Przetwarzanie dźwięków i obrazów

ROZPOZNAWANIE SYGNAŁÓW FONICZNYCH

mgr inż. Kuba Łopatka, p. 628
klopatka@sound.eti.pg.gda.pl

Plan wykładu

1. Wprowadzenie
2. Zasada rozpoznawania sygnałów
3. Parametryzacja
4. Klasyfikacja
5. Przykładowe metody

Rozpoznawanie sygnałów

- Uzyskanie wiedzy w sposób automatyczny z liczbowej reprezentacji zjawiska fizycznego (np. dźwięku, obrazu, sygnałów z sensorów)
- Przykłady
  - obraz – rozpoznawanie twarzy, rozpoznawanie obiektów, rozpoznawanie znaków - OCR
  - dźwięk – rozpoznawanie mowy, zdarzeń dźwiękowych, rozpoznawanie muzyki

Rozpoznawanie sygnałów

Warianty rozpoznawania sygnałów

- klasyfikacja – przyporządkowanie sygnału nieznanego typu do danej klasy
- weryfikacja – potwierdzenie przynależności obiektu do klasy
- rozpoznawanie statyczne (np. na klatce obrazu, całym pliku dźwiękowym)
- rozpoznawanie dynamiczne – z uwzględnieniem wewnętrznych zmian w sygnale

Rozpoznawanie sygnałów

- Metody statystyczne

Parametryzacja – opisanie rozpoznawanego obiektu za pomocą skończonego zestawu liczb – wektora parametrów

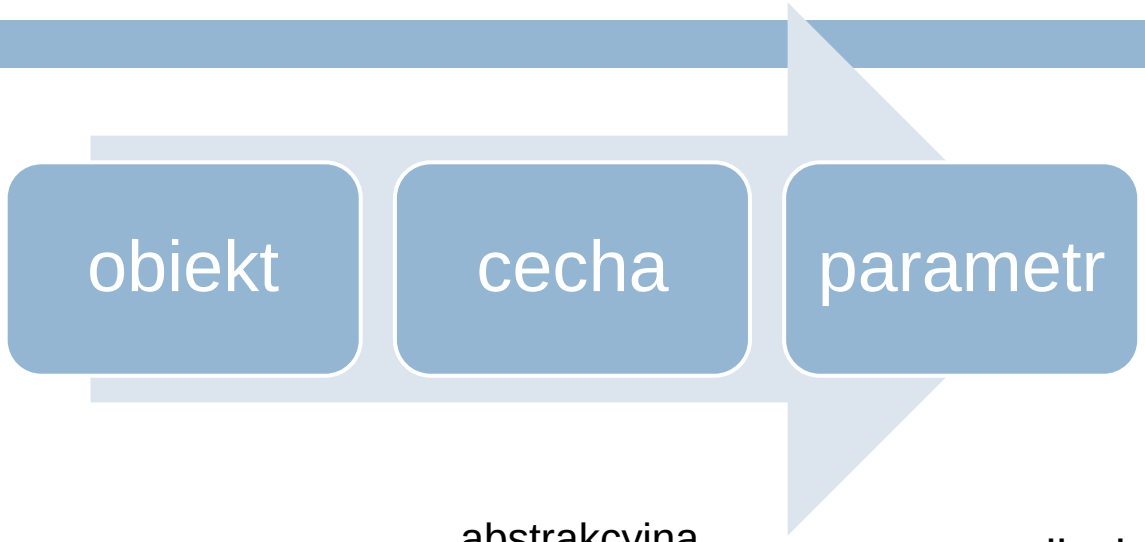
Klasyfikacja – najczęściej z wykorzystaniem inteligentnych systemów decyzyjnych: sieci neuronowe, drzewa decyzyjne, SVM...

Rozpoznawanie sygnałów

Parametryzacja

- Komputery operują na liczbach, nie na abstrakcyjnych cechach.
- Abstrakcyjne cechy obiektu, oparte na subiektywnych wrażeniach, można mnożyć w nieskończoność. Parametryzacja uściśla i formalizuje opis obiektu.
- Wykorzystanie parametrów i ich analizy pozwala nam czasem zauważyć różnice, z których istnienia nie zdawaliśmy sobie sprawy.

