

# Interfejsy głosowe

1. Projektowanie interfejsów głosowych
2. Analiza i synteza mowy

**„Mowa jest źródłem nieporozumień”**

„Mały książę” Antoine de Saint-Exupéry

# Projektowanie interfejsów głosowych



## Interfejs graficzny

- Opiera się na elementach wizualnych
- Pozwala wyświetlić dostępne opcje
- Pozwala pominąć akcje użytkownika niezwiązane z funkcjonalnością

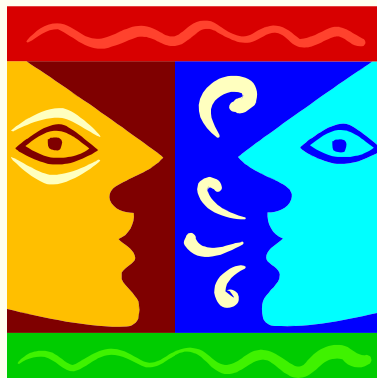
## Interfejs głosowy

- Brak możliwości użycia GUI
- Brak możliwości wizualnej prezentacji,
  - funkcjonalności, z jakiej aktualnie korzysta użytkownik
  - możliwych opcji
- Użytkownik zakłada, że będzie rozumiany tak samo jak w komunikacji z innymi ludźmi

# Wytyczne

1. Udostępnij użytkownikowi informację o tym co może zrobić
2. Informuj użytkownika z jakiej funkcjonalności korzysta i jak ją opuścić
3. Wyrażaj możliwe intencje użytkownika na przykładach – np. pomoc głosowa, instrukcja głosowa
4. Ogranicz liczbę informacji – zadbaj o zwartość i przejrzystość przekazu
  - Nie podawaj więcej niż 3 możliwości na raz
  - Gdy możliwości jest więcej, przedstaw tylko te najczęściej używane i informację o tym jak przejść do pozostałych
5. Używaj wizualnego sprzężenia zwrotnego do zasygnalizowania, że interfejs jest aktywny

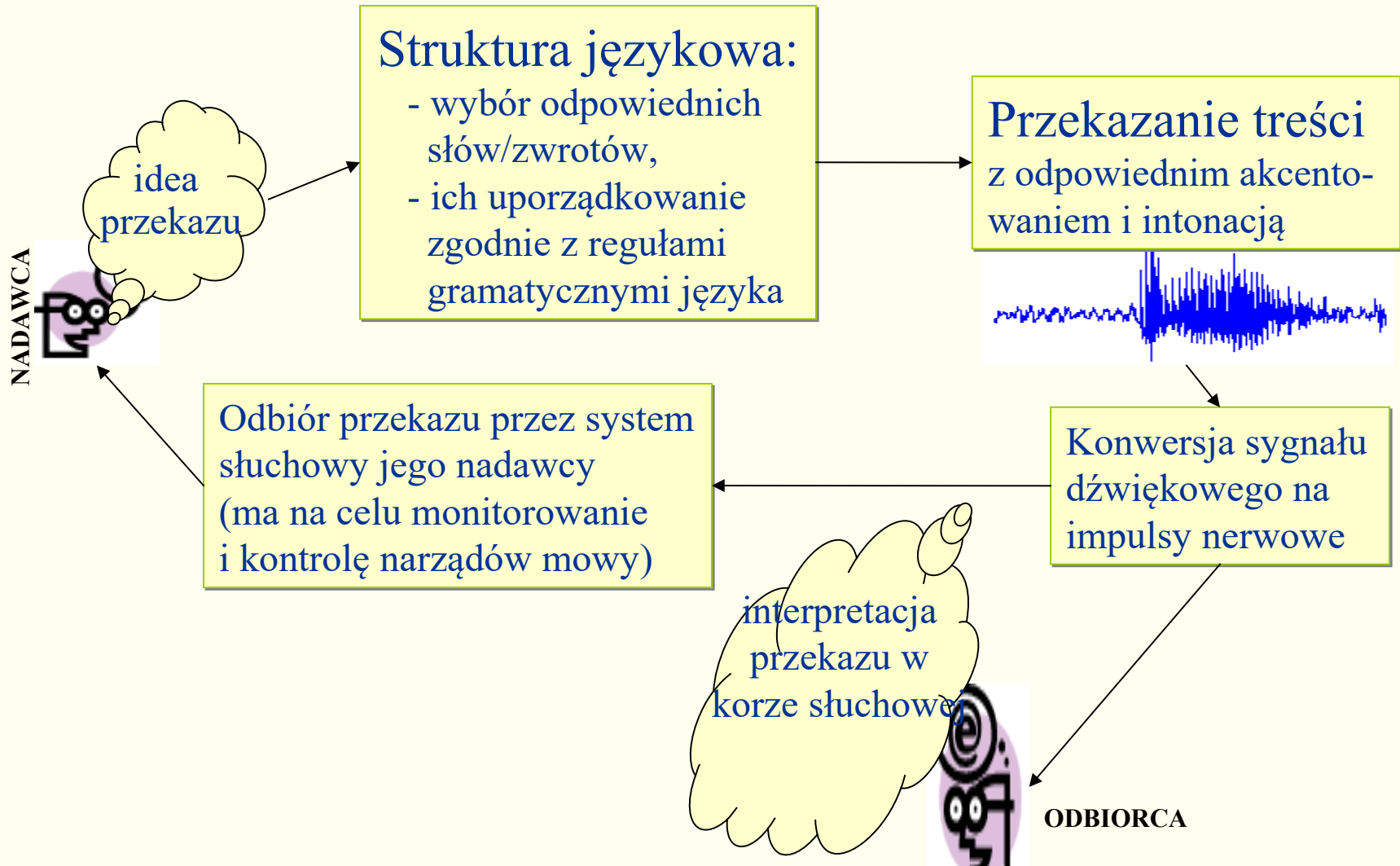
# **Analiza i synteza mowy**



# Rodzaje systemów wykorzystujących mowę

- Rozpoznawanie mowy
  - ◆ Rozpoznawanie mowy ciągłej
  - ◆ Rozpoznawanie pojedynczych wyrazów/poleceń
- Identyfikacja mówcy – określenie, która ze znanych systemowi osób mówi
- Weryfikacja mówcy (sprawdzenie, czy mówca jest rzeczywiście tym za kogo się podaje)
- Poprawa jakości sygnału mowy (redukcja szumów)
- Kodowanie sygnału mowy (do transmisji sygnału)
- Analiza głosu (diagnozowanie chorób układu mowy)
- Synteza mowy (mowa generowana przez komputer)

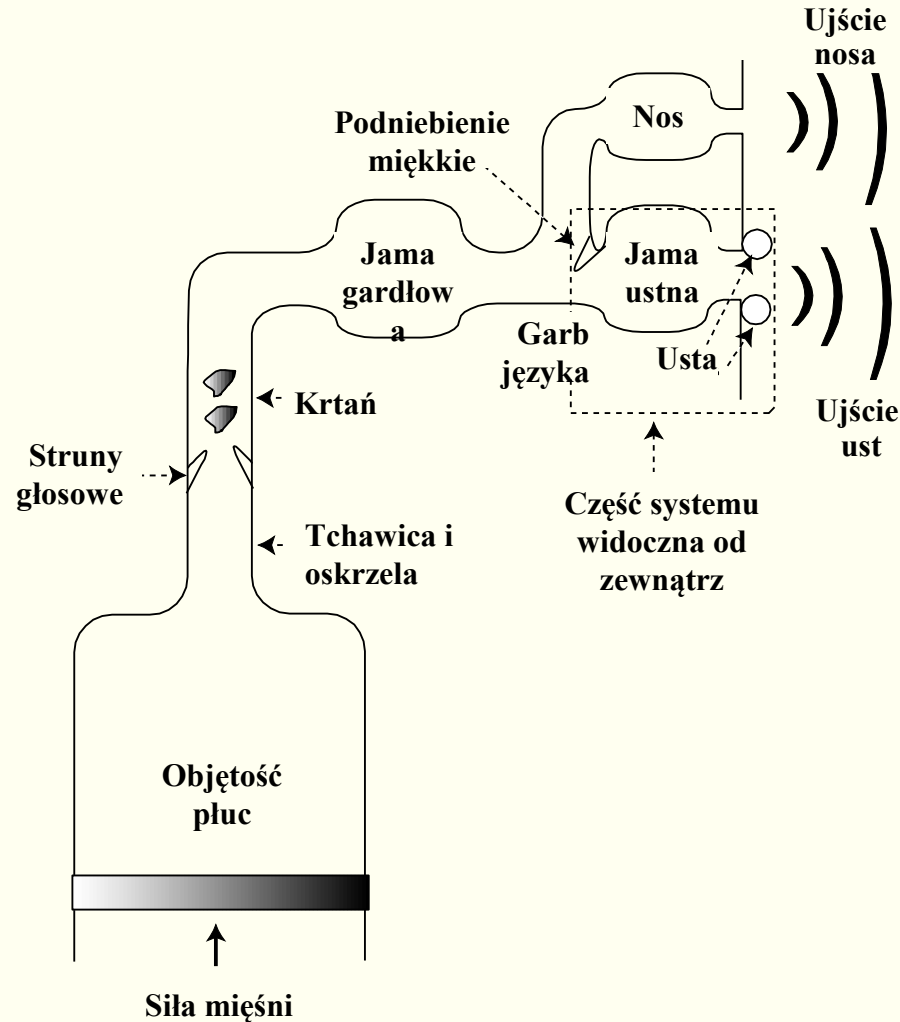
# Komunikacja z użyciem mowy



# Cechy systemu słuchowego

- Odgrywa istotną rolę w tworzeniu i odbiorze informacji
  - ◆ Pozwala odbierać informacje przekazywane za pomocą mowy
  - ◆ Pozwala na kontrolę jakości własnej wypowiedzi (sprzężenie zwrotne)
  - ◆ Brak sprzężenia zwrotnego u osób z wadami słuchu powoduje u nich gorszą wymowę
- Selektywność
  - ◆ osoby słyszące tylko jednym uchem nie posiadają tej cechy
- Niemożność rozróżnienia sygnałów pojawiających się w odpowiednio małych odstępach czasowych lub sygnałów o zbliżonej częstotliwości

