz nadzorcami wykorzystywany jest protokół SNMP.

Zmieniające się usługi sieciowe sprawiają, że przez sieć jest przesyłanych coraz więcej strumieni multimedialnych. Tego typu ruch sieciowy wymaga innych parametrów w stosunku do klasycznych usług sieciowych. W związku z tym zaistniała potrzeba nieco innego podejścia do monitorowania sieci. Pojawiło się pojęcie monitorowania przepływów (flow monitoring). Wzrosło znaczenie parametrów jakościowych QoS oraz poziomu świadczonych usług SLA (Service Level Agreement). Od narzędzi monitorowania sieci oczekiwane jest nie tylko przedstawianie obciążenia sieci ale również rozróżnianie od czego to obciążenie pochodzi. Jakie aplikacje generują ruch oraz w jaki sposób wpływają one na siebie wzajemnie. Pierwsze kroki w kierunku monitorowania przepływów postawiła firma Cisco wprowadzając protokół NetFlow [3,4]. Pierwszym standardem zatwierdzonym przez IETF był IPFIX (IP Flow Information eXport). Jak w każdym standardzie wprowadzono definicję kilku podstawowych pojęć odnoszących się do rozpatrywanej dziedziny. Przepływ jest to zbiór pakietów IP, który w określonym czasie przechodzi przez miejsce obserwacji w sieci. Parametrami takiego przepływu mogą być np. wybrane pola nagłówka (adres IP, port TCP/UDP, pola aplikacji) oraz kierunek ruch pakietu (np. adres następnego węzła). Jeśli pakiet spełnia warunki definiowane przez parametry to należy do danego przepływu. Przepływ jest badany w punkcie obserwacji, gdzie przepływają pakietu. Na podstawie ich nagłówków proces monitorujący dokonuje ich klasyfikacji oraz znakowania czasowego

zmieniając jednocześnie rekordy określające monitorowane przepływy. Kolejny proces zwany procesem eksportującym wysyła przechowywane rekordy do procesów zbierających, w których są one gromadzone oraz przetwarzane. Pomiędzy procesem eksportującym oraz zbierającym dane są przesyłane protokołem IPFIX. Typowa architektura NetFlow została przedstawiona na rys. 6.2.



Rys. 6.2. Architektura NetFlow

6.5. Zarządzanie awariami

Awarie i błędy zdarzają się z różnych powodów. Podstawowym zadaniem administratora jest wykrycie błędu oraz przywrócenie systemu do poprawnej pracy. Jest to nie proste zadanie wymagające określenie przyczyny na podstawie obserwowanych jej efektów, które mogą być obserwowane przez system monitorowania. Informacje diagnostyczne dostępne przez różnorodne urządzenia oraz aplikacje jak do tej pory nie zostały poddane standaryzacji. W złożonych sieciach budowane są specjalne systemy zarządzania błędami, w których każde zdarzenie jest przetwarzane oraz analizowane z wykorzystaniem odpowiedniej bazy danych, dzięki którym określić można przyczynę zaistniałego zdarzenia.

W przypadkach kiedy wymagana jest dużą niezawodność oraz wysoka dostępność stosowane mogą być rozwiązania mające funkcję samo naprawiania. System taki jest w stanie automatycznie wykryć oraz dokonać korekcji wybranych rodzajów błędów. O jakości takiego rozwiązania mówi prawdopodobieństwo pokrycia błędów, czyli jaka część możliwych do pojawienia się błędów będzie przez system obsłużona i nie spowoduje dalszej propagacji błędu.

Sieć oraz systemy komputerowe można zaprojektować w taki sposób by były w stanie unikać błędów. Rozwiązania takie przeważnie posiadają zaimplementowane określonego rodzaju nadmiarowość (redundancję). Dla sieci protokoły routingu potrafią przysyłać ruch innymi ścieżkami w przypadku awarii jednego łącza lub węzła. Nadmiarowość dotyczyć może łącza sieciowego, węzła sieciowego, elementu urządzenia takiego jak karta sieciowa. Element nadmiarowy może pracować jako rezerwa aktywna biorąc aktywny udział w obsłudze obciążenia, jako rezerwa gorąca, która może szybko przejąć obciążenie ale tylko w przypadku awarii elementu podstawowego. Zimna rezerwa potrzebuje dłuższego czasu na przejęcie obciążenia [12]..

6.6. Zarządzanie siecią oparte o zasady

Jednym z nowszych trendów w podejściu do zarządzania siecią jest zarządzanie oparte o zasady PBNM (Policy Based Network Management) [5,6]. Podejście takie można zdefiniować jako model przekładający zasady biznesowe na konfigurację sieci. Jest to nowa metodologia podejścia do zarządzania, w której modelowane są zbiory obiektów oraz zależności pomiędzy nimi. Obiekty reprezentują informację o elementach sieci. Od strony biznesowej firma sprzedaje usługi a klient płaci za określoną ich jakość. Manager wymaga więc by administratorzy zapewnili klientowi to za co firma zobowiązała się w umowie sprzedaży usługi. Manager musi posiadać możliwość sprawdzenia czy owe wymagania są spełnione i czy ewentualne zażalenia ze strony klientów są rzeczywiście adekwatne do sytuacji. Jakość usług QoS (Quality of Service) nie jest pojęciem jednoznacznym ponieważ parametry je definiujące wraz z ich wartościami krytycznymi zmieniają się w zależności od rodzaju świadczonej usługi. Określenie zasad zarządzania pozwala na uproszczenie zarządzania dla złożonych i rozbudowanych sieci. Zasady pozwalają na wykonywanie zadań w zdefiniowany wcześniej sposób, co jest szczególnie przydatne w szybkiej reakcji na zdarzenia. Przestrzeganie zasad jest szczególnie istotne w zarządzaniu bezpieczeństwem.

Z drugiej strony PBNM jest zbiorem zasad, które należy zastosować w przypadku wystąpienia określonego warunku lub warunków. Oznacza to automatyzację zadań i zwiększenie szybkości reakcji. Służy do konfiguracji dużej liczby elementów sieciowych. Zasada składa się z jednej lub wielu reguł pozwalających podjąć decyzje właściwe do zaistniałej sytuacji związane ze zmianą stanu zarządzanego elementu lub elementów sieci.

6.7. Ochrona dostępu do sieci

W systemie Windows 2008 wprowadzono kilka ciekawych rozwiązań dotyczących sieci. Jedną z takich nowości technologia ochrony dostępu do sieci NAP Network Access Protection. Pozwala ona administratorowi sieci wprowadzić zasady bezpieczeństwa odnośnie stacji roboczych w całej sieci. Dzięki czemu możliwa jest ochrona i zabezpieczanie całej sieci przed złośliwym oprogramowaniem. Serwery służą do centralnej konfiguracji, przechowywania definicji stanu zdrowia systemów w sieci oraz wymuszania spełniania wymogów przed otrzymaniem dostępu do sieci. Zapobiega to rozprzestrzeniania się zagrożenia z zainfekowanego komputera na całą sieć.

Sieci korporacyjne są z reguły dobrze zabezpieczone za pomocą różnego rodzaju systemów, z których najpopularniejsze to ściany ogniowe. Usługi, które powinny być widoczne z Internetu przeważnie znajdują się na serwerach w wydzielonej sieci. Brak bezpośredniego dostępu do stacji roboczych z sieci nie oznacza ich zupełnego bezpieczeństwa. Możliwe są inne drogi dostępu złośliwego oprogramowania, które atakuje stacje robocze, chociażby poprzez strony www, pocztę elektroniczna, dostęp komputerów przenośnych oraz zdalny dostęp użytkowników poprzez VPN.

Ochronę stacji roboczych zapewnić mogą:

- programy antywirusowe,
- aktualizacje automatyczne dla systemów operacyjnych,
- zapory ogniowe na stacjach roboczych.

Z punktu widzenia sieci działające prawidłowo ściany ogniowe na stacjach roboczych chronią przed skanowaniem portów oraz wysyłaniem niebezpiecznego ruchu w kierunku komputerów. Najważniejsze jest więc z punktu widzenia całej sieci by tylko takie komputery, które posiadają włączoną całą możliwą ochronę mogły korzystać z zasobów sieciowych, podczas gdy pozostałe komputery nie zostaną dopuszczone do komunikacji. Samo włączenie wymienionych trzech składników ochrony nie jest wystarczające. Zarówno pobrane i zainstalowane aktualizacje dla systemu operacyjnego jak też definicje wirusów dla programów antywirusowych muszą być aktualne. Administrator z wykorzystaniem NAP może określić jakie warunki powinien spełniać system aby uznać go za zdrowy. Wymagania te mogą się różnić w zależności od środków ochrony dostępnych w danej organizacji oraz konfiguracji systemów operacyjnych. Następnie wymagania te muszą być sprawdzone a na ich podstawie podjęta zostaje decyzja o dostępie sieciowym danej stacji roboczej. Działanie takie nazywane jest narzucaniem wymagań odnośnie kondycji systemu [7]. W ogólności system może

spełniać warunki, co oznacza że jest zdrowy lub też warunki mogą nie zostać spełnione co powinno skutkować możliwością stosowania odpowiednich metod leczniczych w postaci dostępu do specjalnej infrastruktury sieciowej zwanej siecią naprawczą. Najczęściej jako taka sieć występują serwery pozwalające na pobranie aktualizacji systemu operacyjnego wraz z serwerami umożliwiającymi aktualizację bazy danych wirusów. Platforma NAP pozwala na stosowanie aplikacji, które będą czuwały nad innymi aspektami bezpieczeństwa lub konfiguracji.

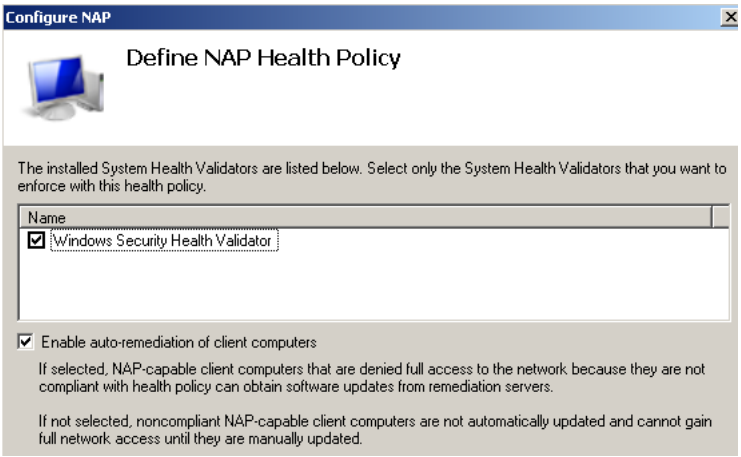
Aby można było wprowadzić technologię NAP potrzebna jest odpowiednia infrastruktura sieciowa. Wymagane elementy infrastruktury różnią się w zależności od metody narzucania (enforcement) zasad. Możliwe są 4 główne metody narzucania oraz jedna dodatkowa [7]:

- wykorzystanie serwera DHCP do przyznawania dostępu z wykorzystaniem konfiguracji adresów przyznawanych klientom,
- wykorzystaniem połączeń IPSec do przyznawania chronionego dostępu do infrastruktury sieciowej,
- wykorzystywanie urządzeń obsługujących standard IEEE 802.1X,
- ochrona dostępu sieciowego dla klientów zdalnego dostępu VPN,
- sprawdzanie kondycji oraz narzucanie przy dostępie poprzez bramę usług terminalowych (wymieniana często jako metoda dodatkowa).

Każdą z wymienionych powyżej metod można stosować jako metodę osobną lub też można je łączyć. W każdej z metod narzucania potrzebne są następujące elementy [8]:

- Active Directory pozwalająca na konfigurację grupy komputerów za pomocą zasad grup jak również scalająca poszczególne elementy infrastruktury,
- klienci kompatybilni z platformą NAP, do których należą systemy XP SP3, Vista oraz Windows 7,
- elementy pozwalające na dostęp do sieci realizujące narzucanie NAP, pozwalają na przełączenie komputera do sieci właściwej lub do sieci z ograniczeniami,
- serwery Windows 2008 z usługą NPS Network Policy Server stanowiący bazę zasad odnośnie kondycji sprawdzanych systemów oraz jednocześnie sprawdzający kondycję na podstawie zasad skonfigurowanych przez administratora,
- serwery dostarczające wymagań odnośnie kondycji określające aktualne dane odnośnie bazy wirusów oraz aktualizacji systemu. Dane te są przesyłane do serwerów zasad kondycji,

- serwery naprawcze stanowiące główną część sieci naprawczej, do której należeć mogą również komputery nie kompatybilne z NAP oraz nie spełniające wymogów sieciowych.



Rys. 6.3. Definiowania zasad wymogów zdrowia dla systemów klienckich

Komputery klienckie od Windows XP SP3 posiadają wbudowany składnik zwany agentem kondycji systemu (SHA System Health Agent), który sprawdza ustawienia Centrum zabezpieczeń systemu. Zebrane przez SHA dane służą do tworzenia świadectwa kondycji. Każdy agent SHA odpowiada za jeden składnik bezpieczeństwa (aktualizacje, antywirus, ściana ogniowa). Przy dowolnej zmianie każdego ze składników system wystawia nowe świadectwo kondycji systemu składające się z poszczególnych świadectw kondycji oraz wersji NAP.

Świadectwo kondycji systemu jest oceniane przez serwer NPS z zasadami, na którym pracują moduły weryfikacji kondycji systemu SHV System Health Validator. NPS tworzy świadectwo odpowiedzi kondycji stanu systemu w celu przesłania ich do sprawdzanego systemu, na którym poszczególni agencji SHA nie spełniający wymogów mogą podjąć działania w celu naprawy swojego stanu kondycji.

Na kliencie sieciowym działa klient narzucania NAP oddzielny dla każdej metody wymuszania (DHCP, VPN, IPSec, 802.1X, TS Gateway). Klient taki żąda dostępu do sieci poprzez punkt narzucania, który taki dostęp oferuje przekazując tam swoją aktualną kondycję.

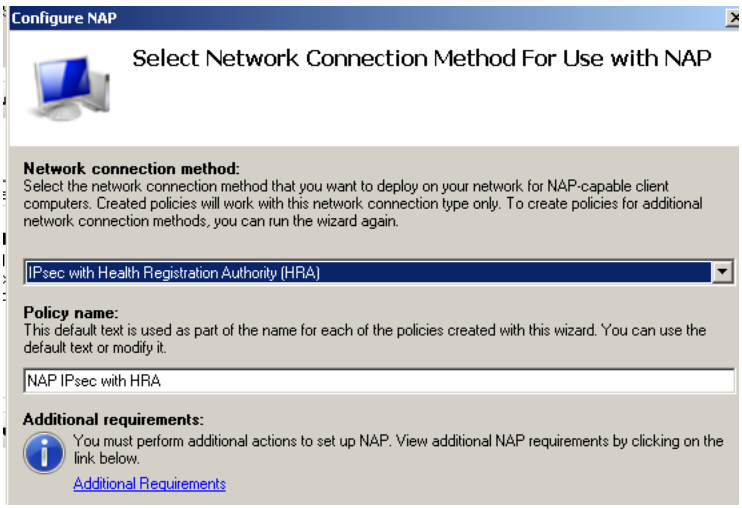
Po stronie serwerowej działają moduły weryfikujące kondycję SHV System Health Validator dla każdego elementu SHA po stronie klienta. SHV jest elementem pozwalającym na dostęp do sieci na określonym przez administratora poziomie oraz wymuszającym ten poziom dla komputera klienta. Po stronie serwerowej istnieją tylko trzy składniki narzucania (DHCP,IPSec, TS Gatway) ponieważ

narzucanie 801.1X jest realizowane przez odpowiedni sprzęt a dla połączeń VPN poprzez server dostępu. Administrator na serwerze NPS poprzez odpowiednie ustawienia steruje wymaganiami co do stanu kondycji komputera klienta. Wymiana informacji pomiędzy punktami narzucania oraz serwerami zasad kondycji NPS następuje za pomocą komunikatów protokołu RADIUS.