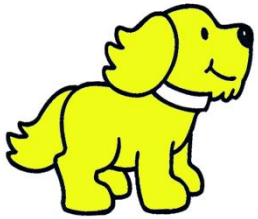
Parametr a cecha obiektu



coś, co potrafimy wyróżnić – dotknąć, nazwać, wskazać

abstrakcyjna jakość związana z obiektem, odróżniająca go od innych obiektów

liczbowe wyrażenie cechy obiektu

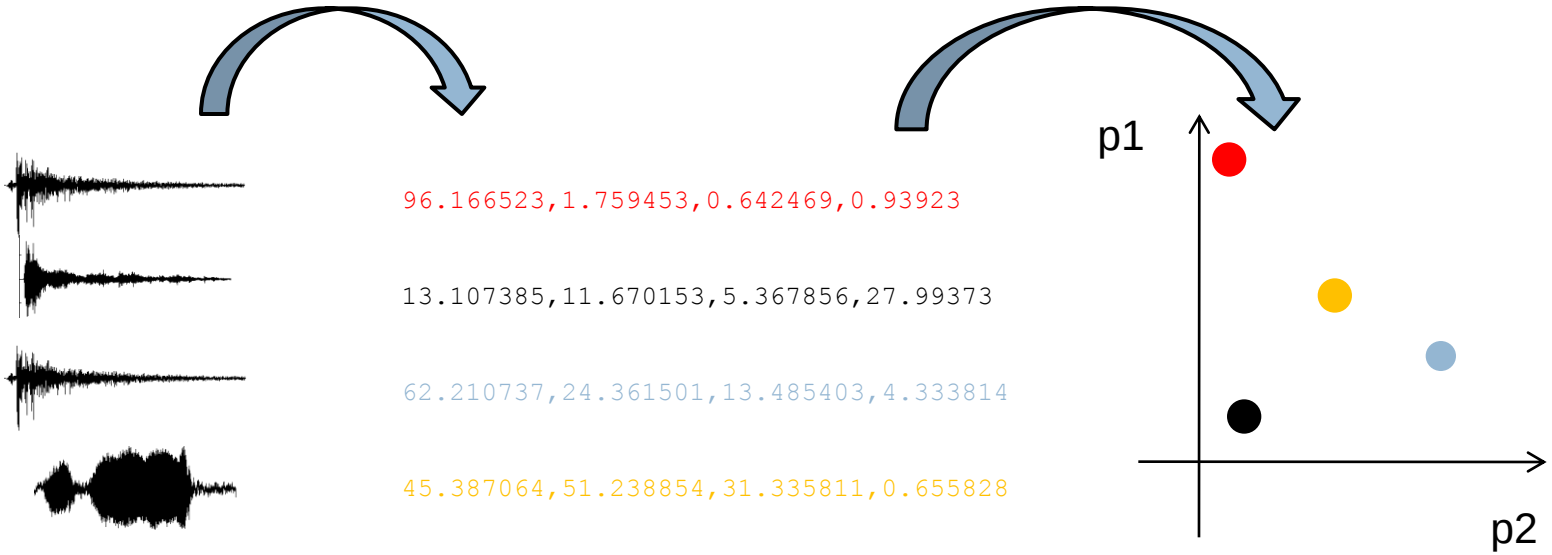


Fazor

żółty
krótkie łapy
duże uszy

kolor = 0xEFFD16
długość łap = 0.3
powierzchnia uszu = 2.5

Istota parametryzacji

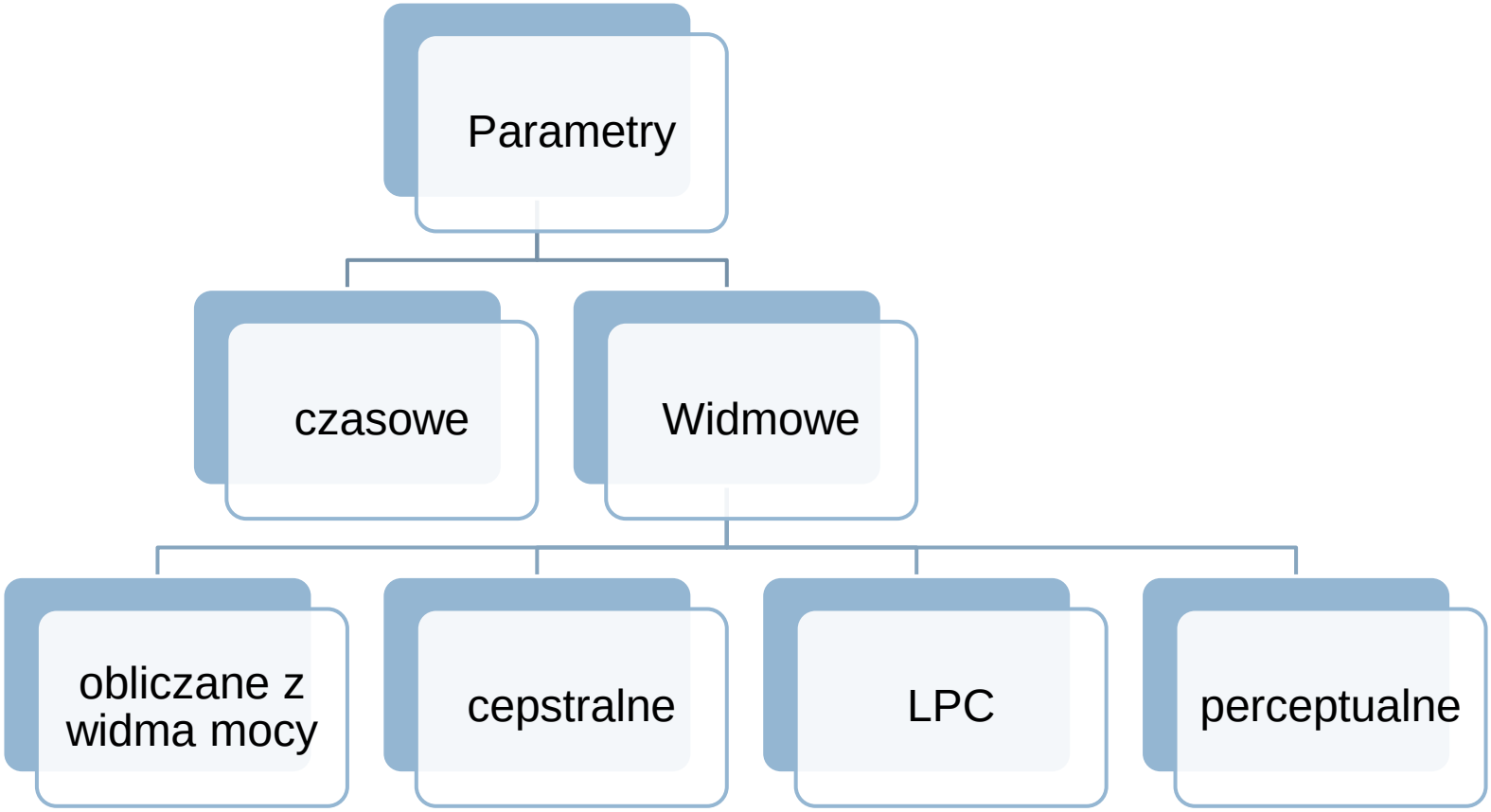


Odwzorowanie dowolnych obiektów za pomocą wektorów liczb o skończonej długości - przejście z przestrzeni o nieskończonym wymiarze do przestrzeni K parametrów.

Cel parametryzacji

- Odróżnienie od siebie obiektów różnych klas
- Rozpoznanie obiektu nieznanej klasy
- Weryfikacja przynależności obiektu do klasy

Parametry dźwięku



Parametry czasowe

Parametry **czasowe** są to parametry, które są wyznaczane wyłącznie na podstawie postaci czasowej sygnału.

Przykłady:

- Energia sygnału
- Środek ciężkości sygnału
- Obwiednia sygnału
- Gęstość przejść przez zero