# Uproszczony model systemu akustycznego człowieka





# Cechy systemu akustycznego człowieka

## ■ Naturalne filtry akustyczne:

- ◆ Jama gardłowa
- ◆ Jama ustna
- ◆ Jama nosowa

## ■ Przeciętna długość ścieżki akustycznej:

- ◆ U osoby dorosłej płci męskiej/żeńskej: 17/14 cm
- ◆ U dziecka: 10 cm

## ■ Kształt narządów artykulacyjnych określa właściwości filtru akustycznego

# Podstawy analizy akustycznej

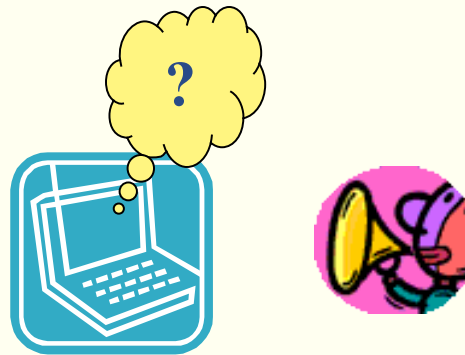
- Charakterystykę widmową fali dźwiękowej cechuje zmienność w czasie, gdyż system fizyczny (narządy mowy) zmienia się szybko w czasie
- Stąd mowę można podzielić na segmenty o podobnych właściwościach akustycznych wyznaczonych w krótkich chwilach czasowych, typowo na:
  - ◆ samogłoski – nie mają ograniczeń wynikających z przepływu powietrza w systemie akustycznym człowieka
  - ◆ spółgłoski – posiadają znaczne ograniczenia, w efekcie czego ich amplituda jest niższa i mają większe zakłócenia
- Zakres fal dźwiękowych produkowanych i odbieranych przez człowieka: 7 – 8 kHz

# Podstawy analizy akustycznej

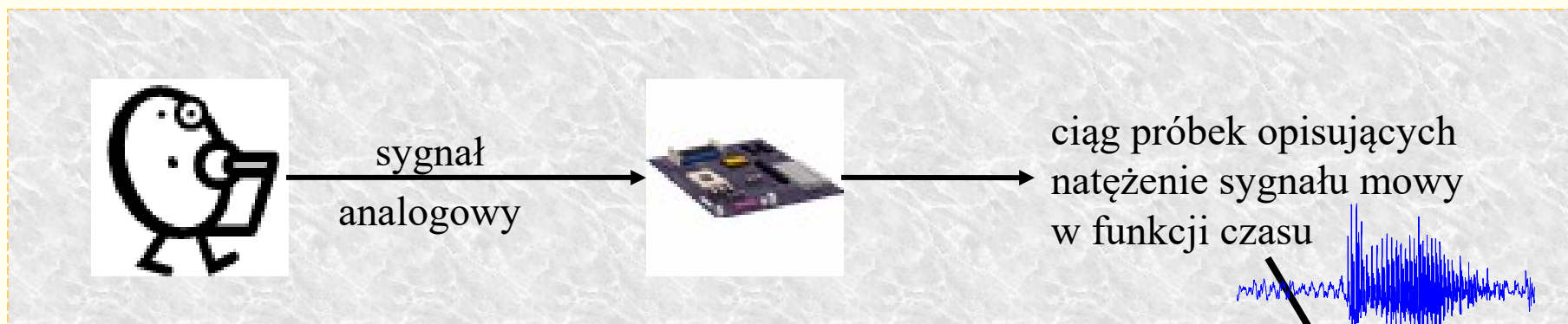
## c.d.

- Sygnał dźwiękowy może służyć do określenia periodyczności, intensywności, czasu trwania i granic między poszczególnymi dźwiękami
- Mowa nie jest ciągiem dyskretnych, łatwo rozróżnialnych dźwięków, lecz raczej serią „docelowych” dźwięków (czasami bardzo krótkich) z przejściami reprezentującymi formowanie się następnego dźwięku (jest to tzw. koartykulacja)

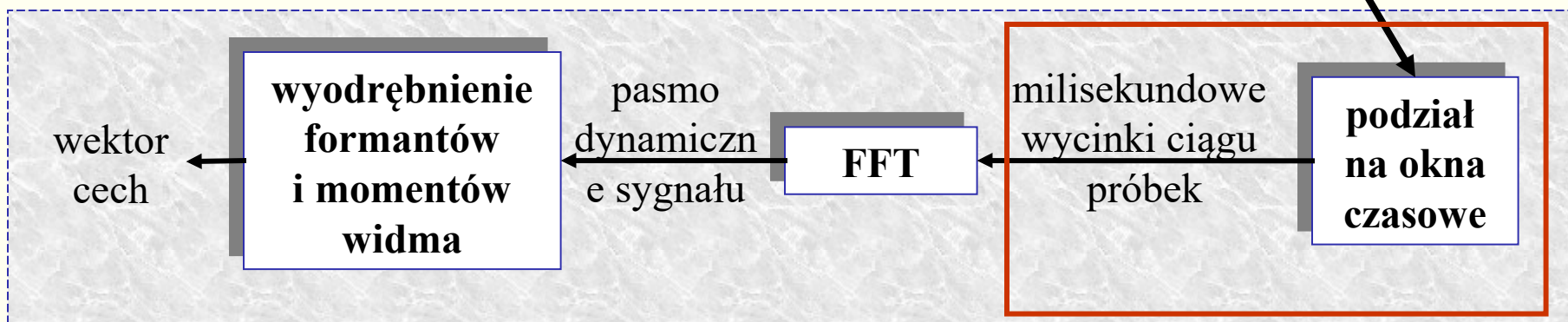
# Rozpoznawanie mowy



# Pozyskiwanie i przetwarzanie sygnału akustycznego



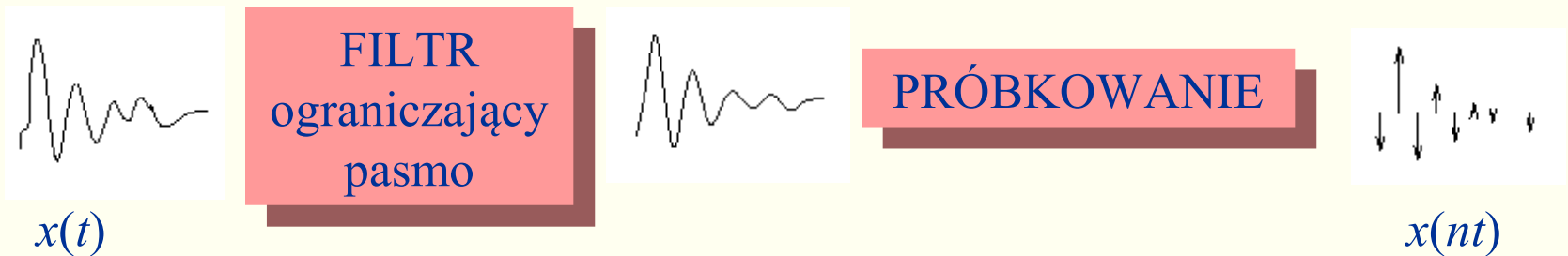
akwizycja



przetwarzanie

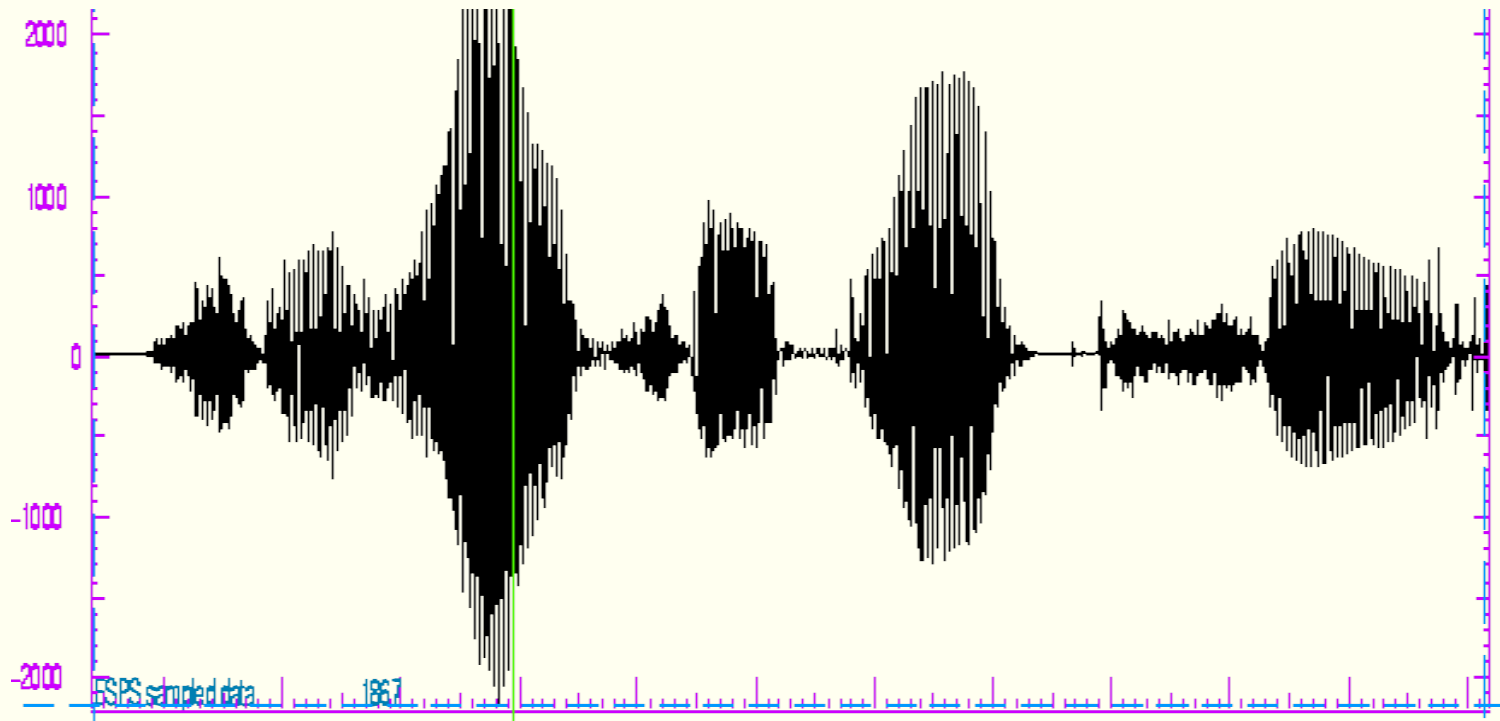
**SEGMENTACJA**

# Próbkowanie sygnału analogowego



- Teoretycznie częstotliwość próbkowania powinna być 2-krotnie większa niż największa częstotliwość, która ma być reprezentowana w sygnale
- W praktyce stosowane są większe częstotliwości próbkowania:
  - ◆ W telefonii 8 kHz
  - ◆ W analizie i syntezie mowy za wystarczającą przyjmuje się częstotliwość 16 kHz
  - ◆ Standardy audio to 44.1 kHz (CD) i 48 kHz (cyfrowa kaseta audio)