Najprostszą metodą wymuszania jest wykorzystanie serwera DHCP, w którym w zależności od spełnienia lub nie wymagań kondycji klient otrzymuje inną konfigurację adresu. Stan klienta jest monitorowany na bieżącą i w przypadku jego zmiany następuje zmiana jego konfiguracji IPv4. Użytkownicy mający prawo do zmiany konfiguracji adresu na swoich komputerach mogą obejść mechanizmy NAP DHCP. Stan kondycji komputera z jego świadectwem jest przesyłany w żądaniu do serwera DHCP, który przesyła je do serwera ze zdefiniowanymi zasadami kondycji NAP. Serwer NPS sprawdza w jaki sposób wymagania zostały przez klienta spełnione i wysyła o tym informacje do serwera DHCP. Jeśli klient nie spełnia wymagań to serwer DHCP otrzymuje informację o potrzebie zastosowania ograniczonego dostępu jak również informacje do klienta jakie działania powinien podjąć by wymagania zostały spełnione. Przyznana w tym przypadku konfiguracja IP będzie pozwalała na dostęp do jedynie wybranego naprawczego fragmentu sieci. Po ewentualnym wykonaniu działań naprawczych klient ponowi do serwera DHCP komunikat z informacjami o stanie zdrowia, które zweryfikuje serwer NPS wysyłając serwerowi DHCP konfigurację adresu z pełnym dostępem.

Wymuszanie IPSec jest metodą bardzo bezpieczną. W sieci musi zostać skonfigurowana metoda ochrony zasobów (izolacja poszczególnych serwerów lub całej domeny, dostęp do określonych adresów IP, dostęp na określonych portach) IPSec. Klient nie spełniający wymagań stanu zdrowia nie otrzyma certyfikatu potrzebnego do rozpoczęcia bezpiecznej komunikacji, ale też nie zainfekuje sieci. Certyfikaty sieciowe są przyznawane przez urząd rejestrowania kondycji na serwerze narzucania IPSec, służą one następnie do uwierzytelniania oraz szyfrowania połączeń. Podczas próby dostępu do sieci komputer klienta wysyła swoje świadectwo kondycji do urzędu rejestrowania kondycji, który przesyła je dalej do serwera NPS z zasadami. Serwer zasad kondycji ocenia otrzymane świadectwo wysyłając do urzędu rejestrowania świadectwo odpowiedzi na stan kondycji. Jeśli klient nie spełnia wymagań w świadectwie tym znajdować się będą informacje o możliwościach naprawczych. Urząd rejestracji nie wystawi wtedy klientowi certyfikatu a więc nie będzie on mógł podjąć bezpiecznej komunikacji z siecią wyłączywszy serwery naprawcze. Jeśli jego stan ulegnie poprawie to świadectwo kondycji zostanie ponownie wysłane do punktu narzucania a następnie serwera NPS. Spełnienie wymagań zaowocuje wysłaniem świadectwa odpowiedzi

na stan kondycji z informacją o zgodności klienta z wymaganiami. W efekcie urząd rejestrowania wystawi klientowi certyfikat zezwalający na komunikację IPSec z siecią właściwą.



Rys. 6.4. Wybór metody wymuszania NAP

Wymuszanie VPN działa dla klientów łączących się za pomocą serwera dostępu zdalnego. Komputer nawiązuje połączenie VPN, podczas którego zostaje sprawdzony jego stan kondycji. Jeśli komputer jest zdrowy to otrzymuje pełny dostęp do zasobów sieci lokalnej. Komputer nie spełniający wymagań odnośnie kondycji otrzymuje dostęp ograniczony za pomocą filtrowania. Weryfikacja stanu zdrowia klienta następuje po uwierzytelnieniu użytkownika, po którym komputer kliencki wysyła świadectwo zdrowia do serwera NPS. Wynik weryfikacji jest jednocześnie wysyłany do klienta jak też do serwera VPN. Klient otrzyma świadectwo odpowiedzi na stan kondycji z ewentualnymi wymaganiami odnośnie działań naprawczych. Serwer VPN otrzyma informacje o filtrach pakietów. Dopiero w tym momencie zakończony zostanie proces ustanawiania połączenia VPN. Dla klienta spełniającego wymagania serwer VPN przyzna pełen dostęp poprzez zdalne połączenie. Dla klienta nie w pełni spełniającego wymagania dostęp zostanie ograniczony za pomocą odpowiednich filtrów, co w efekcie spowoduje dostęp jedynie do serwerów naprawczych. Klient będzie mógł na podstawie świadectwa stanu kondycji wysłać do serwerów naprawczych żądanie aktualizacji systemu operacyjnego lub też bazy wirusów dla programu antywirusowego. Po udanej aktualizacji komputer klienta musi ponownie wykonać procedurę zestawiania połączenia VPN, w której wyśle swoje świadectwo do NPS. Serwer zasad stwierdzi

zgodność wymagań po czym prześle serwerowi VPN informację o pozwoleniu na pełen dostęp. Po czym serwer VPN dokończy zestawianie połączenia o pełnych możliwościach.

Ostatni rodzaj narzucania wymaga urządzeń pracujących w standardzie 802.1X takich jak przełączniki sieciowe lub punkty dostępu bezprzewodowego. Na urządzeniu administrator konfiguruje odpowiednie profile dostępu wykorzystując listy kontroli dostępu oparte o filtry pakietów lub sieci VLAN. Jak w każdej z metod narzucania sprawdzenie stanu kondycji systemu następuje w momencie podejmowania próby dostępu oraz w trakcie działania w sieci. Proces uzyskania dostępu zaczyna się od klasycznego uwierzytelniania 802.1X. Komputer kliencki wysyła do serwera NPS poświadczenia użytkownika oraz komputera w celu uwierzytelnienia. Dopiero po prawidłowym uwierzytelnieniu wstępnych serwer NPS żąda od komputera klienckiego jego świadectwa zdrowia, które następnie weryfikuje wystawiając tak jak przy innych metodach narzucania świadectwo weryfikacji kondycji klienta. Wynik weryfikacji jest wysyłane do klienta by ewentualnie mógł wykonać działania naprawcze oraz do urządzenia 802.1X. Punkt dostępowy lub przełącznik dokończy proces uwierzytelniania ograniczając dostęp do sieci lub przyznając pełen dostęp. Dostęp ograniczony powinien oczywiście pozwalać na korzystanie z serwerów naprawczych. Klient wykorzysta je do pobrania potrzebnych mu aktualizacji, po czym wyśle nowe świadectwo kondycji do serwera NPS. Proces uwierzytelniania 802.1X zacznie się wtedy od początku, a zakończy się przy spełnieniu wszystkich warunków otrzymaniem przez klienta pełnego dostępu.

Ponieważ platforma NAP nie jest kompatybilna ze starszymi systemami administrator może podjąć decyzję o wyłączeniu tych komputerów z określonych wymagań poprzez utworzenie odpowiedniej zasady wykluczającej.

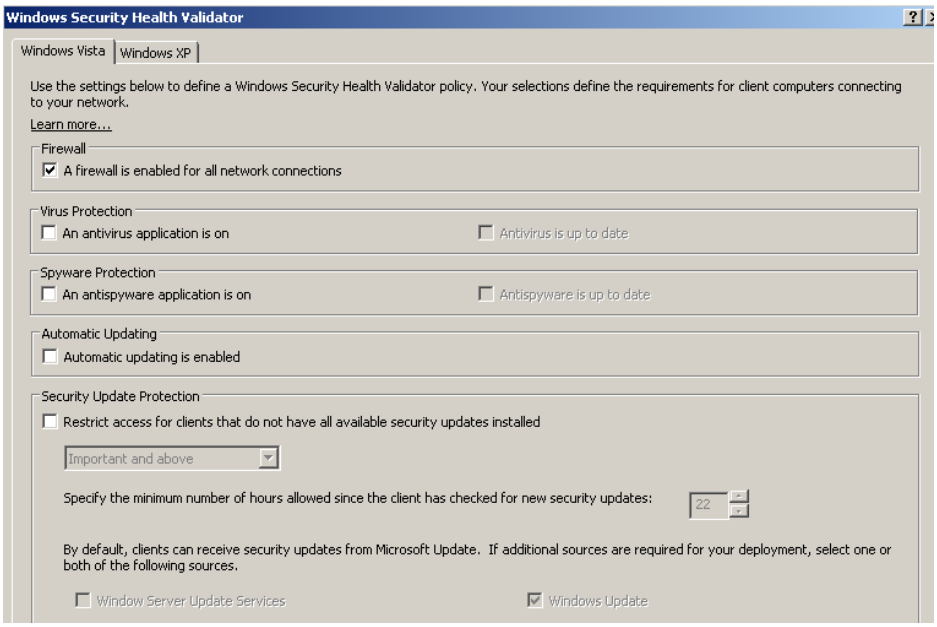
Wdrożenie platformy NAP polega na skonfigurowaniu serwera NPS, którym może być już istniejący serwer RADIUS. Punkty narzucania NAP (przełączniki 802.1X, serwer VPN, serwer DHCP, urząd rejestrowania kondycji dla IPSec) należy skonfigurować jako klientów RADIUS (na serwerze NPS). Następnie należy skonfigurować wymagania kondycji NAP w sensie weryfikacji z nimi zgodności klientów oraz jak należy obsłużyć klientów niezgodnych lub też takich, którzy nie obsługują platformy NAP. Do wymagań kondycji zaliczane są następujące składniki [7]:

- Zasady żądania połączeń (Connection Request Policies) służą do ustanowienia reguł NPS, które mają za zadanie stwierdzenie czy dane żądanie od klienta powinno zostać przetworzone lokalnie czy też przesłane do innego serwera RADIUS. W tym drugim przypadku należy najpierw

stworzyć grupę zdalnych serwerów RADIUS. Dla platformy NPS żądania będą przetwarzane lokalnie.

- Zasady kondycji (Health Policies) określają moduły SHV jakie będą brane pod uwagę przy rozpatrywaniu kondycji klienta oraz warunki jakie te moduły stawiają komputerom klienckim. Warunki modułów SHV mogą sprawdzać status kondycji klienta dla poszczególnych modułów w następujący sposób:

- klient musi spełniać warunki dla wszystkich SHV – służy do wybrania komputerów w pełni zgodnych z wymaganiami,
- klient nie spełnia warunków dla żadnego modułu SHV – służy do wybrania komputerów nie zgodnych w żadnym wymaganym elemencie,
- klient spełnia warunki dla wybranych (co najmniej jednego) modułów SHV służy do wyboru komputerów częściowo zgodnych z wymaganiami,
- klient nie spełnia warunków dla jednego lub większej ilości modułów SHV pozwala na wybranie komputerów, które będą określone jako niezgodne gdy dowolny moduł nie będzie zgodny.



Rys. 6.5. Widok wyboru modułów weryfikacji kondycji systemu

Ochrona dostępu do sieci (NAP w konsoli NPS) zawiera dwa elementy:

- moduły weryfikacji kondycji systemu (System Health Validators) – konfigurowalny zestaw wymagań, domyślnie z modułem weryfikacji

Windows (Windows Security Health Validator) odpowiadającym ustawieniom z centrum zabezpieczeń systemu,

- grupy serwerów naprawczych (Remediation Server Group) – czyli serwery naprawcze dostępne przy wymuszaniu z wykorzystaniem serwera DHCP oraz VPN.

Zasady sieci określające jakie warunki zgodności kondycji systemu należy brać pod uwagę oraz co zrobi z komputerami ich nie spełniającymi bądź komputerami niekompatybilnymi z NAP. Ustawienia zasad sieci dotyczących wymuszania NAP (NAP Enforcement) pozwalają na następującą konfigurację:

- zezwolenie na pełny dostęp sieciowy, który jest typowo stosowany dla klientów zgodnych z wymaganiami i spełniających wszystkie postawione warunki,
- zezwolenie na pełny dostęp sieciowy ale przez określony czas z jedoczesnym powiadomieniem użytkowników o terminie z wymaganą pełną zgodnością,
- dostęp z ograniczeniami stosowany dla klientów nie spełniających wymagań z jednoczesnym zawiadomieniem o ich nie spełnianiu,
- automatyczna naprawa komputerów (auto-remediation).

Konfigurowanie usługi NAP zostanie przedstawione na przykładzie metody wymuszania DHCP. Do jej wdrożenia oprócz klientów usługi kompatybilnych z NAP potrzebny jest element wymuszający (narzucający), którym w tym przypadku jest serwer DHCP oraz serwer z zadami odnośnie kondycji systemu, którym jest serwer z usługą NPS [9]. Przed rozpoczęciem konfiguracji należy określić czy w sieci będą działały komputery klienckie, które nie będą podlegały wymuszaniu NAP. Ponieważ domyślnie nie mogły by one korzystać z sieci jeśli używały by do konfiguracji serwer DHCP administrator powinien skonfigurować na serwerze NAP odpowiednie wykluczenie. Polega ono na stworzeniu zasady sieciowej przyznającej dostęp dla grupy komputerów. Grupę taką należy stworzyć i dodać do niej wszystkie komputery, które chcemy wykluczyć z obszaru działania platformy NAP. Grupę taką można stworzyć pod warunkiem posiadania usługi katalogowej Active Directory, z obszarem sieci, do której należeć muszą rozpatrywane komputery. W sieciach z ograniczoną ilością zasobów jeden serwer Windows 2008 może współdzielić rolę DHCP oraz usługę roli NPS. Na serwerze DHCP do obsługi klientów nie spełniających warunków stworzona jest specjalna klasa użytkownika Default Network Access Protection Class, której opcje mogą służyć do szczegółowej konfiguracji tego typu klientów. Domyślnie dostępne są dwie opcje DHCP dla klientów niezgodnych

tj. maska podsieci oraz trasy statyczne bez klas kierujące klienta do serwerów naprawczych. Konfiguracja klienta polega na następujących krokach [10]:

- włączenie centrum zabezpieczeń systemu Windows,
- włączenie klienta egzekwowania kwarantanny (klient narzucania) dla komputerów klientów,
- włączenie automatycznego uruchamiania usługi agenta NAP na komputerze klienta,

Powyższe zadania można w systemie domenowych skonfigurować za pomocą zasad grup dla wszystkich komputerów w sieci.