# Sygnal mowy - przykład



Sygnal dla zdania „She had your dark”

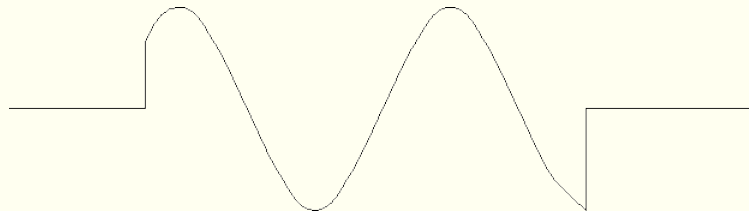
# Podział na bloki (windowing)

**Założenia (dla celów wykonania DFT):**

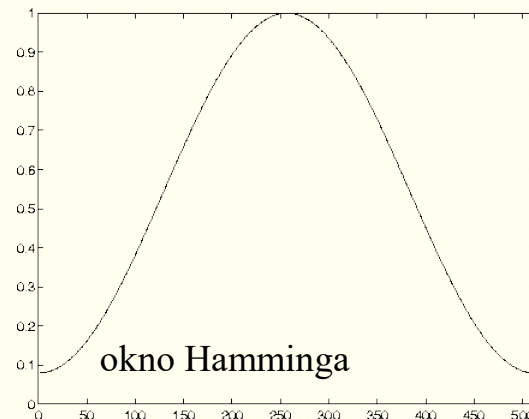
- Sygnał jest stacjonarny w krótkich chwilach czasowych

**Możliwe rozwiązania:**

- Przemnożenie sygnału przez funkcję okna posiadającą wartości 0 poza określonym przedziałem – powoduje powstanie nieciągłości na brzegach przedziału
- Dlatego w analizie mowy wykorzystuje się okno Hamminga



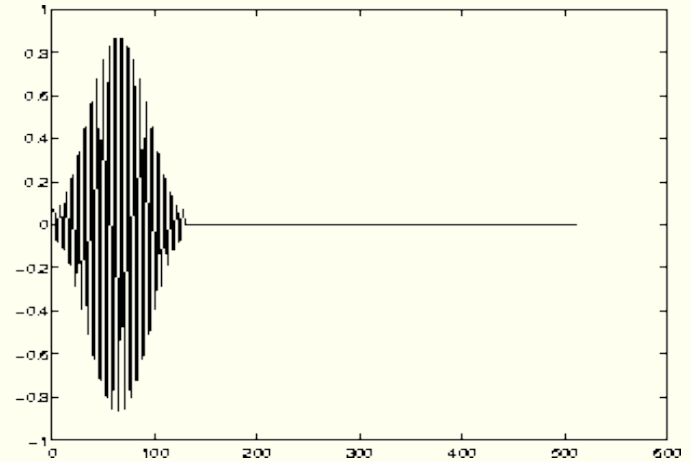
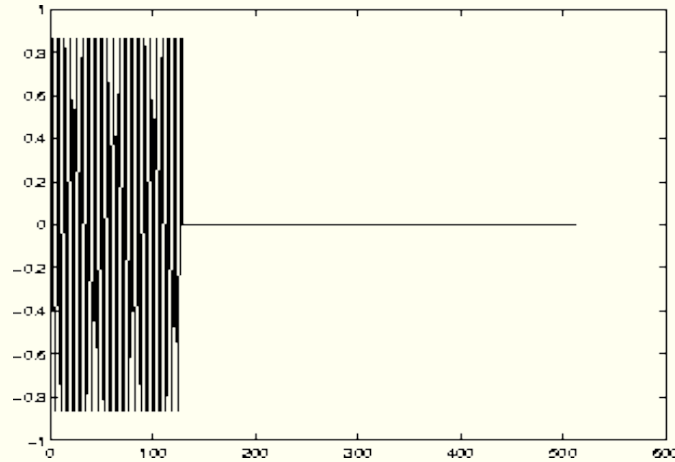
nieciągłości wynikające z zastosowania okna kwadratowego



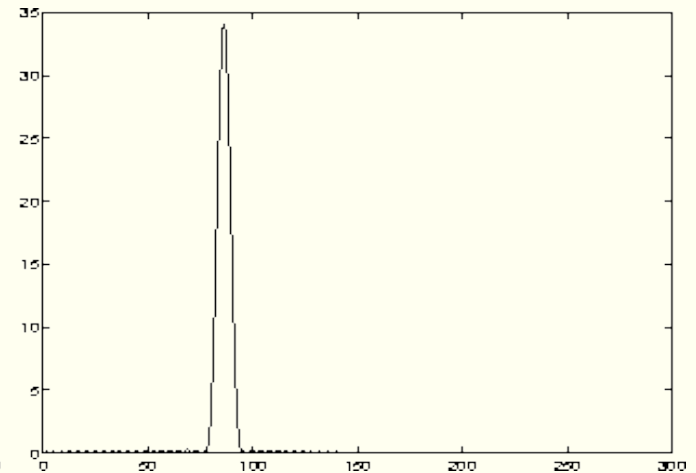
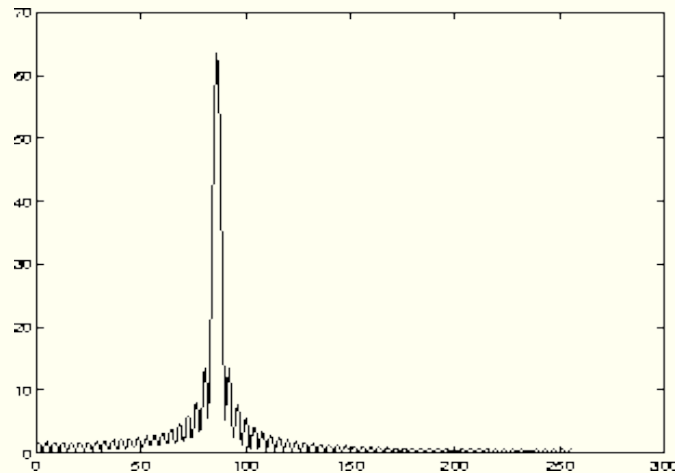


# Przykład windowingu dla funkcji sinus

funkcja sinus  
w oknie



moduł widma  
Fouriera

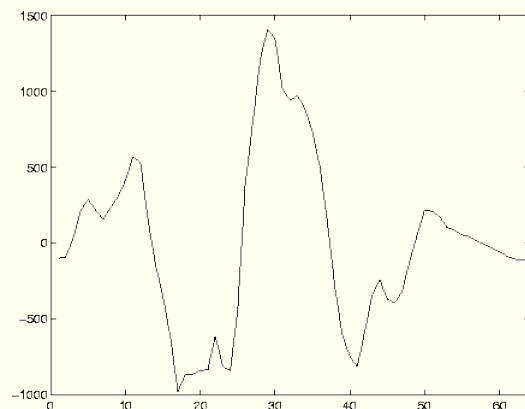
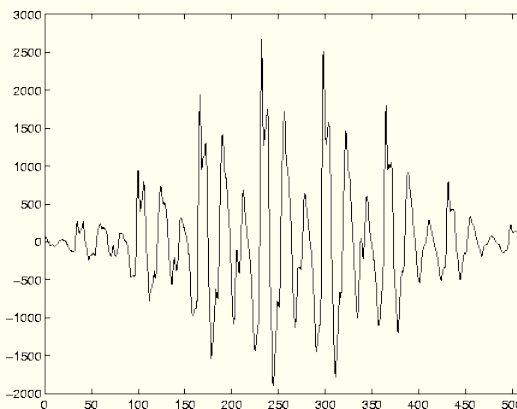
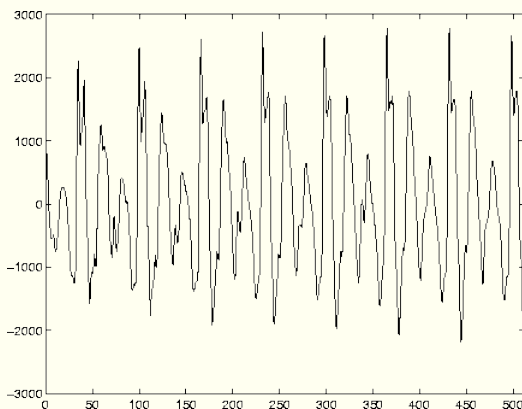


okno kwadratowe

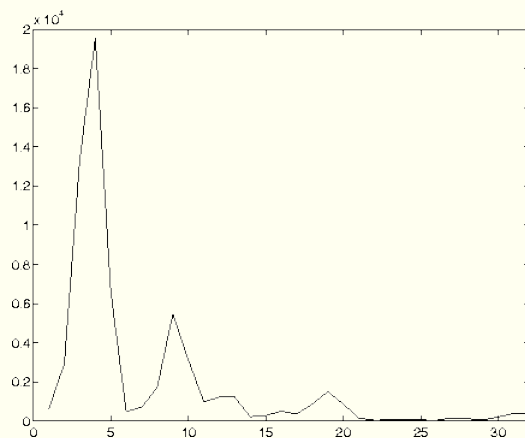
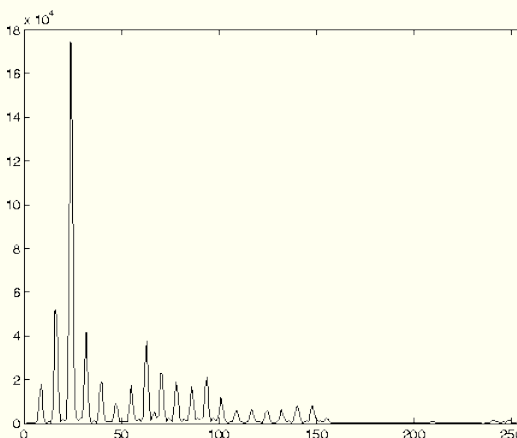
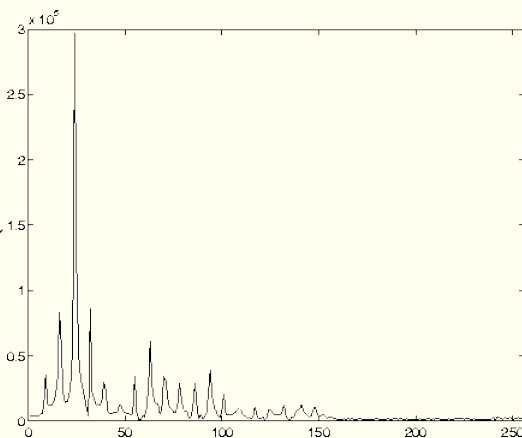
okno Hamminga

# Przykład windowingu dla samogłoski

sygnał dla  
samogłoski



moduł widma  
Fouriera



okno kwadratowe  
dł. 256

okno Hamminga  
dł. 256

okno Hamminga  
dł. 64

# Fonetyka – fonemy i fony

- **Fony** to podstawowe jednostki dźwięku w języku. (syntaktyka, artykulacja)
- Fonem to grupa fonów posiadających to samo znaczenie. (semantyka, język)
- **Alofony** to fony występujące w fonemie.
- Istnieje nieskończona liczba fonów, ale w każdym języku można pogrupować je w 20-60 grup fonemów.
- Brzmienie fonemów jest silnie zależne od osoby je wypowiadającej.

# Fonetyka - koartykulacja

- Fony nie są wypowiedane zawsze w ten sam sposób, ich brzmienie jest zależne od kontekstu.
- Fon jest celem, który chce osiągnąć, ale rzadko osiąga, mechanizm mowy.
- W większości przypadków zbliża się wystarczająco do tego, by być zrozumiałym.
- Systemy syntezy i rozpoznawania muszą uwzględniać koartykulację (dwufony, trójfony).

# Fonematyka - fonemy

- Fonem – podstawowa jednostka mowy mogąca zmienić znaczenie słowa.  
W języku amerykańskim wyróżnia się 42 fonemy. Są to: samogłoski, semisamogłoski, dwugłoski i spółgłoski (nosowe, szczelinowe, zwartoszczelinowe)
- Każdy fonem można traktować jak kod złożony z unikalnego zbioru gestów artykulacyjnych (gest artykulacyjny zawiera rodzaj i położenie pobudzenia dźwiękowego oraz położenie i ruch narządów mowy)
- Około 50-ciu fonemów wystarcza do wypowiedzenia zdania w dowolnym ziemskim narzeczu

# Fonematyka – przykład alofonu

- Przykład alofonu można znaleźć w słowach “pin” i “spin”.
- W drugim z nich głoska /p/ bardziej przypomina w brzmieniu głoskę /b/.
- Alofony są zależne od kontekstu.
- Nie ma różnicy w znaczeniu.

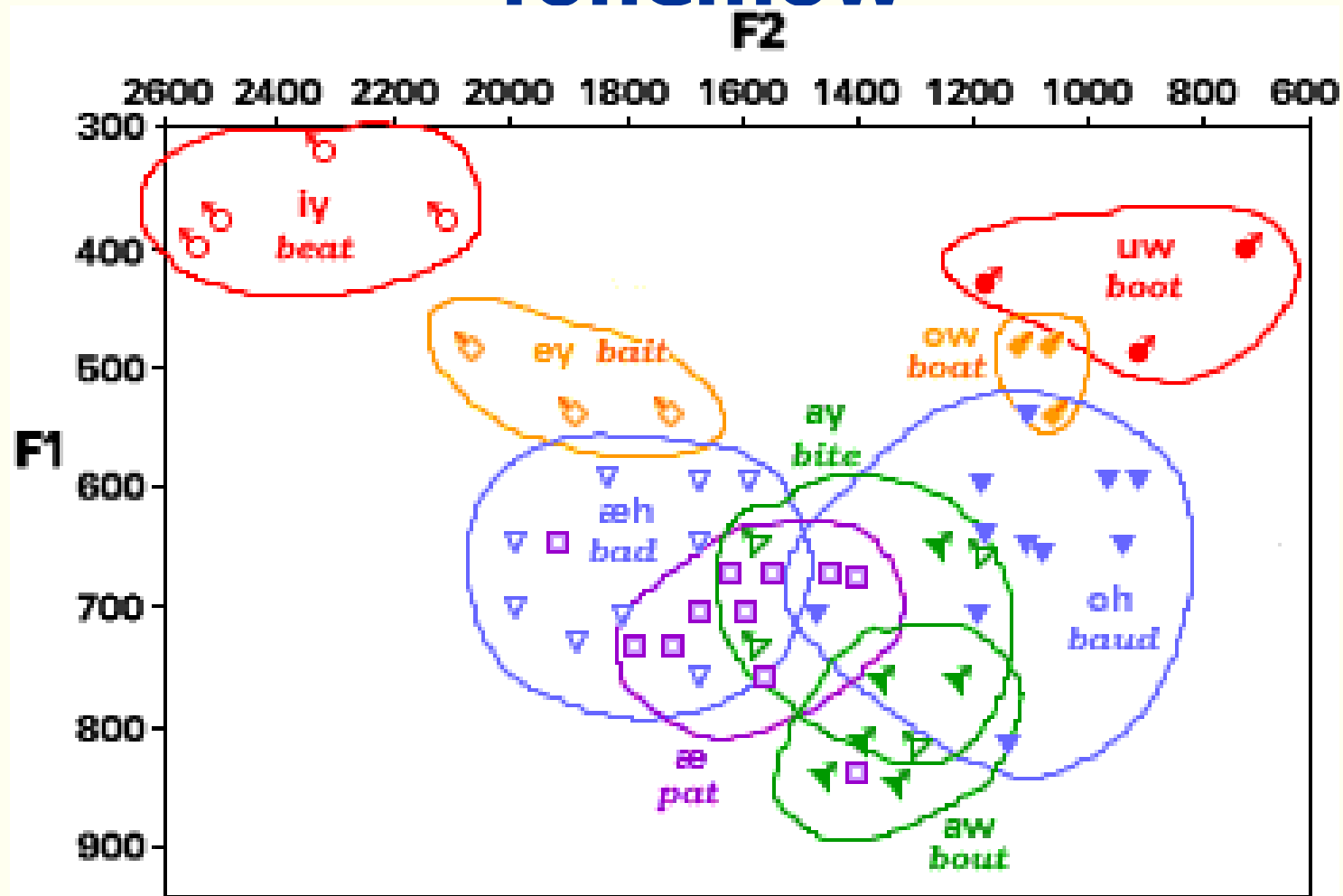
/s/ /p/ /ih/ /n/

⇕

/s/ /b/ /ih/ /n/

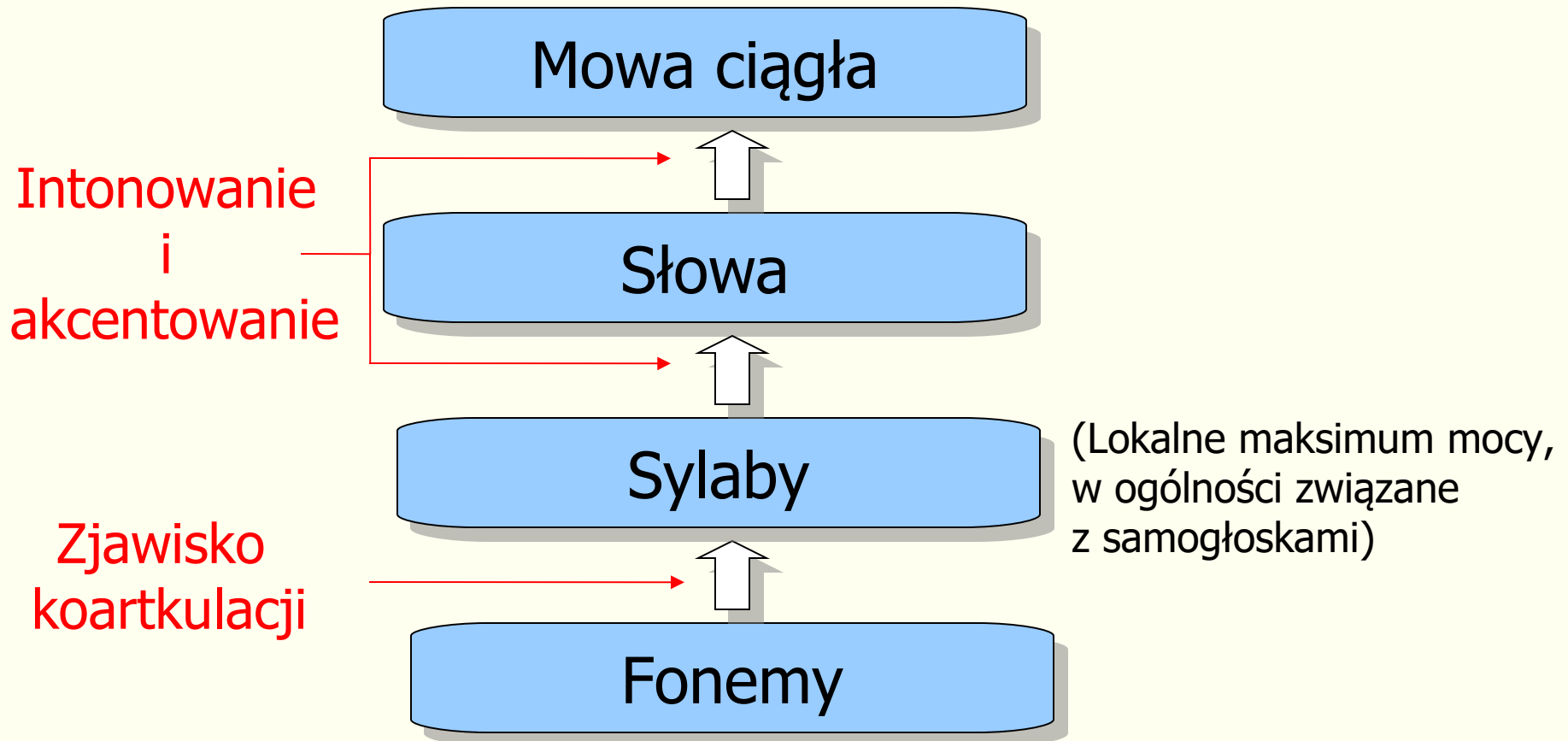
= “spin”

# Fonematyka – grupowanie fonemów



Labov 1994: 168

# Sylaby i słowa





# Intonacja i akcent

- Akcent odzwierciedla stopień nacisku z jakim wypowiedziana jest sylaba lub słowo.
- We frazach i zdaniach intonacja różnicuje wagę (znaczenie) poszczególnych słów.
- Wymaga dalszych badań.