Na serwerze NPS należy skonfigurować ustawienia serwera RADIUS oraz ustalić zasady wymagane dla określenia poprawnej kondycji NAP. Jeśli w danej organizacji istnieje potrzeba rozszerzenia modułów poza standardowy SHV będący odzwierciedleniem centrum zabezpieczeń systemów klienckich to należy takie rozszerzenia zainstalować. Konfiguracja serwera RADIUS polega na dodaniu serwera DHCP jako klienta usługi RADIUS, koniecznie z uwzględnieniem opcji obsługi ochrony dostępu do sieci dla klienta. Do konfiguracji zasad wymagania dla kondycji systemu zalecane jest użycie kreatora, który zapewni pełną konfigurację uwzględniającą wszystkie niezbędne do działania systemu ustawienia. W kreatorze wybierane są m.in. następujące ustawienia:
metoda połączenia sieciowego używana z NAP – w metodzie DHCP wybieramy DHCP,

- wybór i dodanie klientów RADIUS – w tym przypadku serwery DHCP,
- zakresy DHCP używane dla klientów,
- serwery naprawcze (grupa zabezpieczeń oraz adres URL),
- wybór właściwych zasad kondycji poprzez wybór modułów SHV porównujących stan poszczególnych zabezpieczeń sprawdzanych przez NAP.

Konfiguracja serwera DHCP polega na następujących krokach [10]:

- włączenie obsługi NAP przez serwer DHCP we właściwościach IPv4 serwera na zakładce NAP,
- konfiguracja dodatkowych opcji w klasie domyślnej dla ochrony dostępu do sieci, z której korzystają komputery nie spełniające warunków zdrowia,
- dla zakresów używanych przez NAP należy skonfigurować nazwy profili, które będą powiązane z zasadami NPS dla klientów z ograniczeniami nie spełniającymi wymogów kondycji.

Po zaprojektowaniu i skonfigurowaniu usługi następuje faza testowania, w której usługa NAP nie musi ograniczać dostępu dla komputerów lecz działać

w trybie rejestrowania. Jeśli testy wypadną poprawnie można przejść od rejestrowania do faktycznego działania z ograniczeniami. Typowymi zadaniami administracyjnymi w tej fazie jest monitorowanie usługi oraz ewentualnie konfiguracja nowych komputerów klienckich do obsługi NAP. Najtrudniejszym zadaniem administracyjnym jest właściwa reakcja na pojawiające się problemy. W przypadku problemów użyć można narzędzi diagnostycznych. Najprostsze z nich to ipconfig /all, w którym otrzymujemy informację o stanie systemu jako bez ograniczeń lub z ograniczeniami. Kolejne przydatne polecenie to netsh z kontekstem nap. Dokładną konfigurację klienta określić można poleceniem netsh nap client show configuration, natomiast stan klienta włączając w to poszczególne SHA poleceniem netsh nap client show state. Monitorowanie usługi polega również na przeglądaniu zdarzeń w dzienniku zdarzeń w dzienniku aplikacji odpowiadającemu NAP. Diagnozowanie problemów polega przede wszystkim na kolejnym zmniejszaniu zakresu elementów funkcjonujących nieprawidłowo. Po pierwsze administrator powinien wyizolować problem zaczynając od tego czy znajduje się on po stronie komputera klienta, po stronie serwera NPS, serwera DHCP czy też problemem są połączenia sieciowe pomiędzy składnikami infrastruktury.

NAP jest złożoną technologią integrującą w sobie mechanizmy sieciowe z systemami operacyjnymi. Podstawowym celem jej stosowania jest zwiększenie bezpieczeństwa sieci komputerowej opartej na systemach operacyjnych Windows poprzez wymuszenie ochrony dostępu do sieci. Jej planowanie, konfiguracja, testowanie, monitorowanie a więc szeroko rozumiana administracja wymaga znajomości wielu technologii sieciowych takich jak: usługa katalogowa, IPSec, infrastruktura klucza publicznego PKI, technologia 802.1X, dostęp zdalny VPN, DHCP, DNS oraz wiedzy dotyczącej systemów operacyjnych i ich styku z siecią w postaci interfejsów oraz usług sieciowych.

6.8. Niezawodność z perspektywy usług sieciowych

Świadczenie niezawodnych usług Internetowych zależy od niezawodności elementów na drodze łączącej usługodawcę z klientem. Awaria lub duże opóźnienie na fragmencie łącza powoduje całościowy spadek jakości usług. Analizę niezawodności serwera Web rozważać należy w zależności od jego zawartości. Znacznie prostsze a zarazem bardziej niezawodne są serwery ze statyczną zawartością. Nie mają one jedna funkcjonalności wymaganej w dzisiejszych czasach. Wydajność takiego serwera zależy od tego czy zawartość statyczna składa się z wielu niezależnych stron czy jest raczej kilkoma dużymi fragmentami danych.

Przy dużej ilości żądań ograniczeniem wydajności serwera może być procesor, RAM lub karta sieciowa. Jeśli z mała ilością stron związana jest dużą ilość przesyłanych danych to problem w wydajności staje się system plików. Większość dzisiejszych aplikacji Web jest związanych z zawartością dynamiczną, która wymaga znacznie większej wydajności serwera. Zawartość wysyłana do klienta jest tu generowana z kodu aplikacji oraz z zawartości baz danych dostępnych najczęściej za pomocą serwerów lub sieci typu SAN.

Od strony dostawcy niezawodność zależy od obciążenia jego sieci szkieletowej i zapasu obciążenia jakim dysponuje przy obciążeniu typowym i maksymalnym.
Kolejną metodą zwiększania dostępności jest geograficzne równoważenie obciążenia. Metody takie opierają się na usłudze DNS. W tym przypadku używane są specjalne geograficzne urządzenia balansujące, które wybierają odpowiedni adres w zależności od adresu klienta. Składają się one z wielu komponentów rozlokowanych w różnych miejscach sieci. Urządzenie to na podstawie zapytania decyduje, który lokacja z serwerem lub serwerami jest dostępna i najbardziej użyteczna dla danego klienta.

Przechodząc bliżej samej usługi (serwerów) używane są metody round-robin DNS, które wybierają dla danej nazwy DNS na każde żądanie inny adres IP docelowego serwera z usługą. Rozwiązanie takie jest proste lecz równe obciążenie może zapewnić tylko w sensie statystycznym w dłuższym odcinku czasowym. Nie jest ono w stanie uwzględnić faktyczne obciążenie obliczeniowe poszczególnych serwerów oraz ewentualne awarie serwera.

Najbliżej samych serwerów uwieszone może być fizyczne urządzenie równoważące obciążenie serwerów. Urządzenia takie mogą uwzględniać awarię serwera i nie wysyłać tam żądań klientów. Poprzez instalację agentów na serwerach mogą również badać obciążenie poszczególnych serwerów i stosować bardziej zaawansowane algorytmy dystrybucji żądań. W celu sprawdzania stanu zdrowia serwerów układ równoważący może sam nawiązywać połączenia i sprawdzać ich wynik dzięki czemu będzie w stanie sam ocenić działanie poszczególnych serwerów. W takim rozwiązaniu sam układ równoważenia staje się niestety pojedynczym punktem awarii. Dodatkowym problemem w takim rozwiązaniu jest możliwość wykorzystywania prze jednego użytkownika całego szeregu stron z pewnym wspólnym kontekstem np. poświadczeniami logowania. W takim przypadku dana transakcja użytkownika musi być kierowana zawsze do tego samego serwera. Mało wydajnym rozwiązaniem jest przechowywanie tego kontekstu i dostępu dla niego dla wszystkich serwerów. Dużo lepiej sprawdza się metoda podtrzymania sesji, która polega na przechowywaniu informacji o sesji dla danego użytkownika prze

urządzenie równoważące ruch. Sesje mogą być określna na podstawie plików cookie.

Patrząc z perspektywy usługi brak połączenia i przetworzenia żądania dla klienta może być spowodowane dwoma czynnikami[11, 19]:

- awaria serwera spowodowana przez sprzęt lub oprogramowanie,
- brak dostępności serwera w wyniku przekroczenia możliwości serwera reprezentowanej jako bufor serwera.

Przy czym zauważyć należy, że w przypadku awarii serwera pozostałe serwery w klastrze muszą przejąć jego rolę co zwiększa ich obciążenie a zaraz może powodować brak obsługi żądania drugiego typu.

Istnieją rozwiązania, w których aktywne połączenie przy awarii serwera zostaje przejęte przez inny serwer razem ze stanem połączenia. Jest to metoda przywrócenia poprawnego stanu usługi dla klienta, która jest dla niego samego nie widoczna. Ruch przychodzący do serwera jest modelowany procesem Poissona z poziomem λ żądań na sekundę. Maksymalny rozmiar bufora serwera wynosi b, przy czym serwer obsłużyć może μ żądań na sekundę co oznacza że system może być modelowany kolejką $M/M/1/b$.

Ciekawe do rozważania są wartości proponowane w modelu wydajności. Serwer jest w stanie obsłużyć 5 żądań na sekundę, średni czas do awarii usługi wynosi 10 dni, czas wykrycia awarii to 20 s, średni czas naprawy usługi wynosi 200s a rozmiar bufora obsługi żądań serwera wynosi 20. Najważniejszy wniosek z przeprowadzonych badań [12] jest taki, że dostępność usługi nie zależy wiele od ilości serwerów pod warunkiem, że ich ilość jest wystarczająca do obsłużenia napływającego ruchu. Jeśli nie zapewnimy odpowiedniej wydajności serwerów zwiększając ich liczbę nastąpi spadek niezawodności. Kolejną cenną informacją, jest to że wydajność zależy od charakterystyki ilości żądań. Jeśli żądania napływają w sposób wybuchowy może to zmniejszyć wydajność systemu. Z punktu widzenia klienta duże opóźnienie w danej sesji czy to w żądaniu czy w dalszym przetwarzaniu może być rozumiane jako brak dostępności usługi. Jednym z rozwiązań proponowanych w literaturze jest wykorzystanie metod QoS w zmianie dostępności dla usług w zależności od wymagań klientów[11]. Wydaje się, że czas oczekiwania dłuższy od kilku sekund jest od strony klienta nie do zaakceptowania.

Osiągi serwerów a zarazem usług www są zależne od wielu powiązanych ze sobą czynników takich jak: sprzęt serwera, system operacyjny serwera, aplikacja udostępniająca serwis www, połączenie sieciowe, zastosowane metody proxy, rozmiar przesyłanych treści, rozkład wielkości rozmiaru przesyłanych danych, obciążenie serwera, ilość żądań klientów, charakter żądań klientów, obciążenie sieci

związane z innymi usługami. Inną metodą oceny osiągów (wydajności, performance) jest podział na modele związane z poszczególnymi fazami osiągania połączenia (transakcji)[22]:

- połączenie TCP (gniazdo TCP), po czym następuje wysłanie transakcji http,
- połączenie http, przetwarzanie aplikacji,
- model związany z obsługą I/O karty sieciowej serwera,

Modele te są związane z odpowiednimi modelami kolejkowania.

Podstawowym parametrem związanym z osiągami serwera jest czas opóźnienia i jego zależność od ilości żądań klientów. Typową zależność przedstawia rys 1.

Po przekroczeniu wartości krytycznej czas odpowiedzi nie rośnie. Następuje jedna zwiększanie ilości odrzuconych żądań co zmniejsza niezawodność usługi.

Ostanie prace dotyczące wydajności serwerów skupiają się na problemie właściwego modelowania ruchu sieciowego[15]. Wynika z nich podstawowy problem związany z dużą zmiennością i wybuchowością żądań użytkowników. Jako poziom wybuchowości ruchu określa się jego wartość szczytową do wartości średniej. Po przekroczeniu określonej specyficznej dla każdego serwera ilości żądań jego możliwości rozumiane jako ilość transakcji na jednostkę czasu maleje. Niestety znaczący wpływ ma tutaj wybuchowość ruchu. Nawet krótkie chwile z kilkukrotnym przekroczeniem wartości średniej mogą znacznie zmniejszyć wydajność transakcji.

Niezawodność aplikacji Web jest rozumiana jako prawdopodobieństwo wykonania wszystkich operacji związanych obsłużeniem i zakończeniem żądania. Inaczej mówiąc jest to brak awarii podczas wykonywania operacji związanych z sesją określonego użytkownika[21]. Ilość błędów i awarii jest wyraźnie skorelowana z ilością żądań odpowiadającą zwiększeniu obciążenia serwerów[21]. Przykładowe wartości niezawodności usługi na podstawie pomiarów wynoszą R=0,962 a średni czas pomiędzy awariami wynosi MTBF 26,6 obsłużonych żądań. Awarie usługi mogą wynikać z następujących źródeł[21, 18]:

- awarie sieci, przeglądarki, serwera, które mogą być rozpatrywane klasycznym metodami,
- błędy związane z problemami zawartości na serwerach Web i ich przetwarzaniem,
- błędy po stronie użytkownika,

Rozważania na temat niezawodności mają na celu jej podniesienie. Usługodawcom zależy oczywiście na niezawodności odnoszącej się do ich części przetwarzania i przesyłania danych. Na błędy po stronie użytkownika usługodawca nie ma wpływu. Podstawowy parametr niezawodności średni czas pomiędzy uszkodzeniami może być określony za pomocą ilości wszystkich awarii w stosunku

do ilości pracujących węzłów z uwzględnieniem sumarycznego czasu pracy wszystkich serwerów[18, 21]:

$$R = (n - f)/n = 1 - r$$

gdzie:

R - niezawodność według modelu Nelsona,

f - całkowita ilość awarii,

n - ilość pracujących serwerów,

r - intensywność uszkodzeń.

Dla modelu wzrostu niezawodności oprogramowania (SRGM software reliability growth model) niezawodność może być określona parametrem oczekiwanej skumulowanej ilości uszkodzeń (expected cumulative failures)[18, 21] w czasie:

$$m(t) = N(1 - e^{-bt})$$

gdzie:

N - całkowita ilość uszkodzeń w sytemie,

b - stała estymowana na podstawie obserwacji.

W praktycznych środowiskach zachodzi pytanie ile serwerów powinno pracować w sieci by obsłużyć przewidywaną liczbę klientów z odpowiednim poziomem jakości usług. Zazwyczaj ilość serwerów dobierana jest na podstawie statystyk monitorowania pracy serwera przez określony czas. Problemem jest optymalny dobór ilości serwerów tak by ich zasoby były właściwie wykorzystane. Gdyby ilość żądań była równomiernie rozłożona w czasie wybór optymalnej liczby serwerów byłby stosunkowo prosty. Niestety wybuchowe właściwości ilości żądań wymagają w szycie obciążenia dużo większej ilości zasobów niż w pozostałym czasie. Przy czym wydaje się, że w szczycie obciążenia musimy liczyć się z pewnym spadkiem jakości usług gdyż inaczej należało by stosować dużą nadmiarowość, która z innych względów była by nie do zaakceptowania.

W systemach rozproszonych optymalizacja oprócz ilości serwerów uwzględniać musi właściwe ich rozmieszczenie. Ciekawym zagadnieniem jest w przypadku świadczenia kilku usług dystrybuowanie obciążenia pomiędzy nimi. Zagadnienie to jest ciągle rozwijane chociażby w systemach typu cloud computing gdzie takie metody są rozważane. W przypadku systemu z równoważeniem obciążeń znaną właściwością jest zmniejszanie się wydajności systemu wraz ze zwiększaniem się ilością serwerów. Jeśli wydajność systemu rozumiana jako wydajność serwerów pomnożona przez ich liczbę jest taka sama to lepszą wydajność całościową będzie posiadał system z mniejszą ilością serwerów o większej mocy[19]. Jest to związane ze spadkiem możliwości podsystemu równoważenia obciążenia. Rozważając jednak rozwiązanie pod względem tolerancji

awarii system z większą ilości serwerów jest lepszy. Mała ilość serwerów jest trudniej skalowalna. Przestój serwera związany z jego obsługą jest bardziej widoczny niż dla większej jej liczby.