# Schemat systemu rozpoznawania mowy

1. Usunięcie fragmentów sygnału nie zawierających sygnału mowy (detekcja mowy)
2. Wydzielenie cech
3. Klasyfikacja w oparciu o wyznaczone cechy i bazę wzorców

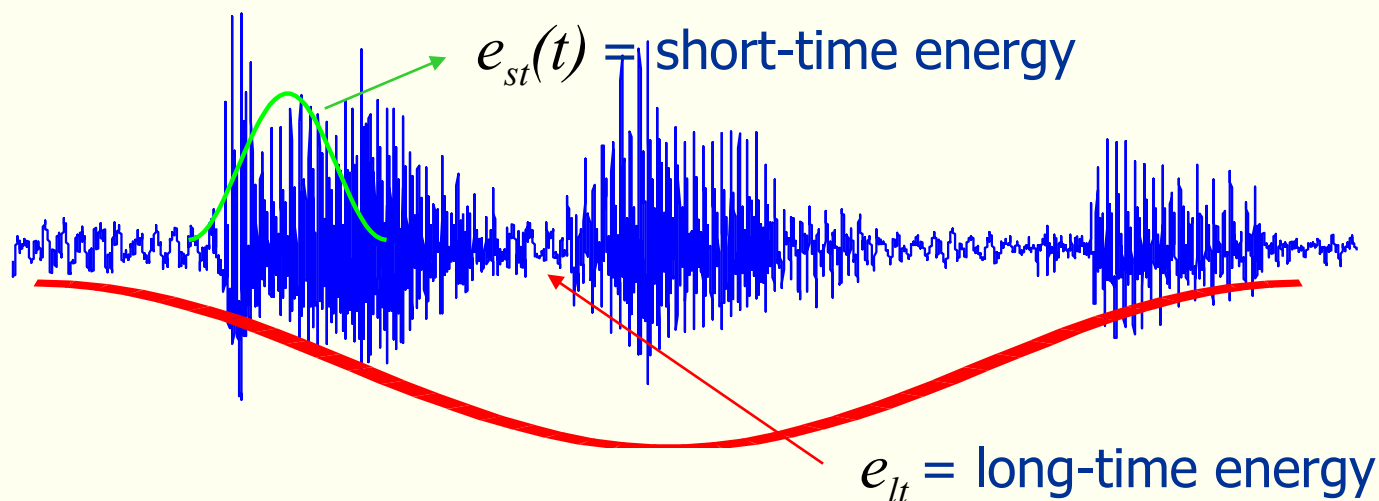
## Kompensacja kanału

- Pojawia się, gdy warunki w jakich przeprowadza się testy różnią się od warunków w jakich następowało uczenie
- Dodatkowym problemem są:
  - ✓ Dźwięki nieświadomie wydawane przez człowieka (cmokanie, ciężki oddech, pstrykanie palcami)
  - ✓ Zakłócenia w systemie transmisyjnym
  - ✓ Warunki środowiskowe (praca maszyn, trzaskanie drzwiami, ruch uliczny, TV, radio, inne rozmowy w tle, nieprzyjazne środowisko-stres)
- Rozwiązaniem może być użycie cech niewrażliwych na powyższe problemy

# Detekcja mowy

- Istnieje wiele technik usuwania „ciszy”
- Wyróżnia się metody:
  - ◆ Zależne od tekstu (wykorzystują statystyczne modele ciszy)
  - ◆ Niezależne od tekstu (oparte na energii sygnału)
- W ogólności detektor mowy usuwa 20-25% sygnału

# Detektory mowy oparte na energii



SNR w chwili  $t$   
można oszacować  
wg wzoru:

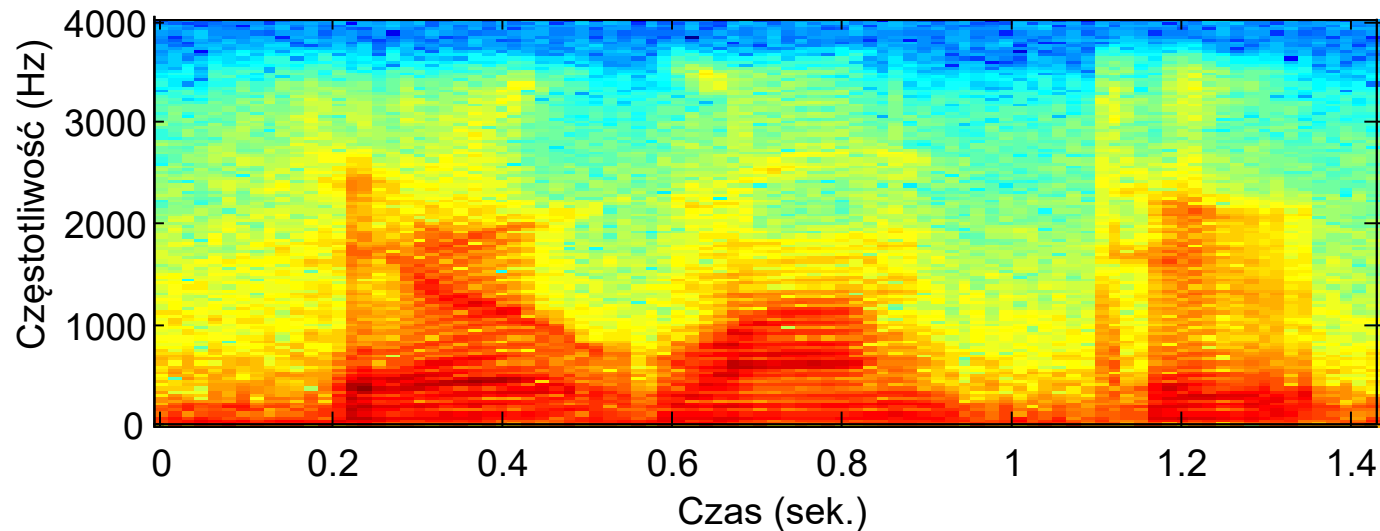
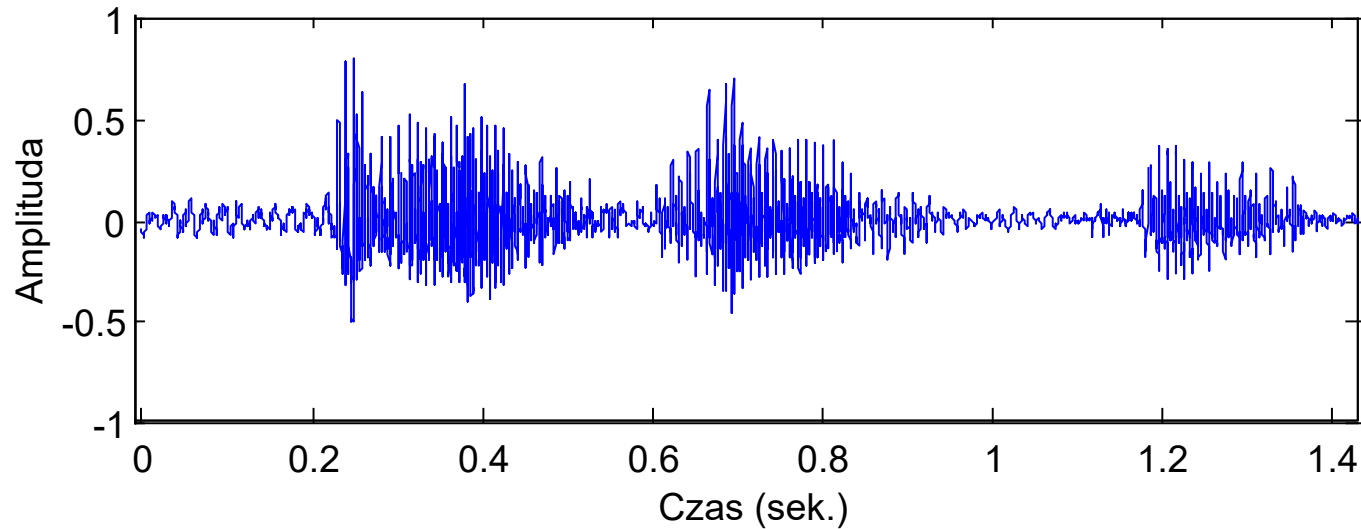
$$\Rightarrow \frac{10\log[e_{st}(t)]}{10\log[e_{lt}]}$$