Dużo bardziej skomplikowana jest analiza wydajności serwera z uwzględnieniem jego interakcji z serwerami baz danych. W mniejszych rozwiązaniach razem z aplikacją web znajdować się może sama baza danych. Pewnym zaskoczeniem jest również to, że dla serwerów z łączoną zawartością statyczną i dynamiczną pod względem obsługi żądań dodatek zawartości dynamicznej nie zmniejsza ilości możliwych do obsługi żądań dla konkretnego serwera [14]. Autorzy wnioskują, że jest to wynik lepszego wykorzystania serwera związanego z większym wykorzystaniem procesora ale mniejszym układów I/O. Te podsystemy sprzętowe pracują równolegle. Przy dynamicznej zawartości znacznie większe znaczenie ma ilość procesorów i całkowita moc obliczeniowa serwera niż przepustowość I/O. W przypadku obciążenia przekraczającego możliwości pewnym rozwiązaniem może być brak akceptacji kolejnych żądań tak by te zaakceptowane zostały obsłużone na akceptowanym poziomie jakości.

Ze względu na duże dobowe zmiany obciążenia serwerów Web w usługach typu Cloud Computing, kupowanych przez klientów od dużych firm stosowane są metody zarządzania zasobami poprzez ich dostosowanie do aktualnych potrzeb. Metody te można ogólnie nazwać metodami adaptacyjnego skalowania zasobów, gdzie przykładowym parametrem branym pod uwagę jest czas odpowiedzi serwerów wirtualnych[16]. Dodatkowo zauważyć należy, że w przypadku wirtualizacji wydajność maszyn wirtualnych wraz ze wzrostem obciążenia maleje dodatkowo w wyniku samej obecności innych maszyn wirtualnych głównie w wyniku narzutu związanego z samą wirtualizacją, połączeń z rdzeniem CPU oraz połączeń z pamięcią[17]. Tematem badań są również metody uwzględnienia awarii maszyn wirtualnych w dynamicznym zarządzaniu zasobami. Metoda ta polega na dodatkowym monitorowaniu działania maszyn wirtualnych oraz przenoszeniu zadań pomiędzy nimi [13].Techniką, która pozwala bardzo aktywnie działać na tym polu jest możliwość migracji maszyn wirtualnych pomiędzy serwerami fizycznymi bez konieczności zatrzymywania ich pracy.

Wydaje się, że atrakcyjnym rozwiązaniem są klastry używane jednocześnie przez kilka usług. Rozwiązanie tego typu wraz z wirtualizacją pozwala przenosić zasoby do usług w zależności od zapotrzebowania. Przy czym usługi należy podzielić pod względem wrażliwości odnośnie jakości. Zasoby należy kierować uwzględniając najpierw potrzeby najbardziej jakościowo wrażliwej usługi i dalej według wyznaczonej hierarchii. Przy czym pożądaną sytuacją jest obsługa jeden usługi z QoS oraz pozostałych, których czas wykonania zadania nie jest krytyczny. Wydaje

się, że zarządzanie zasobami opierać się może o statystyczne modele obciążenia uwzględniające pomiar on-line wybranych parametrów. W literaturze proponowane są metody podziału zasobów na przypisane na stałe do usług niezależnie od obciążenia oraz zasoby zależne od obciążenia[20].

Chmury to przede wszystkim dynamiczne i autonomiczne zarządzania zasobami w centrum danych. Chmura zapewnia mapowanie pomiędzy środowiskiem aplikacji a zbiorem współdzielonych zasobów[23]. Proces jest dynamiczny ze względu na zmiany obciążenia oraz zmiany w dostępnych zasobach. A celem jest zawsze zapewnienie odpowiednich osiągów i zapewnienie poziomu usług. Konsolidacja zasobów w sensie zarządzania centrum danych może być rozpatrywana następująco[23]:

- konsolidacja zasobów w centrum danych z określoną ilością serwerów – za pomocą dwu warstwowej architektury bez wirtualizacji
- konsolidacja zasobów na jednym dużym serwerem.

Literatura

1. S. Abeck at all: Network Management Know It All, Morgan Kaufmann Publishers, Elsevier, 2009,
2. W. Stallings: Protokoły SNMP I RMON Vademecum profesjonalisty, Helion, 2003,
3. NetFlow Services and Applications, White Paper, Cisco Systems, 1999, http://mauigateway.com/~surfer/library/netflow_wp.pdf,
4. NetFlow Services Solution Guide, http://trendmap.net/support/wp/nfwhite.pdf,
5. R. Boutaba, I. Aib: Policy-based Management: A Historical Perspective, J Netw Syst Manage 15, 2007, pp. 447-480,
6. F. Caldeira, E. Monteiro: Policy-based networking: applications to firewall management, Ann. Telecommun. 59 no 1-2, 2004, pp. 38-54.
7. J. Davies, T. Northrup, Microsoft Networking Team: Microsoft Windows Server 2008 Ochrona dostępu do sieci (NAP), APN Promise, Warszawa, 2008.
8. Architektura platformy NAP. http://technet.microsoft.com/pl-pl/library/architektura-platformy-network-access-protection.aspx (sierpień 2011).
9. Windows Server Essential, http://www.techotopia.com/index.php /Windows_Server_2008_Essentials (sierpień 2011).
10. Microsoft Corporation: Step By Step Guide: Demonstrate DHCP NAP Enforcement in a Test Lab, February 2008, www.microsoft.com

11. H. Chen, P. Mohapatra: Overload control in QoS-aware web servers, Computer Networks 42, 2003, pp. 119-133,
12. M. Martineloo, M. Kaaniche, K. Kanoun: Web service availability – impact of error recovery and traffic model, Reliability Engineering and System Safety 89, 2005, pp. 6-16,
13. S. Fu: Failure-aware resource management for high-availability computing clusters with distributed virtal machines, J. Parallel Distrib. Comput 70, 2010, pp. 384-393,
14. X. He, Q. Yang: Performance Evaluation of Distributed Web Server Architectures under E-Commerce Workloads, In Proc. International Conference on Internet Computing, 2000, pp. 285-292,
15. E. Hernandez-Orallo, J. Vila-Carbo: Web server performance analysis using histrogram workload models, Computer Networks 53, 2009, pp. 2727-2739,
16. W. Iqbal, M. Dailey, D. Carrera: SLA-Driven Adaptive Resource Management for Web Application on a Heterogeneous Compute Cloud, in M.G. Jaatun, G. Zhao, C. Rong (Eds.): CloudCom 2009, Springer-Verlag, Berlin Heidelberg, pp. 243-253,
17. R. Iyer, R. Illikkali, O. Tickoo, L. Zhao, P. Apparao, D. Newell: VM3: Measuring, modeling and managing VM shared resources, Computer Networks 53, 2009, pp. 2873-2887,
18. R. Manjula, E.A. Sriram: Reliability evaluation of web applications from click-stream data, International Journal of Computer Applications, Vol. 9 No. 5, 2010, pp. 23-29,
19. J.H Son, M.H. Kim: An analysis of the optimal number of servers in distributed client/server environments, Decision Support Systems 36, 2004, pp. 297-312,
20. M. Stillwell, D. Schanzenbach, F. Vivien, H. Casanova: Resource allocation algorithms for virtualized service hosting platforms, J. Parallel Distrib. Comput. 70, 2010, pp. 962-974,
21. J. Tian, Z. Li.: Evaluating web software reliability based on workload and failure data extracted from server logs, IEEE Transactions on Software Engineering, Vol. 30, No. 12, 2004, pp.
22. R.D. Van Der Mei, R. Hariharan, P.K. Peeser: Web serwer performance modeling, Telecommunications System 16, 2001, pp. 361-378,
23. Y.O. Yazir, C. Matthews, R. Farahbod, S. Neville, A. Guitouni, S. Ganti, Y. Coady: Dynamic Resource Allocation in Computing Clouds through Distributed Multiple Criteria Decision Analysis, Cloud Computing 2010 IEEE 3rd International Conference, pp. 91-98

7. Sztuczna inteligencja i systemy ekspertowe

7.1. Wprowadzenie

Początków sztucznej inteligencji można doszukiwać się w odległych wiekach, nawet u starożytnych filozofów, biorąc pod uwagę filozoficzne aspekty tej nauki. W niej odległej czasowo pierwszej połowie XIX wieku – Charles Babbage zaproponował koncepcję maszyny analitycznej, a w 1950 roku Alan Turing zaproponował test mający stwierdzić, czy dany program jest inteligentny. W początkowym okresie, przy braku formalnych podstaw nauki sztucznych sieci neuronowych zainteresowania badaczy osadzonych w dziedzinie neurofizjologii czy bioniki skierowane były na opisie mechanizmów działania mózgu czy pojedynczych komórek układu nerwowego. Niezależnie, dziedzina sieci neuronowych zaistniała wraz z wydaniem pracy [21], w której po raz pierwszy wprowadzono matematyczny opis komórki neuronowej oraz podjęto próbę określenia zasad przetwarzania informacji opartego na kanwie modelu sztucznego neuronu. Niedoścignionym wzorem i źródłem inspiracji dla kolejnych pokoleń badaczy były pozycje [38] oraz [36], będące kanonem wyznaczającym rozwój wiedzy o sieciach neuronowych, odpowiednio od strony rozwiązań technicznych, źródłem wiadomości biologicznych.

Pierwszym szeroko znanym przykładem modelu neuronu jako prostej sztucznej sieci neuronowej był perceptron [30]. Sieć ta, jako układ częściowo elektromechaniczny a częściowo elektroniczny została zbudowana w 1957 w Cornell Aeronautical Laboratory i miała za zadanie rozpoznawać znaki. Programowanie tego rozwiązania oparto na zasadach prostego uczenia, co stanowiło o wielkim kroku na przód w dziedzinie sieci neuronowych. Po ogłoszeniu wyników przez twórców nastąpił gwałtowny rozwój tego typu sieci neuronowych na całym świecie. Wśród naśladowców pierwotnego rozwiązania znaleźli się i tacy, którzy twórczo przekształcili pomysły Rosenblatta i Wightmana.

Istotnym przykładem jest rozwiązanie zaproponowane przez Bernarda Widrowa sieć Madaline, zbudowana w 1960 roku. Składała się ona z pojedynczych elementów Adaline (ang. Adaptive linear element), które powielone oraz połączone pozwoliły uzyskać układ Madaline (ang. Many Adaline) [41], [42].

Tempo rozwoju badań nad problematyką sztucznych sieci neuronowych zostało gwałtownie zahamowane na początku lat 70-tych po publikacji książki [24], która zawierała formalny dowód, że sieci jednowarstwowe (podobne do perceptronu) mają bardzo ograniczony zakres zastosowań. Przez ponad dekadę utrzymywał się stan impasu, do ukazania się serii publikacji, wykazujących, że sieci nieliniowe

wolne są od ograniczeń pokazanych w pracy [24]. Jednocześnie mniej więcej w tym czasie ogłoszono kilka bardzo efektywnych, formalnych przepisów na uczenie sieci wielowarstwowych.

Okres lat 70-tych nie jest jednak zupełnie bezpłodny jeśli chodzi o tworzone nowe konstrukcje sieci neuronowych. Wymienić tu należy, chociażby zbudowaną przez Stephena Grossberga na uniwersytecie w Bostonie sieć Avalanche. Służyła ona do rozpoznawania mowy oraz sterowaniem ramieniem robota. Z kolei w MIT powstaje Cerebellatron skonstruowany przez badaczy Davida Mara, Jamesa Albusa i Andresa Pollioneze, służący także do sterowania robota. Odmienne zastosowanie miała sieć Brain State in the Box, zbudowana przez Jamesa Andersona z Uniwersytetu Browna w 1977 roku. Funkcjonalnie była odpowiednikiem pamięci asocjacyjnej z dwustronnym dostępem (BAM), ale jej działanie nie było związane z iteracyjnym procesem poszukiwania, lecz polegało na szybkich zależnościach typu wejście - wyjście.

Opracowanie technologii wytwarzania sztucznych modeli komórek nerwowych w postaci układów scalonych przyniosło w latach 80-tych pierwsze konstrukcje o dużych rozmiarach oraz znaczących mocach obliczeniowych.

W tym czasie, zaistniały pierwsze sieci ze sprzężeniami zwrotnymi, których przykładem jest opracowana przez Johna Hopfielda sieć wykorzystywana do odtwarzania obrazów z ich fragmentów, a także stosowana do rozwiązywania zadań optymalizacyjnych.

Renesans szerokiego zainteresowania tematyką sieci neuronowych datuję się na drugą połowę lat 80-tych, kiedy to J.A. Anderson publikuje pracę [3], a J.J.Hopfield wprowadza rekurencyjną architekturę pamięci neuronowych i opisuje w swoich pracach obliczeniowe własności sieci ze sprzężeniem zwrotnym. Z kolei wzrost zainteresowania możliwościami sieci warstwowych przyniosło publikowanie [20] przez McClellanda i Rumelharta monografii o równoległym przetwarzaniu rozproszonym. W latach 1986-87 wzrasta liczba zagadnień, które można rozwiązać przy użyciu sieci neuronowych oraz obserwowana jest znaczna liczba projektów badawczych w tej dziedzinie. Powoduje rozszerzanie się zakresu ich zastosowań [8].

Na początku lat 90. ideę uczenia sieci neuronowych zaadoptowano do uczenia systemów rozmytych. W ten sposób powstały struktury neuronowo-rozmyte i zaproponowano różne kombinacje sieci neuronowych, systemów rozmytych i algorytmów ewolucyjnych. W ten sposób wydzieliła się oddzielna gałąź nauki, określana w literaturze anglojęzycznej Computational Intelligence, znane w nomenklaturze polskiej jako inteligencja obliczeniowa. Termin ten rozumiany jest jako rozwiązywanie różnego rodzaju problemów sztucznej inteligencji

z wykorzystaniem komputerów wykonujących obliczenia numeryczne [31], związanych z zastosowaniem następujących technik: sieci neuronowe [35, 44], logika rozmyta [16, 40], algorytmy ewolucyjne [14, 21], zbiory przybliżone [28, 29], zmienne niepewne [4, 6, 7].

Opracowane teorie inteligencji obliczeniowej szybko znalazły liczne zastosowania w wielu dziedzinach inżynierii, analizy danych, prognozowania, w biomedycynie i innych. Stosuje się je w obróbce i rozpoznawaniu obrazów oraz dźwięków, przetwarzaniu sygnałów, wizualizacji wielowymiarowych danych, sterowaniu obiektami, analizie danych leksykograficznych, systemach wnioskujących w bankowości, systemach diagnostycznych, ekspertowych i wielu praktycznych wdrożeniach.

Istotą systemów opartych na inteligencji obliczeniowej jest przetwarzania i interpretacja danych o zróżnicowanym charakterze. Mogą to być dane numeryczne, symboliczne, binarne, logiczne lub odczytane wprost, niezakodowane obrazy. Mogą one również składać się z uporządkowanych sekwencji elementów lub tablic i zawierać elementy opisane w bardzo nieprecyzyjny, lub wręcz subiektywny sposób.