aktywność  
↓  
>  
<  
↑  
brak aktywności

$\theta_{SNR}$  ← Próg określany doświadczalnie

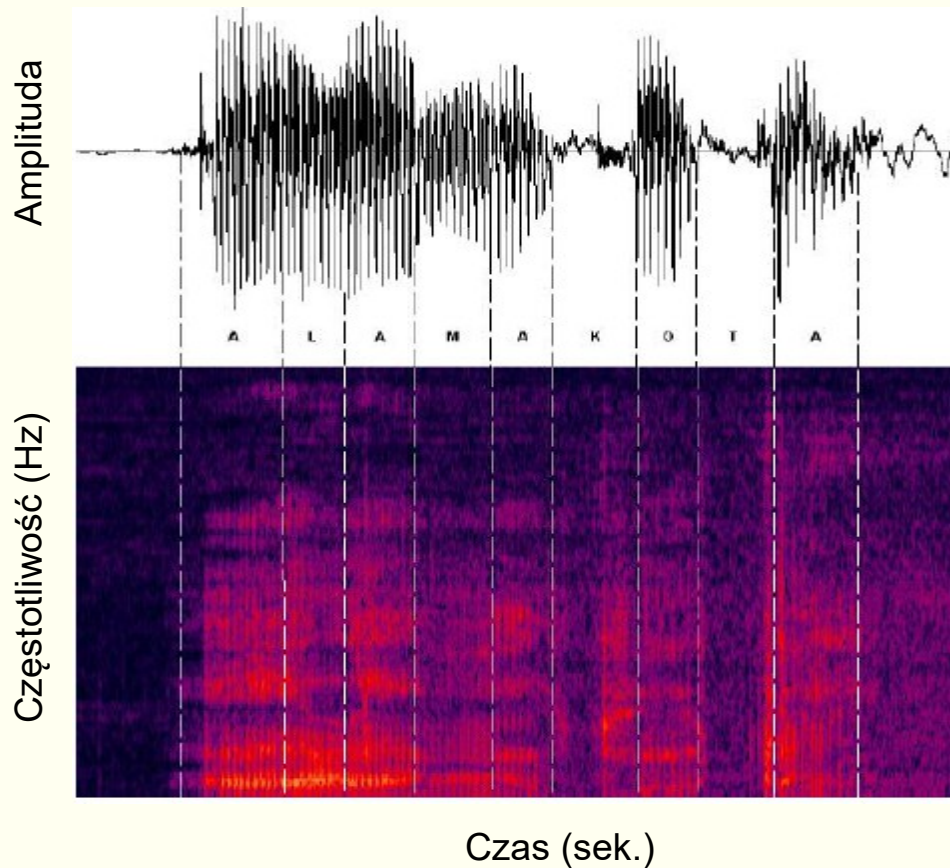
# Spektrogram

Jest to wykres zależności mocy widma w różnych zakresach częstotliwości od czasu



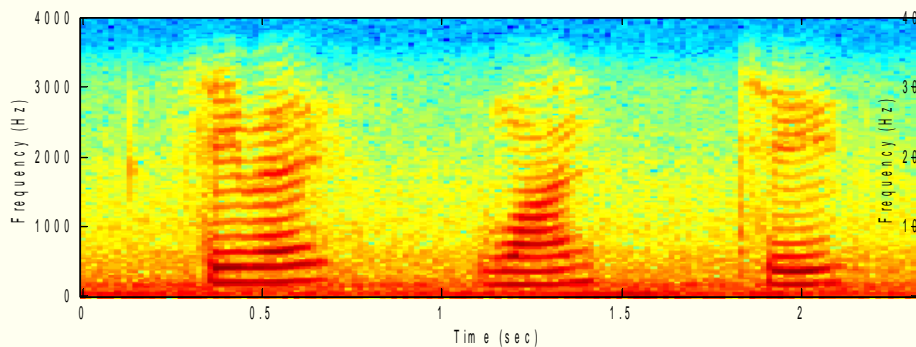
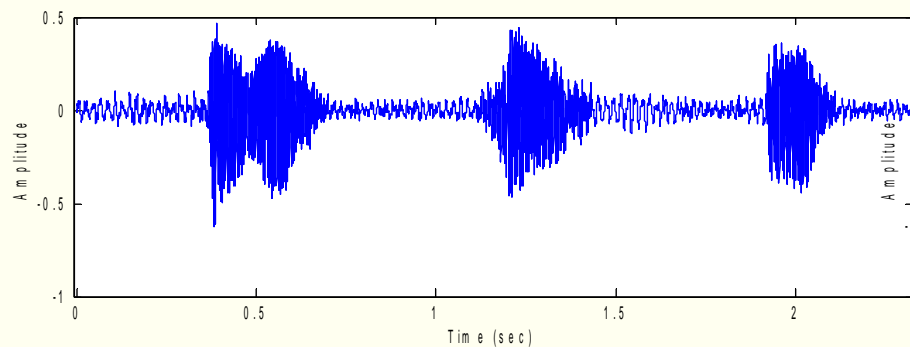
# Spektrogram

Jest to wykres zależności mocy widma w różnych zakresach częstotliwości od czasu

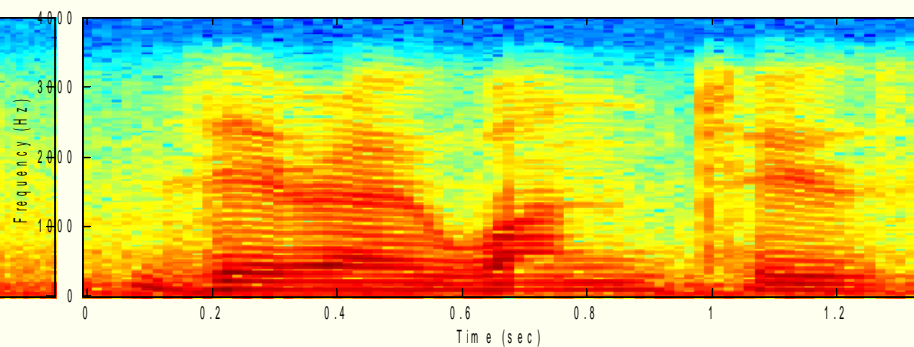
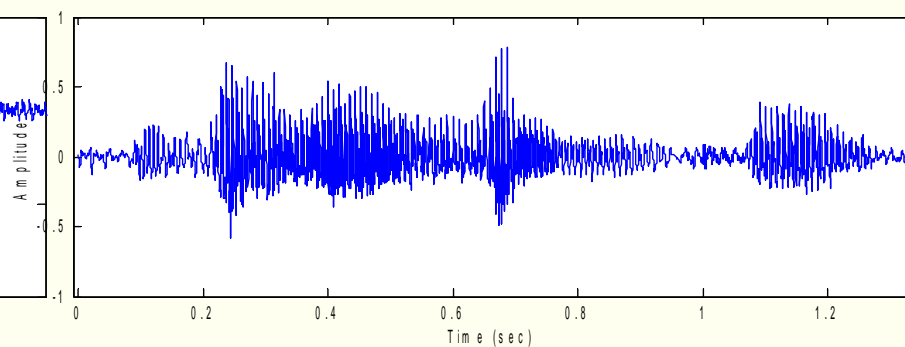


Spektrogram dla zdania *Ala ma kota*.

# Porównanie spektrogramów



Kobieta „0-1-2”



Mężczyzna „0-1-2”

# Formanty

- Formant to maksimum lokalne obwiedni widma sygnału mowy, a częstotliwość przy której występuje to częstotliwość formantowa.
- Główna zaleta formantów polega na ich charakterystycznej konfiguracji, możliwej do określenia w charakterze wzorca dla większości głosek (w tym głównie samogłosek) – niezależnie od tego, kto je wypowiada, jak szybki jest proces artykulacji, jakie towarzyszą mu emocje itp.



# Cechy sygnału mowy

- Istnieje ogólna zgoda, że najlepsza reprezentacja sygnału mowy oparta jest na analizie spektrum
- Techniki analizy spektrum różnią się sposobem kwantyzacji spektrum
- Podstawowe podejścia do kwantyzacji spektrum wykorzystują:
  - ◆ Predykcję liniową (LP – linear prediction)
  - ◆ Bank filtrów

# Cechy sygnału mowy

## ■ Estymaty autokorelacji

- ◆ Autokorelacja
- ◆ Kowariancja
- ◆ Moc widma
- ◆ Korelacja skrośna i kros-PDS

## ■ Funkcja średniej różnic amplitud

## ■ Miary wykorzystujące zerowanie się 2-giej pochodnej

## ■ Moc i energia

## ■ Cechy wykorzystujące analizę Fouriera

# Analiza z użyciem predykcji liniowej

- Z historycznego punktu widzenia jest to jedna z najważniejszych technik analizy mowy
- Następna próbka określana jest na podstawie ważonej sumy  $p$  poprzednich próbek:

$$\hat{s}_n = \sum_{i=1}^p a_i s_{n-i}$$

- Sprowadza się to do zamodelowania procesu mowy z użyciem następującego filtra:

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

Wtedy

$$X(z) = G(z)H(z);$$

gdzie:

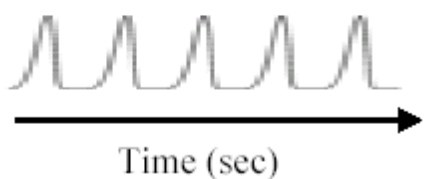
$X$  – spektrum okna sygnału mowy

$G$  – widmo pobudzenia krtaniowego

$z$  – częstotliwość



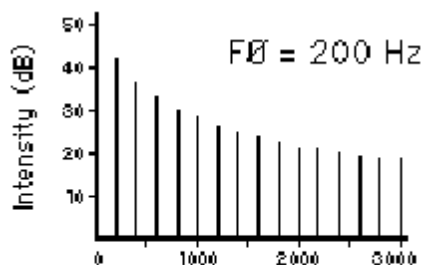
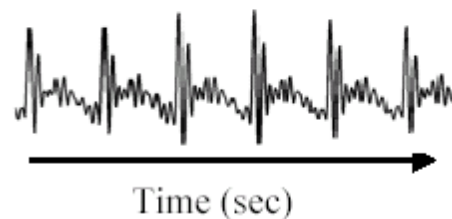
Impulsy  
krtaniowe



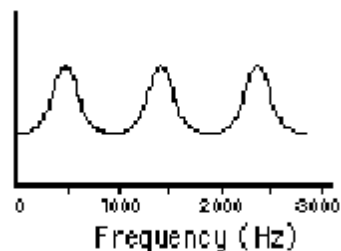
Trakt wokalny



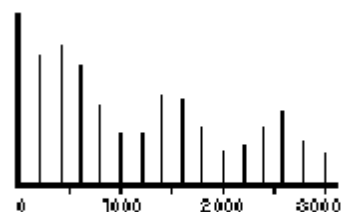
Sygnal mowy



**WIDMO  
ŹRÓDŁOWE**



**FUNKCJA FILTRU**



**WIDMO ENERGII  
WYJŚCIOWEJ**

# Analiza z użyciem predykcji liniowej c.d.

- Polega na doborze takich współczynników predykcji liniowej (**LPC** – Linear Prediction Coefficients), by błąd średniokwadratowy w danym fragmencie sygnału mowy (oknie) był minimalny
- Moduł odpowiedzi filtru reprezentuje spektralną obwiednię okna sygnału mowy
- Uwzględnienie odpowiednio dużej liczby współczynników wystarcza na aproksymację obwiedni widma dla dowolnego dźwięku mowy