Wspólną cechą systemów inteligencji obliczeniowej jest przetwarzanie przez nie informacje w przypadkach trudnych do przedstawienia w postaci algorytmów i czynią to w powiązaniu z symboliczną reprezentacją wiedzy. Mogą to być relacje dotyczące jakiegoś obiektu znanego tylko na podstawie skończonej liczby pomiarów stanu wyjścia i wejścia (pobudzennia). Mogą to być również dane wiążące najbardziej prawdopodobna diagnozę z szeregiem zaobserwowanych symptomów, często opisowych. W innych przypadkach mogą to być dane charakteryzujące zbiory względem ich pewnych, specjalnych cech, które są dla użytkownika początkowo nieuchwytne, aż do momentu ich wydobycia z danych i określenia jako cech dominujących. Systemy te mają zdolność odtwarzania zachowań zaobserwowanych wciągach uczących, potrafią formułować reguły wnioskowania i generalizować wiedzę w sytuacjach, kiedy oczekuje się od nich predykcji, bądź zaklasyfikowania obiektu do jednej z zaobserwowanych wcześniej kategorii.

7.2. Sztuczne Sieci Neuronowe

Naukowcy od dawna próbują poznać budowę i sposób działania mózgu, ale wciąż pozostaje ona nie do końca odkrytą, fascynującą zagadką. Jako obiekt badań, sztuczne sieci neuronowe stanowią bardzo uproszczony model rzeczywistego biologicznego systemu nerwowego. Składają się one z połączonych ze sobą neuronów, a ich istotną cechą jest możliwość uczenia się - to jest modyfikowania

parametrów charakteryzujących poszczególne neurony w taki sposób, by zwiększyć efektywność sieci przy rozwiązywaniu zadań określonego typu.

Sieci neuronowe mogą być bardzo skuteczne jako narzędzia obliczeniowe, zwłaszcza w rozwiązywaniu zadań, z którymi typowe komputery (programy) sobie nie radzą. Wynika to po pierwsze z faktu, że obliczenia są w sieciach neuronowych wykonywane równolegle, w związku z czym szybkość pracy sieci neuronowych może znacznie przewyższać szybkość obliczeń sekwencyjnych. Drugą zaletą sieci jest możliwość uzyskania rozwiązania problemu z pominięciem etapu konstruowania algorytmu rozwiązania problemu.

Sieci nie trzeba programować. Istnieją metody uczenia i samouczenia sieci pozwalające uzyskać ich celowe i skuteczne działanie nawet w sytuacji, kiedy twórca sieci nie zna algorytmu, według którego można rozwiązać postawione zadanie. W sieciach neuronowych zarówno program działania oraz informacje stanowiące bazę wiedzy, dane, jak i sam proces obliczania są całkowicie rozproszone.

Sieć działa zawsze jako jednolita całość, a wszystkie jej elementy mają wkład w realizację działań/czynności, które realizuje. Istotną tego konsekwencją jest jej zdolność do poprawnego działania nawet po wystąpienia uszkodzenia znacznej części wchodzących w jej skład elementów.

Struktura sieci powstaje w ten sposób, że wyjścia jednych neuronów łączy się z wejściami innych. Z jednej strony, od rodzaju stawianego zadania, wynika konkretna topologia sieci. Z drugiej zaś strony, decyzje dotyczące struktury sieci nie wpływają na jej zachowanie w stopniu decydującym. Zachowanie sieci determinowane jest przede wszystkim przez proces uczenia. Sieć musi jednak mieć wystarczający stopień złożoności, żeby w jej strukturze można było w toku uczenia "wykrystalizować" potrzebne połączenia i struktury. Zbyt mała sieć może mieć zbyt mały "potencjał intelektualny" do realizacji określonego zadania.

Sztuczne neurony charakteryzują się występowaniem wielu wejść i jednego wyjścia. Sygnały wejściowe xj (i = 1, 2 n) oraz sygnał wyjściowy y mogą przyjmować wartości, odpowiadające pewnym informacjom. W ten sposób zadanie sieci sprowadzone do funkcjonowania jej podstawowego elementu polega na tym, że neuron przetwarza informacje wejściowe xj na pewien wynik y.

Każdy neuron dysponuje pewną wewnętrzną pamięcią (reprezentowaną przez aktualne wartości wag i progu) oraz pewnymi możliwościami przetwarzania wejściowych sygnałów w sygnał wyjściowy.

Z powodu bardzo ubogich możliwości obliczeniowych pojedynczego neuronu - sieć neuronowa może działać wyłącznie jako całość. Wszystkie możliwości

i właściwości sieci neuronowych są wynikiem kolektywnego działania bardzo wielu połączonych ze sobą elementów (całej sieci, a nie pojedynczych neuronów).

Cykl działania sieci neuronowej podzielić można na etap nauki, kiedy sieć gromadzi informacje potrzebne jej do określenia, co i jak ma robić, oraz na etap normalnego działania (nazywany czasem także egzaminem), kiedy w oparciu o zdobytą wiedzę sieć musi rozwiązywać nowe, konkretne zadania. Możliwe są dwa warianty procesu uczenia: z nauczycielem i bez nauczyciela.

Uczenie z nauczycielem polega na tym, że sieci podaje się przykłady poprawnego działania, które powinna ona potem naśladować w swoim bieżącym działaniu (w czasie egzaminu). Przykład należy rozumieć w ten sposób, że nauczyciel podaje konkretne sygnały wejściowe i wyjściowe, pokazując, jaka jest wymagana odpowiedź sieci dla pewnej konfiguracji danych wejściowych. Występuje tu para wartości - przykładowy sygnał wejściowy i pożądanym (oczekiwanym) wyjściem, czyli wymaganą odpowiedzią sieci na ten sygnał wejściowy. Zbiór przykładów zgromadzonych w celu ich wykorzystania w procesie uczenia sieci nazywa się ciągiem uczącym. Stąd, w typowym procesie uczenia sieć otrzymuje od nauczyciela ciąg uczący i na jego podstawie uczy się prawidłowego działania, stosując jedną ze znanych strategii uczenia.

Obok opisanego wyżej schematu uczenia z nauczycielem występuje też szereg metod tak zwanego uczenia bez nauczyciela (albo samouczenia sieci). Metody te polegają na podawaniu na wejście sieci wyłącznie szeregu przykładowych danych wejściowych, bez podawania jakiejkolwiek informacji, dotyczącej pożądanych czy chociażby tylko oczekiwanych sygnałów wyjściowych. Odpowiednio zaprojektowana sieć neuronowa potrafi wykorzystać same tylko obserwacje wejściowych sygnałów i zbudować na ich podstawie sensowny algorytm swojego działania - najczęściej polegający na tym, że automatycznie wykrywane są klasy powtarzających się sygnałów wejściowych i sieć uczy się (spontanicznie, bez jawnego nauczania) rozpoznawać te typowe wzorce sygnałów.

Jednym z istotnych niebezpieczeństw przy konstruowaniu zbiorów jest nieświadome uczenie sieci ludzkich błędów systematycznych. Sieć uczy się takich błędów tak samo jak innych wzorców. Jeśli ciąg uczący obciążony jest błędami systematycznymi, może to doprowadzić do sytuacji, gdy nauczona sieć działająca poprawnie na takim ciągu uczącym, zawiedzie przy pracy na danych nieobciążonych błędem systematycznym.

Zastosowania metod inteligencji obliczeniowej ograniczone są często do problemów, którymi zajmuje się rozpoznawanie struktur (pattern recognition) [33]. Większość prac skupia się przy tym nad zagadnieniami zdefiniowanymi w ramach paradygmatu przestrzeni cech, określającej własności obiektów. Sieci neuronowe

potrzebują danych w postaci wektorów liczb o ustalonej liczbie składowych. Odpowiada to funkcjom kory zmysłowej, podświadomym mechanizmom rozpoznawania podstawowych cech obiektów, wykrywaniu cech wyższego rzędu i kategoryzacji na tej podstawie. Mózgi zajmują się wyłącznie sygnałami, mającymi strukturę czasoprzestrzenną, sekwencjami sygnałów, podczas gdy metody CI najczęściej danymi statycznymi.

Tymczasem wiele problemów nie da się w ogóle przedstawić w tej postaci. Należą do nich zagadnienia wymagające złożonych metod reprezentacji wiedzy, opis obiektów o zmiennej strukturze (organizacji, przedsiębiorstw, cząsteczek chemicznych), sekwencji symboli (liter, wyrazów, zdań, par zasad DNA lub aminokwasów białek), zmieniającego się stanu obiektów (pacjenta, gier planszowych, gier wojennych). Niektóre z tych zagadnień wchodzą w zakres zainteresowań sztucznej inteligencji. Niezwykle ambitne projekty, takie jak General Problem Solver [26, 43, 44], od początku wytyczyły w tej dziedzinie dobrze określone cele. Stworzenie programu wykazującego się ogólną inteligencją okazało się bardzo trudne, jednakże również inteligencja ludzka nie okazała się tak uniwersalna, jak początkowo sądzono. Uczenie się rozwiązywania problemów w jednym kontekście nie prowadzi automatycznie do osiągnięcia lepszych rezultatów dla podobnych problemów w odmiennym kontekście [2]. W dobrze określonej dziedzinie daje się utworzyć ontologie zawierające opis używanych pojęć i utworzyć bazę wiedzy w oparciu o powiązania miedzy nimi. Przykładem systemu, którego kompetencje znacznie przewyższają możliwości ludzkiego intelektu jest EcoSys [17], zawierający oparty na regułach produkcji model procesów metabolicznych i genetycznych zachodzących w bakterii Escherische Coli.

Takie zastosowania stawiają przed inteligencją obliczeniową szereg wyzwań obejmujących między innymi sposób wykorzystania wiedzy zdobytej w oparciu o analizę danych do systematycznego rozumowania. Stworzenie systemu do wspomagania diagnoz medycznych to jedynie pierwszy krok do planowania i monitorowania terapii. Takie działania wymagają rozważenia szeregu wariantów, a więc procesów szukania optymalnych rozwiązań. Najłatwiej jest je wykonać w systemach opartych na regułach. Jeśli z danych można wyciągnąć niewielką liczbę stosunkowo prostych reguł to da się je wykorzystać w algorytmie planującym. Zrozumienie danych, zarówno w sensie odkrywania reguł logiki klasycznej lub rozmytej, lub też szukania prototypów wystarczających do kategoryzacji przez podobieństwo, nie było dotychczas celem statystyki. W tym celu zastosować można wiele metod inteligencji obliczeniowej [9].

Trudne do określenia jest jakiej metody użyć, jeśli liczba cech, istotnych dla opisu danych z analizowanej bazy nie jest ustalona (i nie można się posłużyć

paradygmatem przestrzeni wektorowej). W niektórych przypadkach, problem może zostać pomyślnie przeanalizowany w kilku przestrzeniach (np. wstępnych testów po których nastąpią bardziej zaawansowane testy różnego rodzaju, zależnie od wyników oceny początkowych testów). Niezbędne wówczas są różne modele, służące do otrzymania końcowego rezultatu. Nie zawsze wystarczające jest jednak posiadanie różnych modele, służące do otrzymania końcowego rezultatu. Cząsteczki chemiczne można w bardzo uproszczony sposób zapisać w postaci grafów, których struktury da się analizować za pomocą sieci z rekurencją [13]. W nieco bardziej ogólny sposób można zdefiniować operatory przekształcające obiekty lub stany opisu problemu w siebie i obliczyć podobieństwa jako sumę kosztów elementarnych operacji. W tym przypadku koszty mogą być parametrami adaptacyjnymi, pozwalającymi na upodobnienie obiektów należących do tej samej klasy do siebie [19]. Mając daną macierz podobieństw można do niej zastosować wiele metod klasyfikacji, np. metody oparte na podobieństwie lub analizę dyskryminacyjną Fishera [8].

W realnych sytuacjach, na ogół istnieją wyłącznie powiązania elementów, regularności łączące kilka zmiennych, które można się nauczyć na prostych przykładach. Jednak zagadnienie stanowi korzystanie z wiedzy na temat podproblemów przy rozwiązywaniu złożonego zadania. Ekspert analizując różnego rodzaju problemy korzysta w intuicyjny sposób z takiej wiedzy, prowadząc dłuższe rozumowanie. Nawet, jeśli możliwy jest opis problemu w przestrzeni cech, to początkowo znane są tylko nieliczne z nich, a brakujące elementy są uzupełnianie na podstawie fragmentarycznej wiedzy. Wykorzystanie takiej wiedzy jako heurystyk pozwala uniknąć eksplozji kombinatorycznej w procesach szukania rozwiązań [10, 11, 27].

7.3. Systemy Rozmyte

Celem wprowadzenia pojęcia i teorii zbiorów rozmytych była potrzeba matematycznego opisu zjawisk i pojęć, które maja charakter wieloznaczny i nieprecyzyjny. W teorii tej można mówić o częściowej przynależności punktu do rozważanego zbioru.

Zamiast zdaniami przyjmującymi wartości logiczne prawda lub fałsz używane są zmienne lingwistyczne. Na systemy rozmyte składają się techniki i metody, służące do obrazowania informacji nieprecyzyjnych, nieokreślonych bądź niekonkretnych. Pozwalają one opisywać zjawiska o charakterze wieloznacznym, których nie są ujmowane w teorii klasycznej i logice dwuwartościowej. Charakteryzują się tym, że wiedza jest przetwarzana w postaci symbolicznej i zapisywana w postaci rozmytych reguł. Systemy rozmyte znajdują zastosowanie wszędzie tam, gdzie brak jest

wystarczającej wiedzy o modelu matematycznym danego zjawiska oraz tam gdzie odtworzenie tegoż modelu staje się niemożliwe lub nieopłacalne.

Za pomocą zbiorów rozmytych istnieje możliwość formalnego określenia pojęć nieprecyzyjnych i wieloznacznych. Przed podaniem definicji zbioru rozmytego należy ustalić tzw. obszar rozważań (ang. the uniwerse of discourse), który w dalszej części nazywać będziemy przestrzenią lub zbiorem. Obiektami tej abstrakcyjnej przestrzeni będą zbiory rozmyte. Własność, która przypisuje elementowi przynależność do tego zbioru może być ostra lub nieostra. W pierwszym przypadku jest to wówczas zbiór klasyczny, w drugim - zbiór rozmyty. Tak więc każdy element zbioru rozmytego może do niego należeć, do niego nie należeć lub też należeć w pewnym stopniu [31, 32].

Ze zbiorem rozmytym nierozerwalnie związane jest pojęcie funkcji przynależności (ang. membership function). Odzwierciedlają one na obiektach z przestrzeni rozważań, uporządkowanie wprowadzone przez skojarzenie ze zbiorem pewnej własności. Wartość funkcji przynależności na danym elemencie określa stopień jego należenia do tego zbioru. Operacja przypisywania stopnia przynależności jest raczej subiektywna i zależy od kontekstu sytuacyjnego.

Te ogólne rozważania uzasadniają wprowadzenie definicji

Definicja. Zbiorem rozmytym pewnej (niepustej) przestrzeni X , przy zapisie $A \in \boldsymbol{X}$określa się zbiór

$$A = \{(x, \mu_A(x)); x \epsilon \boldsymbol{X}, \mu_A(x) \epsilon [0,1]\}, \tag{7.1}$$

Jeżeli A jest zbiorem klasycznym, to μ_A z powyższej definicji jest funkcją charakterystyczną tego zbioru.

Najważniejsze pojęcie logiki rozmytej, stanowi definicja zmiennej lingwistycznej.

Definicja. Zmienną lingwistyczną określana jest zmienna, której wartościami są słowa lub zdania w języku naturalnym lub sztucznym. Powyższe słowa lub zdania nazywane są wartościami lingwistycznymi zmiennej lingwistycznej [31].

Dla przykładu: Ta sosna jest wysoka . czy Ta książka jest droga

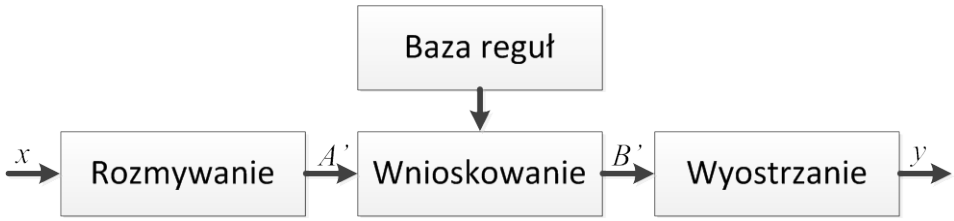
Pojmowanie wartości lingwistycznych (wysoka sosna, czy droga książka) może inne dla różnych osób. Dlatego każdą taka wartość charakteryzowana jest zbiorem rozmytym, np. przedziałem liczbowym określającym wysokość drzewa wyrażoną w metrach, lub cenę książki w EUR.

Zbiory rozmyte i funkcje przynależności dostarczają możliwości reprezentacji nieprecyzyjnych stwierdzeń. Jednak, odwzorowanie wiedzy na zbiór rozmyty jest bezużyteczne przy braku zbioru operatorów, które przekształcają posiadane informacje i wpływają na ich stan.

Operatory rozmyte są uogólnieniem tradycyjnych operatorów boolowskich w teorii zbiorów takich jak fi, U, C, itd.

Opisywanie otaczającego świata za pomocą tradycyjnych metod matematycznych nastręcza wiele kłopotów. Jednocześnie należy zauważyć, że wiele zjawisk można określić przy pomocy języka naturalnego z dostatecznie dobrą dokładnością. Zbiory rozmyte są narzędziem, które umożliwia za pomocą formalnych technik opis zjawisk przy użyciu określeń jakościowych. Podstawowe założenia dotyczące teorii zbiorów rozmytych zostały sformułowane w 1965 roku przez L.A Zadeha. Dzięki stworzonej teorii, wyrażenia z języka potocznego np: duży, wysoki, średni, mały, niski, itp. mogą być zapisane formalnie w postaci odpowiednio skonstruowanych zbiorów rozmytych. Wprowadzone zostało również pojęcie wnioskowania przybliżonego, które jest odpowiednikiem wnioskowania z logiki tradycyjnej. Wnioskowanie z użyciem logiki rozmytej pozwala precyzyjniej modelować sposób wnioskowania człowieka, który ze względu na swoje potencjalne możliwości jest wzorem do naśladowania. Wnioskowanie przybliżone, co do ogólnych zasad jest podobne do wnioskowania z logiki tradycyjnej. Różnice tkwią w budowie reguł oraz w pewnej modyfikacji algorytmu wnioskowania. W celu umożliwienia opisywania zjawisk określeniami znanymi z języka naturalnego wprowadzono pojęcie zmiennej lingwistycznej, która jako wartości może przyjmować określenia np: mały, duży, średni, bliski wartości x, które reprezentowane są przez odpowiednie zbiory rozmyte. Baza reguł systemu wnioskowania przybliżonego składa się z reguł typu JEŻELI TO, przy czym zmienne występujące w przesłankach tych reguł oraz ich konkluzjach są zmiennymi lingwistycznymi [31].

Ze względu na budowę reguł rozmytych proces wnioskowania przybliżonego może zostać przeprowadzony tylko na podstawie informacji w postaci rozmytej. Z tego względu dane wejściowe przed wykorzystaniem w procesie wnioskowania przybliżonego należy poddać procesowi rozmywania. Jednocześnie wyniki uzyskane w procesie wnioskowania są również rozmyte i dalsze ich wykorzystanie wymaga zastosowania operacji wyostrzania, czyli zdefiniowania konkretnej wartości danej zmiennej procesowej. Konieczność wykonania operacji rozmywania na danych wejściowych oraz operacji wyostrzania na wynikach powoduje, że proces wnioskowania przybliżonego dany jest schematem przedstawionym na rysunku 7.1.



Rys. 7.1. Schemat blokowy procesu wnioskowania przybliżonego

Podstawą wnioskowania w rozmytej logice jest tautologia Uogólniony Modus Ponens, natomiast przykład wnioskowania na podstawie reguł dany jest w tabeli 7.1.

Tabela 7.2. Przykład wnioskowania

Pojęcie	Zapis logiczny	Przykład
Przesłanka 1(fakt)	x = A’	Fakt: Pomidor jest prawie czerwony
Przesłanka 2 (implikacja)	JEŚLI x = A TO y = B	Reguła: Jeżeli pomidor jest czerwony to jest dojrzały
Wniosek	y = B’	Wniosek: Pomidor jest prawie dojrzały

gdzie A’, B’ oznacza „bliski A”, „bliski B” odpowiednio, A, A’, B, B’, - zbiory rozmyte, x, y – zmienne lingwistyczne

Należy zauważyć istotną różnicę w działaniu zwykłej reguły w porównaniu z regułą rozmytą. W obu przypadkach występuje relacja implikacji $A \rightarrow B$. Jednakże reguła nierozmyta pozwala na wyznaczanie wniosku tylko wtedy, gdy zdanie A występujące w przesłance jest również obecne w implikacji $A \rightarrow B$. Natomiast w regule rozmytej przesłanka opisana zbiorem rozmytym A’ może być zbliżona do zbioru A z poprzednika implikacji, ale nie musi być koniecznie równa A. Mając dany schemat wnioskowania taki jak na rysunku 7.1, aby wyznaczyć zbiór B’ opisujący wniosek należy przeprowadzić operację kompozycji o postaci:

$$B' = A' \circ (A \rightarrow B). \tag{7.2}$$

Dla przedstawionej operacji wzór na funkcję przynależności zbioru rozmytego B’ przyjmuje następującą postać:

$$\mu B'(y) = \sup\{T[\mu A'(x), \mu A \rightarrow B(x,y)]\}. \tag{7.3}$$

Szczegółowy wzór na funkcję przynależności zbioru B' zależy od przyjętej T-normy oraz od sposobu realizacji implikacji. Poniżej zaprezentowano najczęściej wykorzystywane sposoby przedstawiania funkcji przynależności dla implikacji A→B:

Reguła Mamdaniego:

$$\mu A \rightarrow B(x,y) = \min[\mu A(x), \mu B(y)], \tag{7.4}$$

Reguła Larsena:

$$\mu A \rightarrow B(x,y) = \mu A(x) * \mu B(y). \tag{7.5}$$

Najczęściej stosowane T-normy dane są następującymi wzorami (6- 10):

$$T(x, y) = \min(x,y), \tag{7.6}$$

$$T(x, y) = x * y. \tag{7.7}$$

T-norma Zadeha:

$$T(x,y) = \min\{x,y\} \text{ i } x,y \in [0,1], \tag{7.8}$$

T-norma Mengera:

$$T(x,y) = x \bullet y \quad \text{ i } x,y \in [0,1], \tag{7.9}$$

T-norma Łukaszewicza:

$$T(x,y) = \max\{0, x+y-1\} \quad \text{ i } x,y \in [0,1], \tag{7.10}$$

Istotną częścią systemu wnioskowania rozmytego, która ma decydujący wpływ na jakość działania systemu jest baza reguł. Do tworzenia bazy reguł wykorzystuje się często wiedzę pozyskaną od ekspertów, która następnie zakodowana zostaje w postaci reguł rozmytych typu JEŻELI ..... TO ..... Mimo niewątpliwych zalet pozyskiwania wiedzy od ekspertów (np. możliwość opisu systemów, które nie posiadają modeli matematycznych) istnieją także pewne niedogodności związane z takim jej pozyskiwaniem. Dane przekazane przez eksperta mogą być niekompletne lub błędne, a ocena niektórych zdarzeń subiektywna (dwóch różnych ekspertów to samo zjawisko może opisać inaczej)[31].

7.4. Wprowadzenie do systemów Takagi-Sugeno-Kanga

W zależności od typu reguł rozmytych można mówić o układzie rozmytym typu Mamdaniego lub Takagi-Sugeno. W przypadku układów typu Takagi-Sugeno reguły rozmyte zapisujemy w postaci:

$$\text{Jeżeli } x=A \text{ to } y=f(x). \tag{7.11}$$

Najczęściej przedstawiona reguła przyjmuje postać, w której konkluzja jest kombinacją liniową wejść :

$$\text{Jeżeli } x=A \text{ to } y=ax+b, \tag{7.12}$$

lub w szczególnym przypadku:

$$\text{Jeżeli } x=A \text{ to } y=c. \tag{7.13}$$

W powyższym przypadku procedura wnioskowania rozmytego odbywa się identycznie jak w przypadku rozmytych systemów typu Mamdaniego przy czym następniki reguł rozmytych reprezentowane są przez zbiory rozmyte typu singleton.

Dwa podstawowe podejścia do maszynowego uczenia się to: symboliczne i połączeniowe. Przez długi okres czasu obydwie techniki rozwijały się niezależnie. Od pewnego czasu jednak, pomimo rozwoju niezależnego naukowcy zaczęli szukać metod zintegrowania ze sobą obydwu metod.

Sieci neuronowe mają kilka zalet w porównaniu do systemów indukcyjnego uczenia się, ale również systemy indukcyjnego uczenia się przewyższają w kilku kwestiach sieci neuronowe. Jeżeli możliwe byłoby zintegrowanie metod uczenia się sieci neuronowych ze zdolnością objaśniania w systemach objaśnieniowych, powstałyby lepiej funkcjonujące systemy hybrydowe.

W tabeli 7.2 zamieszczono porównanie systemów ekspertowych z sieciami neuronowymi [25, 34]. Można zauważyć, że obydwa rozwiązania mają swoje mocne i słabe strony. Gdy są połączone, uzupełniają się nawzajem. Stąd systemy neuro-rozmyte są układami łączącymi własności sieci neuronowych (zdolność uczenia się) oraz systemów rozmytych (zdolność do opisu i przetwarzania wiedzy jakościowej). Zaletą systemów neuro-rozmytych jest możliwość wykorzystania w procesie dostrajania parametrów zarówno wiedzy jakościowej jak i ilościowej. Dodatkowo nie ma potrzeby opracowywania nowych algorytmów uczących, ponieważ bez większych modyfikacji można zaadaptować procedury uczące znane dla sieci neuronowych lub algorytmy klasteryzacji danych. Możliwe jest również wprowadzenie do systemu neuro-rozmytego wiedzy pochodzącej od eksperta. System neuro-rozmyty nie musi być rozpatrywany, jako czarna skrzynka ponieważ jej parametry posiadają interpretację w postaci reguł rozmytych.

Tabela 7.3. Porównanie systemów ekspertowych z sieciami neuronowymi

Systemy ekspertowe	Sieci neuronowe
Trudności w uczeniu się i adaptacji. Nie mogą zwiększyć swojej wydajności wraz z rozwojem doświadczenia.	Uczenie się jest dziedziczne. Mogą dopasować się do zmieniającego otoczenia przez przeszkolenie się.
Potrzebny jest ekspert by przygotować wiedzę	Używa przykładów do szkolenia się
Długi czas tworzenia	Czas tworzenia jest krótszy. Wiele z czasu tworzenia poświęca się na wybranie odpowiedniego algorytmu sieci neuronowej i dopasowaniu sieci do systemu.
Trzeba przygotować system w wiodącej dziedzinie. Jeżeli problem wyjdzie poza zakres dziedziny obsługiwanej, wydajność systemu ekspertowego znacząco spada	Szeroki zakres możliwych dziedzin. Dopasowane do udzielania odpowiedzi na tematy, z którymi nigdy nie miały kontaktu
Komunikuje się z użytkownikiem w sposób prosty. Może pobrać od użytkownika brakujące informacje	Brak komunikacji z użytkownikiem
Bardzo dobra umiejętność wyjaśniania. Mogą wyjaśnić rozumowanie zawarte w wyniku lub wytłumaczyć dlaczego zadawane jest kolejne pytanie dla użytkownika	
Nie są odporne na błędne dane	Odporne na błędne i nieważne dane
Długi czas uruchamiania dla rozbudowanych baz reguł	Gdy sieć jest wyszkolona, uruchamia się bardzo szybko.
Zarządzanie dużymi bazami reguł jest trudne. Dodawanie nowych reguł lub aktualizowanie już istniejących w dużych bazach, może być żmudne.	Łatwe w obsłudze i zarządzaniu. Jeśli są zmiany otoczenia, zazwyczaj ponowne uczenie się samej sieci jest wystarczające.

Proces doboru odpowiedniej struktury można częściowo zautomatyzować wykorzystując w tym celu algorytmy klasteryzacji. System neuro-rozmyty można przedstawić jako strukturę składającą się z trzech części. Część pierwsza odpowiedzialna jest za rozmywanie danych wejściowych, cześć druga - za przeprowadzenie operacji wnioskowania oraz część trzecia - za wyostrzenie wyników. Wiedza zakodowana w postaci parametrów rozmytej sieci neuronowej

może być rozpisana na reguły rozmyte. W zależności od typu reguł rozmytych można mówić o układzie neuro-rozmytym typu Mamdaniego lub Takagi-Sugeno.

Logika rozmyta okazała się bardzo przydatna w zastosowaniach inżynierskich, czyli tam, gdzie klasyczna logika klasyfikująca jedynie według kryterium prawda/fałsz nie potrafi skutecznie poradzić sobie z wieloma niejednoznacznościami i sprzecznościami. Znajduje wiele zastosowań, między innymi w elektronicznych systemach sterowania (maszynami, pojazdami i automatami), zadaniach eksploracji danych czy też w budowie systemów ekspertowych.

Metody logiki rozmytej wraz z algorytmami ewolucyjnymi i sieciami neuronowymi stanowią nowoczesne narzędzia do budowy inteligentnych systemów mających zdolności uogólniania wiedzy.

7.5. Algorytmy ewolucyjne

Z definicji algorytm ewolucyjny to taki, który przetwarza populację rozwiązań poprzez wykorzystanie mechanizmów selekcji oraz operatorów różnicowania [31]. Należy dodać, że rozwiązania w populacji oddziałują na siebie i te „lepsze" mają większe szanse na „przeżycie". Ponadto, w celu dokonania selekcji istnieje potrzeba oceny rozwiązań, czyli funkcja oceny, jakości, przystosowania. Natomiast kolejne rozwiązania (pokolenia), po zastosowaniu mechanizmów różnicowania, powinny zawierać część informacji rodziców. Do algorytmów ewolucyjnych zalicza się między innymi: Algorytmy Genetyczne (AG), strategie ewolucyjne, programowanie genetyczne, komitety klasyfikatorów, algorytmy mrówkowe (mrowiskowe), algorytmu rojowe oraz sztuczne systemy immunologiczne.