# Wyznaczanie LPC

Błąd dla próbki  $n$ :  $e_n = s_n - \hat{s}_n = s_n - \sum_{i=1}^p a_i s_{n-i}$

SSE w oknie o długości  $N$ :  $E = \sum_{n=0}^{N-1} e_n^2 = \sum_{n=0}^{N-1} \left( s_n - \sum_{k=1}^p a_k s_{n-k} \right)^2$

$E$  osiąga minimum, gdy  $\delta E / \delta a_j = 0$ :

$$\frac{\partial E}{\partial a_j} = - \sum_{n=0}^{N-1} \left( 2 \left( s_n - \sum_{k=1}^p a_k s_{n-k} \right) s_{n-j} \right) = -2 \sum_{n=0}^{N-1} s_n s_{n-j} + 2 \sum_{n=0}^{N-1} \sum_{k=1}^p a_k s_{n-k} s_{n-j} = 0$$

⇓

$$\sum_{n=0}^{N-1} s_n s_{n-j} = \sum_{n=0}^{N-1} \sum_{k=1}^p a_k s_{n-k} s_{n-j} = \sum_{k=1}^p \sum_{n=0}^{N-1} s_{n-k} s_{n-j}$$

Utwórzmy macierz kowariancji  $\Phi$  o elementach  $\varphi_{i,k}$ :

Można pokazać, że:  $\varphi_{i,0} = \sum_{k=1}^p \varphi_{i,k} a_k$   $\varphi_{i,k} = \sum_{n=0}^{N-1} s_{n-i} s_{n-k}$

W postaci macierzowej:  $\Phi_0 = \Phi \cdot \mathbf{a} \Rightarrow \mathbf{a} = \Phi^{-1} \cdot \Phi_0$

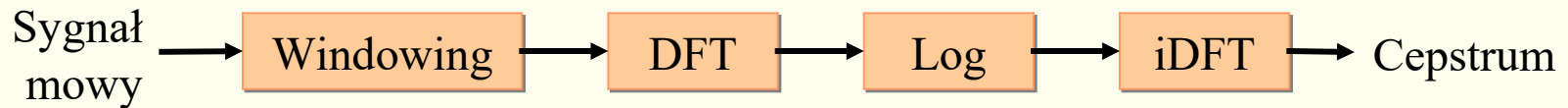
# Cechy analizy z użyciem predykcji liniowej

- LPC są wrażliwe na zakłócenia wynikające z różnych warunków uczenia i testowania – lepiej jest oprzeć się na reprezentacji cepstralnej
- Współczynniki cepstralne liniowej predykcji (**LPCCs** – Linear Predictive Cepstral Coefficients) można wyznaczyć na podstawie współczynników LPC wykorzystując następującą regułę iteracyjną:

$$c_n = a_n + \frac{1}{n} \sum_{i=1}^{n-1} i c_i a_{n-i}$$

# Cepstrum

- Cepstrum to odwrotna transformacja Fouriera (lub podobna) logarytmu mocy widma sygnału



- Definicja

$$C(q) = 2T\{\ln|G(f)| + \ln|H(f)|\}$$

gdzie:

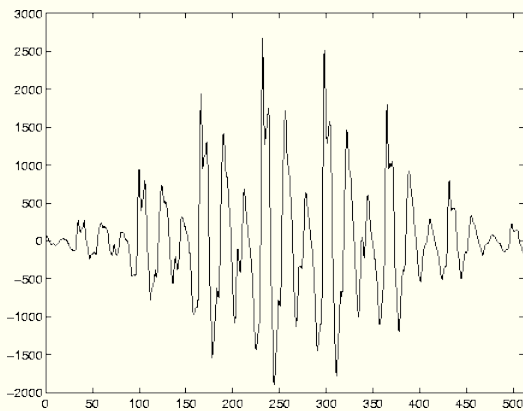
$C(q)$  – cepstrum

$T\{\}$  – transformacja (zazwyczaj DCT)

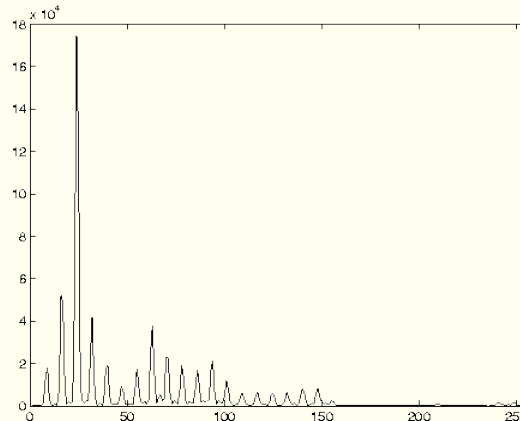
$q$  – quefrequency



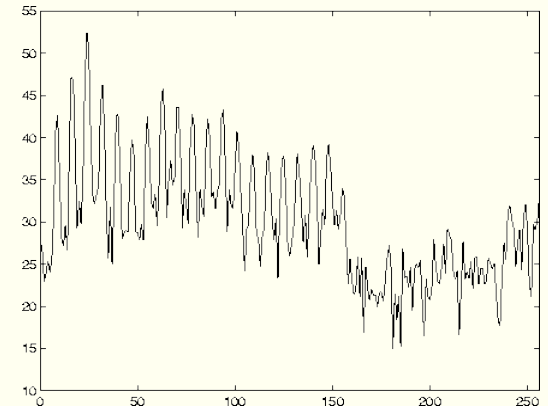
# Cepstrum - przykład



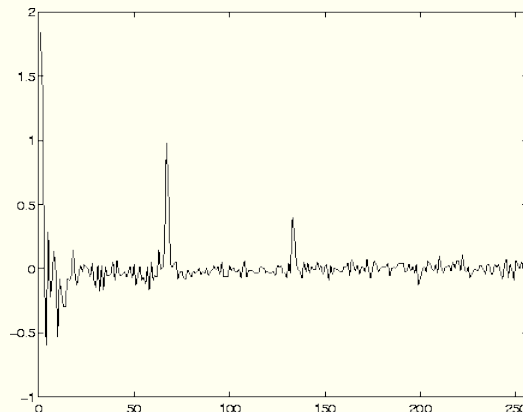
(a)



(b)



(c)



(d)

Segment samogłoski dla okna Hamminga (a)

i odpowiadające mu:

(b) moduł widma

(c) moc widma (w dB)

(d) część rzeczywista cepstrum

# Cechy cepstrum

- Większość szczegółów pojawia się na początku i w wierzchołkach cepstrum  $\Rightarrow$  pierwsze współczynniki zawierają informację o obwiedni mocy widma, a szczegóły dotyczące pobudzenia krtaniowego reprezentowane są w większości przez pojawiające się okresowo wierzchołki
- Cechy są nieskorelowane, ze względu na transformację  $T\{\}$  (zazwyczaj DCT)
- Wygodne i efektywne w obecności zakłóceń
- Ze względu na te cechy prawie zawsze stosowana jest reprezentacja cepstralna

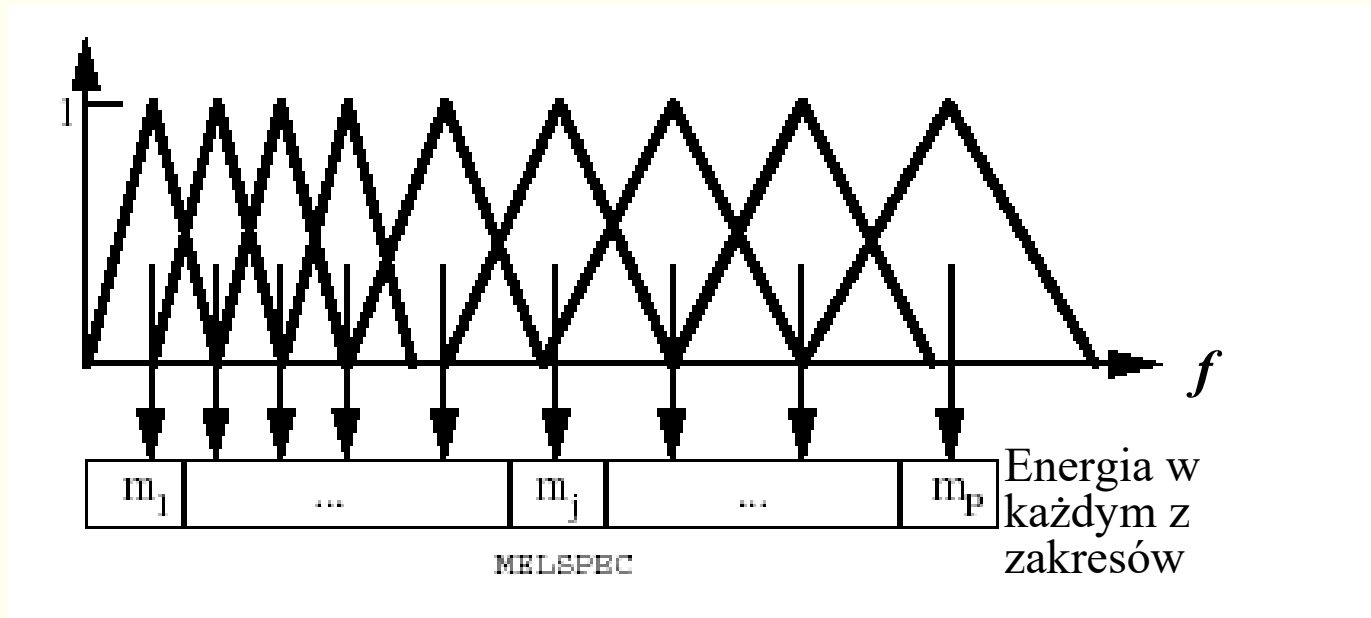
# Cechy oparte na analizie z użyciem banku filtrów