Do cech algorytmów ewolucyjnych w stosunku do klasycznych metod optymalizacji można zaliczyć: przetwarzanie populacyjne, brak wymogu informacji o gradiencie, zdolność do radzenia sobie z pułapkami lokalnych optimów (zależna od konkretnego zadania i dobrego doboru parametrów algorytmu), bardziej ogólne i elastyczne podejście oraz możliwość działania na dowolnej reprezentacji [5]

Algorytm genetyczny stanowi rodzaj algorytmu przeszukującego przestrzeń alternatywnych rozwiązań problemu w celu wyszukania rozwiązań najlepszych. Sposób działania algorytmów genetycznych cechuje analogia do naturalnych zjawisk ewolucji biologicznej. W algorytmach genetycznych stosuje się pojęcia zapożyczone z genetyki naturalnej. Istnieje populacja rozwiązań. Każde rozwiązanie jest nazywane osobnikiem. Na populacji dokonywana jest operacja selekcji (słabe osobniki giną), krzyżowania (czyli tworzenia nowych osobników) i mutacji (dzika karta).

W terminologii algorytmów genetycznych używa się następujących pojęć [15]:

Populacja – czyli zbiór osobników o określonej liczebności. Osobniki populacji w algorytmach genetycznych stanowią zakodowane w postaci tzw. chromosomów zbiory parametrów zadania, czyli rozwiązania, określane też jako punkty przestrzeni poszukiwań.

Chromosomy, nazywane również łańcuchami lub ciągami kodowymi to uporządkowane ciągi genów.

Gen - nazywany też cechą, znakiem, detektorem to pojedynczy element genotypu (w szczególności chromosomu).

Genotyp, czyli struktura - to zespół chromosomów danego osobnika. Osobnikami populacji mogą być genotypy albo pojedyncze chromosomy.

Fenotyp - zestaw wartości odpowiadający danemu genotypowi, czyli zdekodowana struktura. Stanowi on zbiór parametrów zadania.

Allel to wartość danego genu (określana też jako wartość cechy lub wariant cechy).

Locus to położenia danego genu w łańcuchu czyli w chromosomie.

Funkcja przystosowania (także funkcja dopasowania lub funkcja oceny, ang. fitness function) – jest miarą przystosowania (dopasowania) danego osobnika w populacji. Pozwala ona ocenić stopień przystosowania poszczególnych osobników w populacji, aby na jej podstawie wybrać osobniki najlepiej przystosowane (o największej wartości funkcji przystosowania). Odpowiada to ewolucyjnej zasadzie przetrwania „najsilniejszych" (najlepiej przystosowanych) osobników i ma znaczący wpływ na działanie algorytmów genetycznych. Stąd musi być odpowiednio zdefiniowana.

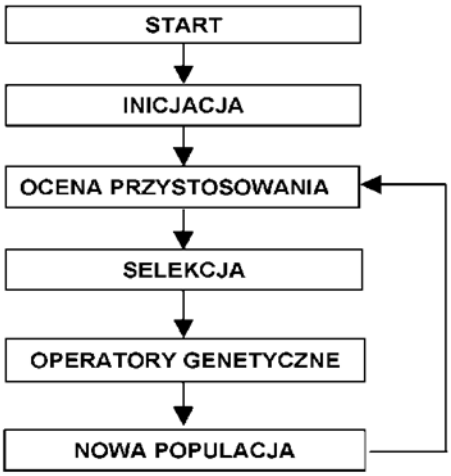
Generacja – odnosi się do kolejnej iteracja w algorytmie genetycznym.

Pokolenie (niekiedy: nowe pokolenie lub pokolenie potomków) - to nowoutworzona populacja osobników.

W algorytmie genetycznym, w każdej jego iteracji, jest oceniane przystosowanie każdego osobnika danej populacji za pomocą funkcji przystosowania. Na tej podstawie tworzona jest nowa populacja osobników, którzy stanowią zbiór potencjalnych rozwiązań problemu, np. zadania optymalizacji.

Klasyczny algorytm genetyczny

Mechanizm działania klasycznego (elementarnego, prostego) algorytmu genetycznego jest prosty. Polega na kopiowaniu ciągów i wymianie podciągów. Na rysunku 7.2 pokazano schemat działania algorytmu genetycznego.



Rys. 7.2. Schemat blokowy działania algorytmu genetycznego.

Składa się ono z następujących kroków [15]:

- Inicjacja - utworzenie populacji początkowej, poprzez losowy wybór ustalonej liczby chromosomów.
- Ocena przystosowania - obliczenie wartości funkcji przystosowania dla każdego chromosomu.
- Selekcja chromosomów - wybór chromosomów, które biorą udział w tworzeniu nowej populacji.
- Zastosowanie operatorów genetycznych - na grupie chromosomów, wybranej drogą selekcji działają operatory genetyczne. W klasycznym algorytmie genetycznym są operatory krzyżowania i mutacji. Przy czym, krzyżowanie powinno zachodzić znacznie częściej niż mutacja.
- Utworzenie nowej populacji - Chromosomy otrzymane jako rezultat działania operatorów genetycznych na chromosomy tymczasowej populacji wchodzą w skład nowej populacji. W klasycznym algorytmie genetycznym cała poprzednia populacja chromosomów jest zastępowana przez tak samo liczną nowa populacje potomków.
- Wyprowadzenie „najlepszego" chromosomu - najlepszym rozwiązaniem jest chromosom o największej wartości funkcji przystosowania.

Nową populację tworzą chromosomy powstałe w wyniku selekcji i działania operatorów genetycznych. Nowa populacja zastępuje w całości starą, i staje się bieżącą w kolejnej generacji (iteracji) algorytmu genetycznego.

Charakterystyczny dla klasycznego algorytmu genetycznego jest sposób kodowania. Chromosomy są zakodowanie binarnie - geny mogą przyjmować tylko wartości 0 i 1. Selekcji dokonuje się z wykorzystaniem metody ruletki, która stanowi jedną z najbardziej podstawowych metod selekcji. Polega na utworzeniu koła ruletki z polami odpowiadającymi poszczególnym chromosomom. Wielkość pól jest proporcjonalna do wartości funkcji przystosowania. Proces selekcji oparty jest na obrocie ruletką tyle razy ile osobników jest w populacji i na wyborze za każdym razem jednego chromosomu do nowej populacji. Pewne chromosomy są wybierane więcej niż jeden raz, niektóre dokładnie raz, a niektóre wcale [22].

Krzyżowanie jednopunktowe - krzyżowanie zachodzi z pewnym Prawdowo-dobieństwem pc; dla każdego osobnika losuje się liczbę i sprawdza, czy zachodzi krzyżowanie. Następnie dobiera się wybrane chromosomy losowo w pary. Losuje się liczbę określającą miejsce krzyżowania i wymienia kod.

Mutacja - również zachodzi z pewnym zadanym prawdopodobieństwem. Dla każdego osobnika losuje się liczbę sprawdzając czy będzie on podlegał mutacji. Jeśli tak to losuje się gen, który będzie zmutowany i dokonuje się mutacji, czyli zamiany wartości genu na przeciwny. Mutacja w klasycznym algorytmie genetycznym odgrywa drugorzędną rolę. Częstość mutacji potrzebna do uzyskania dobrych wyników w empirycznych badaniach nad algorytmami genetycznymi jest rzędu jeden do tysiąca skopiowanych bitów. W naturalnych populacjach częstość jest równie mała - lub nawet mniejsza [22].

Kodowanie binarne przyjęto w klasycznym algorytmie genetycznym. Oznacza to, że allele wszystkich genów w chromosomie są równe 0 lub 1. Michalewicz w [22] pisze: „Reprezentacja rozwiązań w postaci łańcuchów binarnych zdominowała badania nad algorytmami genetycznymi. Kodowanie to ułatwia także analizę teoretyczną i umożliwia wprowadzenie eleganckich operatorów genetycznych. Stosuje się różne modyfikacje tego sposobu kodowania. Kodowanie logarytmiczne jest przykładem modyfikacji kodowania binarnego. Ten sposób kodowania jest stosowany celem zmniejszenia długości chromosomów w algorytmie genetycznym. W kodowaniu logarytmicznym pierwszy bit ciągu kodowego oznacza znak funkcji wykładniczej, drugi bit oznacza znak wykładnika funkcji, a pozostałe bity (bin) są reprezentacją wykładnika funkcji [15].

W algorytmach genetycznych można wyróżnić etap selekcji, który z bieżącej populacji osobników wybiera te o największej wartości funkcji przystosowania do populacji rodzicielskiej. Ze względu na ograniczenia metody ruletki w zakresie

stosowalności tylko do jednej klasy zadań, oraz zbyt wczesnego eliminowanie z populacji osobników o bardzo małej wartości funkcji przystosowania. Może to prowadzić do zbyt wczesnej zbieżności algorytmu [15]. Dlatego stosuje się inne metody selekcji takie jak: selekcja turniejowa, rankingowa i strategia elitarna.

7.6. Biologicznie inspirowane metody optymalizacji

Idea algorytmu rojowego pochodzi z pracy [12] z 1995 roku - jest to więc dość młode zagadnienie. Od tej pory powstało wiele wariacji sposobów rozwiązań problemów związanych z tą ideą.

Zachowanie pojedynczej mrówki, pszczoły czy termita jest zwykle bardzo proste. Jednak o sile tego gatunku decyduje nie jednostka, jednak cała społeczność. Powyższa obserwacja stanowi podstawę algorytmów w dziedzinie Swarm intelligence. Po pojawieniu się pierwszych prac wykorzystujących zachowanie owadów, wielu naukowców rozpoczęło obserwację naturalnego zachowania zwierząt i przekładanie sposobów radzenia sobie z problemami w realnym świecie na problemy dotyczące np. optymalizacji.

Ogólnie można stwierdzić, że takie algorytmy będą bazowały na zbiorze prostych osobników, wykonujących określone czynności, które będą ze sobą się w odpowiedni sposób komunikowały. Tego typu algorytmy nie mają narzuconego z góry sterowania pracą wszystkich osobników, jednak dzięki założonym regułom całość osiąga założony cel.

W pracy [1] wyróżniono kilka zasadniczych cech zachowania, które są zakorzenione w naturze, a z którego korzystają algorytmy oparte na Swarm Intelligence:

- *Jednorodność* każdy element ma z góry określony model zachowań, nie występuje stały lider;
- *Lokalność* tylko najbliższe elementy mają wpływ na zachowanie każdego elementu;
- *Unikanie kolizji* z elementami w okolicy;
- *Dostosowanie prędkości* z elementami które są w okolicy;
- *Centrowanie stada* próba trzymania się w odpowiedniej bliskości do innych elementów.

Mrówki są niewielkimi owadami rozpowszechnionymi na całym świecie. Tworzą one zhierarchizowane kolonie, w których każda mrówka ma przydzielone

ściśle określone obowiązki. Pojedyncza mrówka nie jest w stanie funkcjonować samodzielnie, możliwość przeżycia zyskuje dopiero działając w kooperacji z innymi mrówkami. Dzięki współdziałaniu, mrówki zyskują możliwość budowy gniazda i poszukiwania pokarmu a sposób w jaki to robią okazuje się zarówno zaskakujący, jak i przynoszący duże korzyści.

Kluczem do zagadki mrówek jest feromon – zapachowa, semiochemiczna substancja wydzielana przez owady i gryzonie w różnych celach takich jak: płciowe, obronne, komunikacyjne, markujące terytorium. W przypadku mrówek zakres użycia feromonu przez mrówki jest bardzo duży. Feromon jest podstawowym sposobem komunikacji i pozwala np. rozpoznać dwóm mrówkom czy należą do tej samej kolonii, a także identyfikować królową. Interesujące nas działanie feromonu to znaczenie drogi i odnajdywanie pożywienia. W przypadku mrówek polega ono na pozostawianiu feromonu na drodze przez mrówkę która odnalazła źródło pożywienia a także na wabieniu pozostałych mrówek w miejsce gdzie jego intensywność jest największa. Dzięki temu substancja ta pełni rolę substytutu mózgu całego gniazda. Kolonia mrówek korzystająca z zapachowi feromonu, wydaje się podlegać pewnemu „społecznemu" algorytmowi, który nazywamy algorytmem mrówkowym.

Początki badań nad algorytmami mrówkowymi związane jest z pracą czwórki naukowców J.-L. Deneubourga, S. Gossa, S. Arona oraz J.-M. Pasteels, nad społecznym zachowaniem argentyńskich mrówek. Badacze zostali zainspirowani obserwacjami P.-P Grassa, dotyczącymi porozumiewaniem się termitów przy budowie gniazd. Przeprowadzili oni eksperyment, którego wyniki opublikowali w 1989 roku. Polegał on na połączeniu gniazda mrówek z źródłem pożywienia za pomocą dwóch dróg o bardzo zróżnicowanej długości. Pierwsze mrówki docierające do rozwidlenia dróg dokonywały wyboru przypadkowo – z jednakowym prawdopodobieństwem. Mrówki docierały do pożywienia i wracały do gniazda również przypadkową drogą, zachowując mniej więcej podobną ilość na obydwu drogach. Jednak, po pewnym czasie prawie cały ruch przenosił się na krótszą ze ścieżek dzięki feromonowi. Mrówki wybierające krótsze połączenie szybciej pokonywały drogę gniazdo-pożywienie i mogły dzięki temu pozostawić w sumie więcej feromonu. Kolejne mrówki wyczuwając intensywniejszy zapach krótszej z dróg, wybierały ją pogłębiając proces.

Pierwszą implementację algorytmów mrówkowych w świat komputerów dokonał Marco Dorigo. Nosi on nazwę Systemu Mrówkowego – Ant System. Algorytm ten wprowadza podstawowe definicje i założenia. Pozostałe algorytmy nie tylko korzystają z tych założeń, ale są właściwie modyfikacjami Ant System, różniące się głównie sposobem nanoszenia feromonu.

Algorytmy mrówkowe służą do rozwiązywania trudnych problemów kombinatorycznej optymalizacji, takich jak np. problem komiwojażera. Znajdują również zastosowanie w rozwiązywaniu dyskretnych problemów optymalizacyjnych, np. do wyznaczania tras pojazdów, sortowania sekwencyjnego, czy wyznaczania tras routingu w sieciach telekomunikacyjnych [18].

7.7. Technologie agentowe

Jednym z narzędzi sztucznej inteligencji jest inteligentny agent (ang. intelligent agent), inaczej również nazywany software agent, multiagent, distributed agent, autonomy agent, mobile agent, neural agent, adaptive agent, engineering agent, serach agent, inforamtion agent, robot, softbot, knowbot, taskbot, userbot. Jest to mały, sprytny program pozwalający zautomatyzować wybraną czynność, często podejmujący decyzje w trakcie działania

Pojęcie „agent" zostało w pewnym stopniu odziedziczone po technikach z lat 80. Kolejna generacja programów – inteligentnych agentów ma znacznie większe szanse zaistnienia w świecie informatyki. Ich zaletą jest to, że wykorzystują nowe modele inteligencji komputerowej. Ważnym elementem jest wspólna filtracja, która wykorzystuje myśl ludzką w postaci bazy danych z priorytetami użytkowników. Kilku agentów odwołuje się w swoim działaniu do dziedzin wiedzy takich jak sieci neuronowe, logika rozmyta, przetwarzanie języka naturalnego i wspólna filtracja. Od agentów oczekuje się automatyzacji wykonywania powtarzających się czynności. W Internecie agent po raz pierwszy został użyty do skatalogowania milionów stron tworzących sieć

Systemy agentowe są naturalnie przystosowane do uruchamiania w dużych lub niepewnych środowiskach, np. w sieciach komputerowych, gdzie może zajść awaria łącza, awaria komputera lub ktoś może sabotować obliczenia wysyłając błędne dane. Systemy agentowe nie wymagają zsynchronizowanego zegara, umożliwiają płynne zwiększanie lub zmniejszanie ilości pamięci i procesorów, na których są uruchamiane (bez przerywania obliczeń), tolerują opóźnienia komunikacji i wykorzystują możliwości środowisk heterogenicznych.