- Stosowane banki filtrów odzwierciedlają nieliniową wrażliwość ucha ludzkiego
- Operują bezpośrednio na widmie
- Uwzględniają amplitudy  $m_i$  wyznaczone dla każdego zakresu nieliniowo rozmieszczonych filtrów
- Inżynierowie wykorzystują bank trójkątnych filtrów w skali Mela

# Skala Mela

- Definicja skali Mela:

$$Mel(f) = 2595 \log_{10}(1 + f / 700)$$



# Cechy MFCCs

- Oparte na reprezentacji cepstralnej:

$$c_{mel}(n) = \sqrt{\frac{2}{N}} \sum_{i=1}^N m_i \cos\left(\frac{\pi n}{N}(i - 0.5)\right)$$

gdzie:

$N$  – liczba filtrów w banku

$m_i$  – wartość energii dla  $i$ -tego filtru

- Znane jako współczynniki cepstralne dla częstotliwości Mela (**MFCCs** – Mel-Frequency Cepstral Coefficients)

# Metody klasyfikacji cech

## ■ DTW – Dynamic Time Warping

- ◆ Pozwala obliczyć podobieństwo dwóch sekwencji (np. czasowych) o różnej długości,
- ◆ Metoda niewrażliwa na nieliniowe transformacje wzdłuż osi czasu

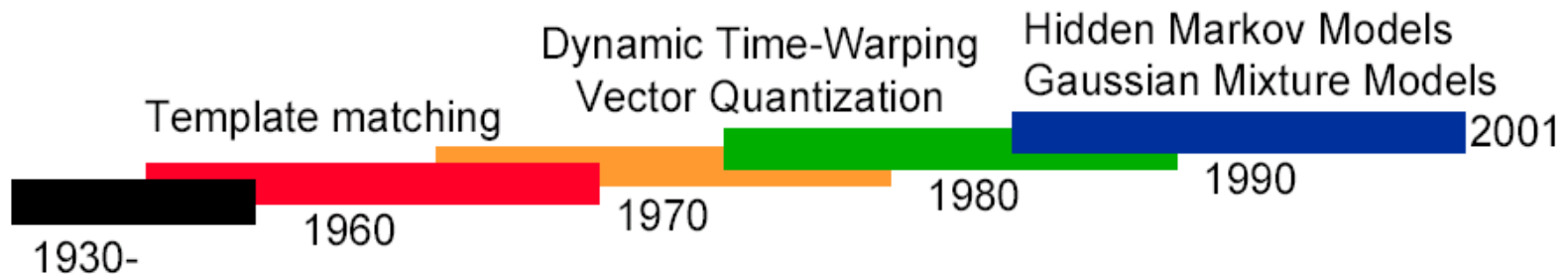
## ■ Ukryte modele Markova (HMM – Hidden Markov Models)

- ◆ Są to modele statystyczne, w których zakłada się, że modelowany system jest procesem Markova, którego stan da się zaobserwować
- ◆ Każde słowo/fonem posiada oddzielny model, którego parametry określa się na drodze uczenia
- ◆ Model, dla którego prawdopodobieństwo na wyjściu jest największe określa klasę

# Ewolucja systemów rozpoznawania mowy

Porównywanie dźwięku  
i spektrogramu z wzorcem

Aplikacje  
komercyjne



Małe bazy danych,  
mowa czysta,  
kontrolowana

Duże bazy danych,  
mowa rzeczywista,  
niekontrolowana

# Narzędzia do rozpoznawania mowy dla programistów

- API do silnika rozpoznawania mowy firmy Google (Google Cloud Speech API)
  - ◆ Rozpoznaje ponad 80 języków
  - ◆ Umożliwia rozpoznawanie mowy niezależnie od platformy
- Open source speech decoder *Julius*:
  - ◆ Daje możliwość rozpoznawanie dużego słownika wyrazów mowy ciągłej
  - ◆ W oparciu o to narzędzie powstał system *Skrybot* ([skrybot.pl](http://skrybot.pl)) rozpoznawania mowy polskiej
- Java Speech API:
  - ◆ Złożone z 3 pakietów:
    - javax.speech,
    - javax.speech.synthesis,
    - javax.speech.recognition



# Narzędzia do rozpoznawania mowy dla programistów c. d.

## ■ HTK

- ◆ Zbiór modułów bibliotecznych i programów narzędziowych wspierających rejestrowanie i przetwarzanie sygnału mowy, konstruowanie złożonych układów HMM, uczenie, testowanie i analizę rezultatów
- ◆ Opracowane w zespole Speech, Vision and Robotics na Uniwersytecie Cambridge
- ◆ Pierwotnie przeznaczone do rozpoznawania mowy, jednak ukryte modele Markova budowane z użyciem jądra HTK mogą być stosowane do modelowania dowolnych przebiegów czasowych (synteza mowy, rozpoznawanie sekwencji DNA, pisma, gestów)

## ■ Microsoft Speech Recognition API:

- ◆ Przestrzeń nazw `Windows.Media.SpeechRecognition`

# Narzędzia do rozpoznawania mowy dla programistów c. d.



## ■ HTK

- ◆ Zbiór modułów bibliotecznych i programów narzędziowych wspierających rejestrowanie i przetwarzanie sygnału mowy, konstruowanie złożonych układów HMM, uczenie, testowanie i analizę rezultatów
- ◆ Opracowane w zespole Speech, Vision and Robotics na Uniwersytecie Cambridge
- ◆ Pierwotnie przeznaczone do rozpoznawania mowy, jednak ukryte modele Markova budowane z użyciem jądra HTK mogą być stosowane do modelowania dowolnych przebiegów czasowych (synteza mowy, rozpoznawanie sekwencji DNA, pisma, gestów)

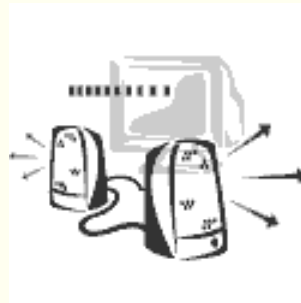
## ■ Microsoft Speech Recognition API:

- ◆ Przestrzeń nazw `Windows.Media.SpeechRecognition`

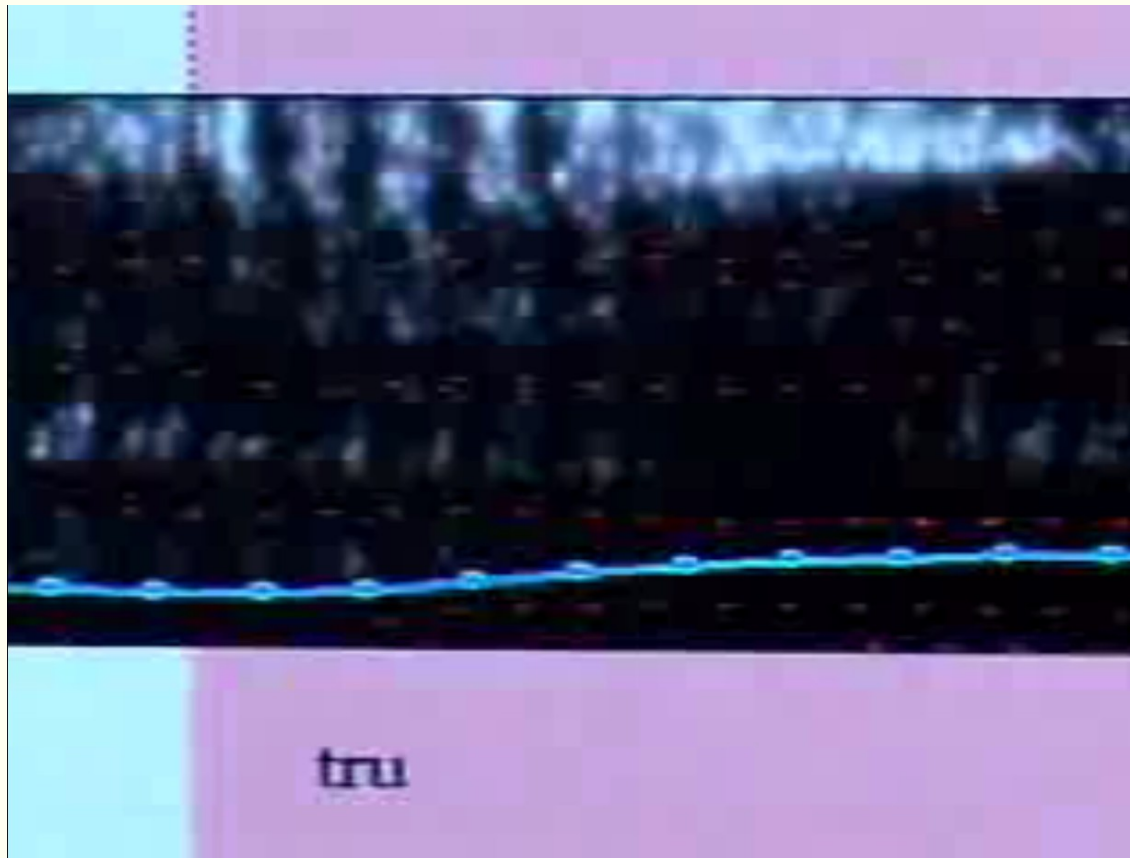
# Systemy komercyjne

- System rozpoznawania mowy polskiej i syntezyator Primespeech (<http://www.primespeech.pl>) - stosowany przez
  - ◆ Giełdę Papierów Wartościowych (telefoniczne informowanie o aktualnych notowaniach spółek giełdowych )
  - ◆ Korporację taksówkową (zamawianie taxi),
  - ◆ infolinię Zarządu Transportu Miejskiego w W-wie,
  - ◆ Polsko-Japońską Wyższą Szkołę Technik Komputerowych (przełączanie głosowe telefonu, telefoniczny serwis informacyjny dla studentów, aktualności )

# Synteza mowy



# Syntezator mowy Krzysztofa Szklanego - prezentacja



# Systemy komercyjne

- Syntezator mowy polskiej IVOVA ([http://www.ivona.com/film IVONA w TVN](http://www.ivona.com/film_IVONA_w_TVN)), cena od 184 zł (IVONA Reader z 2 głosami)
- Syntezator i system rozpoznawania mowy Primespeech (<http://www.primespeech.pl>)

# System rozpoznawania mowy wspomagany informacją wizyjną