Właściwości agentów można w pełni wykorzystać, jeśli zostaną połączone w zespoły nazywane Systemami wieloagentowymi (ang. Multi Agents System). Przykładem tworzenia i zastosowania agentów może być niemiecka firma Spider-Software.net, która dla swoich klientów - banków i biur maklerskich, zaprogramowała i umieściła agenty na giełdach m.in. w Barcelonie, Mediolanie, Frankfurcie, Londynie, Wiedniu i Zurichu.

Agenty posiadają pewną wrodzoną inteligencję, za pomocą, której potrafią porozumieć się z innymi współpracownikami w celu wymiany wiedzy, czy też

uniknięcia błędów popełnionych poprzednio przez kolegów. Filozofia systemów wieloagentowych opiera się o organizacje, grupy inteligentnych agentów, które mają za zadanie wspólnie rozwiązywać problemy. Agenty mają możliwość kooperacji, negocjacji oraz koordynacji za pomocą wymiany komunikatów między sobą. Systemy wieloagentowe zajmują się rozwiązywaniem problemów, które są zbyt duże dla systemów scentralizowanych. Jednocześnie umożliwiają połączenia i współpracę wielu starych systemów. Rozwiązują rozproszone problemy lub problemy, w których eksperci są rozproszeni. W typowych systemach rozproszonych zadanie powinno być podzielne według pewnych kryteriów dekompozycji na zbiór podzadań, które przydzielane są, zgodnie z pewnymi kryteriami przydziału, poszczególnym agentom. Każdy z agentów rozwiązuje przydzielone podzadanie, a rezultaty mogą być połączone w jedno globalne rozwiązanie.

Wzajemna interakcja między agentami oraz między agentem i jego środowiskiem stanowi podstawę systemów wieloagentowych. Interakcja rozumiana jest tu jako współdziałanie agentów, gdzie jeden podejmuje akcję lub decyzję pod wpływem obecności lub wiedzy innego agenta, a także jako reakcja na zmiany środowiska.

Koncepcja inteligentnego agenta w systemach rozproszonych jest postrzegana jako naturalny rozwój idei programowania równoległego.
Możliwości zastosowania technologii agenckich są bardzo szerokie i obejmują różne obszary. Tworzone oprogramowanie jest wyspecjalizowane do rozwiązywania problemów z bardzo wąskiej dziedziny.

Przemysł był jedną z pierwszych dziedzin, w której zaczęto stosować inteligentne agenty. System kontroli lotów OASIS był testowany na lotnisku w Sydney w Australii. Celem systemu jest wspomaganie pracy kontrolera lotów w zarządzaniu przylotami samolotów na dużych lotniskach.

Ważną dziedziną, w której wykorzystywane są agenty jest e-commerce. Programy agenckie mogą być stosowane od samego początku procesu handlowego, od marketingu, poprzez rekomendowanie produktów, wyszukiwanie towarów i dostawców, negocjacje cen, warunków dostawy oraz ilości kupowanego dobra, po udział w aukcjach i rynkach elektronicznych.

Inteligentny agent, to agent, który może podsunąć ofertę, rekomendować towar, a nawet symulować potrzebę posiadania poprzez informowanie o nowych atrakcyjnych produktach. Przykładem takiego oprogramowania jest system Amazon Delivers, który automatycznie wysyła recenzje najciekawszych książek, gdy tylko nowe pojawią się w kategorii interesującej użytkownika

Przy zaletach omawianego oprogramowania należy zdawać sobie sprawę, że większość aplikacji, które wykorzystuje tą technologię mogłaby być zbudowana równie dobrze bez jej udziału. Pełne zalet inteligentne agenty mają pewne ograniczenia. Systemy budowane przy udziale tego oprogramowania nie mogą być stosowane w dziedzinach, gdzie zagwarantowana musi być odpowiedź w czasie rzeczywistym. Agenty nie mają całościowej kontroli systemu. Kolejnym ograniczeniem jest brak perspektywy globalnej, czyli podejmowane są decyzje na podstawie lokalnej wiedzy poszczególnych agentów. Ważnym elementem jest zdobycie zaufania użytkowników w stosunku do inteligentnych agentów, którym powierzanoby odpowiednie zadania [39].

7.8. Perspektywy rozwoju sztucznej inteligencji i systemów ekspertowych

Wszystkie systemy, które realizują obecnie wyższe czynności poznawcze, rozwiązujące problemy czy analizujące wypowiedzi w języku naturalnym, oparte są na technologii systemów ekspertowych (jednakże w pracy [37], opisano system spontanicznie uczący się znaczenia otrzymywanych I wysyłanych symboli, który zapewne da się zrealizować w postaci sieci neuronowej). System CYC (www.cyc.com) zawierający ponad milion faktów i dziesiątki tysięcy koncepcji nie używa sieci neuronowych ani żadnych inspiracji kognitywnych, ograniczając się do metod symbolicznej reprezentacji wiedzy. Inne modele AI, które odniosły znaczny sukces, systemy Soar [26] i Act [2], również opierają się wyłącznie na podejściu symbolicznym. Pozostaje kwestia, czy można je ulepszyć wykorzystując subsymboliczne podejścia z metodami inteligencji obliczeniowej. Sieci Bayesowskie i modele graficzne mogą stanowić pomost pomiędzy tymi technologiami. Jedynym systemem hybrydowym, który wykorzystywał sieci neuronowe dla analizy tekstów, był DISCERN [23]. Chociaż wykorzystano w nim szereg interesujących idei system ten przestał się rozwijać.

Bardzo złożone supersieci, takie jak indywidualne mózgi, można też traktować jako jednostki oddziaływujące ze sobą i tworzące struktury wyższego rzędu, takie jak grupy ekspertów, instytucje, uniwersytety, wykorzystujące ogromną wiedzę, wymaganą do rozwiązywania problemów, z którymi borykają się współczesne społeczeństwa. Burza mózgów jest przykładem takich oddziaływań, które mogą przyczynić się do powstania nowych idei, ocenianych i analizowanych przez grupy ekspertów. Najtrudniejszym zadaniem jest tworzenie nowych idei, twórcze działanie wymagające nowych kombinacji znanych elementów, generalizacji wiedzy na nowe sposoby. Proces ten nie musi się różnić w zasadniczy sposób od

generalizacji wiedzy na niskim poziomie, w sieciach neuronowych, chociaż zachodzi na znacznie wyższym poziomie złożoności. Prawdziwa trudność tworzenia takich systemów może być związana z koniecznością szczegółowej reprezentacji ogromnej wiedzy, pozwalającej na dodawanie nowych kombinacji znanych elementów i tworzenie nowych koncepcji [8].

Naukowcy od lat spierają się o perspektywy rozwoju sztucznej inteligencji, niemniej jednak podsumowując różne poglądu i doświadczenia w tym zakresie można wyrazić następujące opinie na ten temat [31]:

- Żadna ze stworzonych maszyn nie potrafi wyjść poza zestaw zaprogramowanych przez człowieka zasad.
- Sztuczne systemy inteligentne, ze względu na ograniczenia sprzętowe oraz ogromną złożoność mózgu nie są wstanie symulować procesów w nim zachodzących (posługują się ich modelami).
- Maszyny będą mogły przejść test Turing, ale wyłącznie w wąskim zakresie tematycznym
- Maszyny przyszłości, choć przejawiają oznaki świadomości, to nie będą świadome w sensie filozoficznym.
- W perspektywie kilkudziesięciu lat inteligentne maszyny będą pomagać ludziom
  w wykonywaniu czynności domowych lub zawodowych,
- Komputery, projektujące kolejne generacje komputerów oraz robotów odegrają znaczącą rolę w rozwoju inteligencji mieszkańców Ziemi.

Literatura

1. Abraham A., Das S., Roy S. Swarm Intelligence Algorithms for Data Clustering, Oded Maimon and Lior Rokach (Eds.), Springer Verlag, Germany, pp. 279-313, 2007
2. Anderson, J.R.: Rules of the Mind. Erlbaum, Hillsdale, N.J. (1993) Anderson, J.R.: Learning and Memory. J. Wiley and Sons, New York (1995)
3. Anderson J.A., Rosenfeld E. , „Neurocomputing - Foundation of Research", MIT Press, Cambridge, Mass., 1988.
4. Anderson, J.R.: Rules of the Mind. Erlbaum, Hillsdale, N.J. (1993) Anderson, J.R.: Learning and Memory. J. Wiley and Sons, New York, 1995.
5. Bereta M., Algorytmy genetyczne, Materiały labratoryjne, 2007.
6. Bubnicki Z., Analysis and decision making in uncertain systems, Series: Communications and Control Engineering, Springer-Verlag, Heidelberg, 2004.
7. Bubnicki Z.,Teoria i algorytmy sterowania, PWN, Warszawa, 2002.

8. Duch W., Dokąd zmierza inteligencja obliczeniowa?", w (Cierniaka R. red.) Ewolucja czy rewolucja nowoczesne techniki informatyczne, Książka wydana nakładem: Katedry Inżynierii Komputerowej Politechniki Częstochowskiej, Częstochowa, 2003.
9. Duch, W., Adamczak, R., Grąbczewski, K.: Methodology of extraction, optimiza-tion and application of crisp and fuzzy logical rules. IEEE Transactions on Neu-ral Networks 12 (2001) 277-306.
10. Duch, W., Diercksen, G.H.F.: Feature Space Mapping as a universal adaptive system. Computer Physics Communications 87 (1995) 341-371
11. Duch, W.: Platonic model of mind as an approximation to neurodynamics. In: Brain-like computing and intelligent information systems, ed. S. Amari, N. Ka-sabov. Springer, Singapore (1997) 491-512
12. Eberhart, R. C., Kennedy, J. A. New optimizer using particle swarm theory Proceedings of the Sixth International Symposium on Micromachine and Human Science, Nagoya, Japan, pp. 39-43, 1995.
13. Frasconi, P., Gori, M., Sperduti, A.: A General Framework for Adaptive Pro¬cessing of Data Structures. IEEE Transactions on Neural Networks 9 (1998) 768-786
14. Fraser A.S., Simulation of genetic systems by automatic digital computers. I. Introductio, Part I,II, Aust. J.Biol.Sci., 10,484-499, 1957.
15. Goldberg D. E., Algorytmy genetyczne i ich zastosowania (WNT 1998)
16. Kacprzyk J., Zbiory rozmyte w analizie systemowej, PWN, Warszawa 1986.
17. Karp, P.D.: Pathway databases: a case study in computational symbolic theories. Science 293 (2001) 2040-2044
18. Kupczak M., Szołtysek P., Wykorzystanie algorytmów rojowych do klasteryzacji zbioru ofert sprzedaży mieszkań Algorytmy genetyczne - Projekt 2009.
19. Marczak, M,, Duch, W., Grudziński, K., Naud, A.: (2002) Transformation Distan-ces, Strings and Identification of DNA Promoters. Int. Conf. on Neural Ne¬tworks and Soft Computing, Zakopane, Poland, 2002
20. McClelland T.L., RumelhartD.E., and the PDP Research Group, „Paralell Distributed Processing", MIT Press, Cambrideg, Mass. 1986.
21. McCulloch W.S., Pitts W., A logical calculus of the ideas immanent in nen/ous actMty, Bulletin of Mathematica Biophysics, No 5, 1943, pp. 115-133.
22. Michalewicz Z., Algorytmy genetyczne + struktury danych = Programy ewolucyjne, WNT, Warszawa 1999.
23. Miikkulainen, R. Subsymbolic natural language processing: an integrated model of scripts, lexicon and memory. MIT Press, Cambridge, MA, 1993.
24. Minsky M., Papert S., „Perceptrons", MIT Press, Cambridge 1969.
25. Mulawka J.J.: Systemy ekspertowe. WNT Warszawa.
26. Newell, A., Unified Theories of Cognition. Cambridge, MA: Harvard University Press, 1990.

27. Osowski S., Sieci neuronowe w ujęciu algorytmicznym (WNT 1996)
28. Pawlak Z., Rough sets - theoretical aspect of reasoning about data, Kluwer Academic Publishers, Dordrecht 1991.
29. Pawlak Z., Rough sets, proc. on int. Journal of Information and Computer Science,11, 341, 1982.
30. Rosenblatt F. , The perceprton. A theory of statistical separability in cognitive system, Cornell Aeronautical ab. Inc. Rep. No. VG-1196-G-1, 1968.
31. Rutkowski L., Metody i techniki sztucznej inteligencji, PWN, Warszawa, 2006.
32. Rykaczewski K., Systemy rozmyte i ich zastosowania, Materiały do wykładu, 2006.
33. Ryszard Tadeusiewicz, „Sieci Neuronowe", Państwowa Oficyna Wydawnicza RM 1993.
34. Sienkiewicz W., Sieci neuronowe w tworzeniu reguł w systemach ekspertowych, Materiały do wykładu, 2003.
35. Tadeusiewicz R., Sieci neuronowe, Akademicka Oficyna Wydawnicza RM, Warszawa 1993.
36. Taylor W.K. , Computers and the nen/ous system. Models and analogues in biology, Cambridge Univ. Press, Cambridge, 1960
37. Treister-Goren, A., Hutchens, J.L.: Creating AI: A unique interplay between the development of learning algorithms and their education. Technical Report, AI Enterprises, Tel-Aviv 2000. Available from http://www.a-i.com
38. von Neumann J., The Computer and the Brain, Yale Univ. Press, New Haven, 1958.
39. Waldemar Wojcik (red), Sztuczna inteligencja i metody optymalizacji - od teorii do praktyki, Polskie Towarzystwo Informatyczne, Lublin 2008.
40. Whitley D., Applying genetic algorithms to neural network learning, Proceedings of 7th Conference of Society of Artificial Intelligence and Simulatio of Behavior, Sussex, England, Pitman Publishing, 137-144, 1989.
41. Widrow B. , „The Original Adaptive Neural Net Broom-Balancer", Proceedings of 1987 IEEE Int. Symp. on Circuits and Systems, Philadelphia, PA, May 1987 pp. 351-357.
42. Widrow B., Hoff M.E., „Adaptive switching circuits", IRE WESCON Convention Record, New York, 1960, pp. 96-104.
43. Wilamowski B.M., Irwin J.D., Intelligent Systems (The Industrial Electronics Handbook), Publisher: C.R.C Press
44. Winston P.: Artificial Intelligence. 3rd ed, Addison Wesley (1992) Vapnik, V, Statistical learning theory. New York: John Wiley & Sons, 1998.]