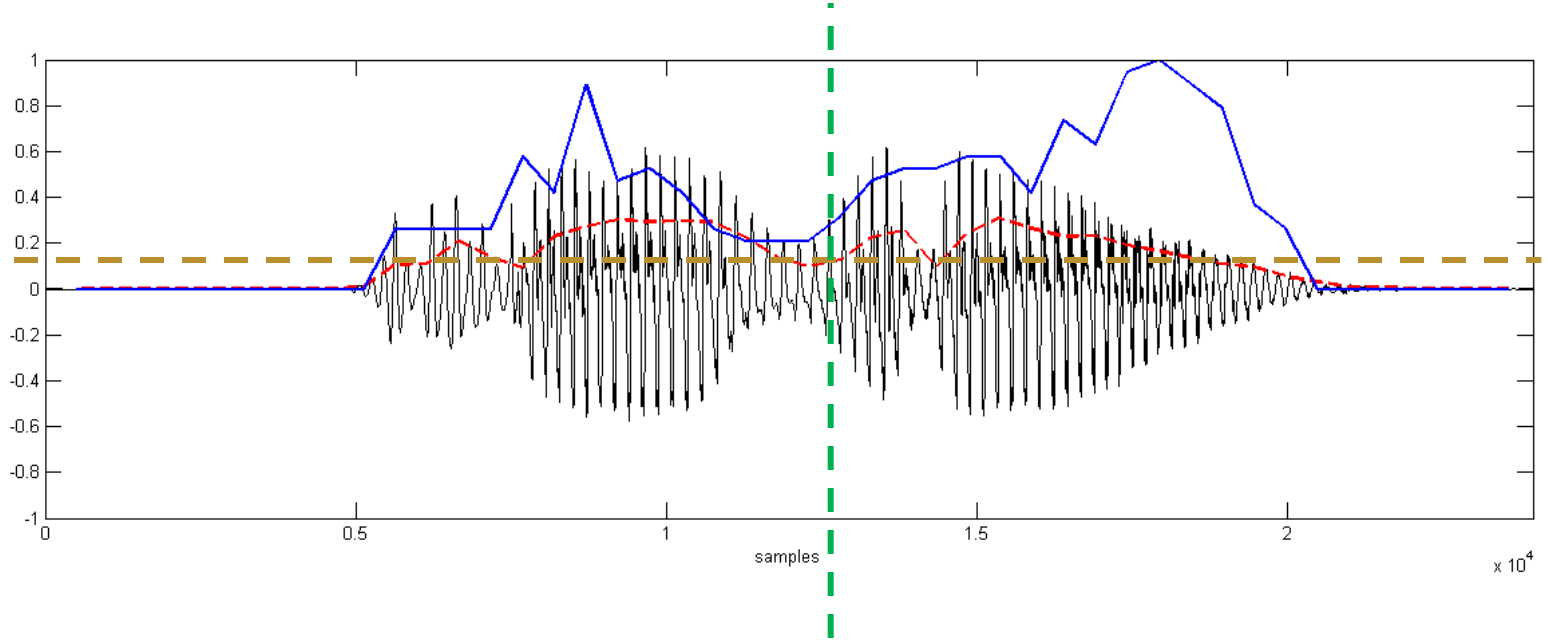
$$E = \sum_{n} x^2(n)$$

$$TC = \frac{\sum_{n=1}^{N} n \cdot O(n)}{\sum_{n=1}^{N} O(n)}$$

$$O(n) = \{o_1 \quad o_2 \quad \ldots \quad o_N\}$$

$$o_n = \sqrt{\frac{1}{K} \sum_{n=1}^{K} x_n^{\,2}(k)}$$

Parametry czasowe



Obwiednia

Gęstość przejść przez zero

E=0.1627

TC=12581 [smpl]

Parametry widmowe

- Parametry widmowe wyznaczane są na podstawie estymaty widma sygnału.

Estymacja widma sygnału:

- Funkcja widmowej gęstości mocy (*power spectral density – PSD):* periodogram, estymator Welcha, autokorelacja – **widmo mocy**
- Moduł DFT sygnału – **widmo amplitudowe**

Momenty widmowe

- Momenty widmowy *m*-tego rzędu definiuje się następująco:

$$M(m)=\sum_{k=0}^{\infty}|G(k)|\cdot[f_k]^m$$

gdzie: *G(k)* – wartość widma mocy dla *k*-tego pasma częstotliwości
f_k – częstotliwość środkowa *k*-tego pasma

- Moment unormowany *m*-tego rzędu

$$M_u(m)=\frac{M(m)}{M(0)}$$

- Moment normalizujący zerowego rzędu ma sens mocy sygnału

$$M(0)=\sum_{k=0}^{\infty}|G(k)|$$

Płaskość widmowa

- Płaskość widmowa (ang. *spectral flatness measure* – *SFM*) – stosunek średniej geometrycznej i arytmetycznej współczynników widma – miara harmoniczności sygnału

$$SFM = 10 \cdot \log \left\{ \frac{\left[\prod_{k=1}^{N/2} P\left(e^{j\frac{2\pi k}{N}} \right) \right]^{1/N/2}}{\frac{1}{N/2} \cdot \sum_{k=1}^{N/2} P\left(e^{j\frac{2\pi k}{N}} \right)} \right\}$$

$P\left(e^{j\frac{2\pi k}{N}} \right)$ - widmowa gęstość mocy

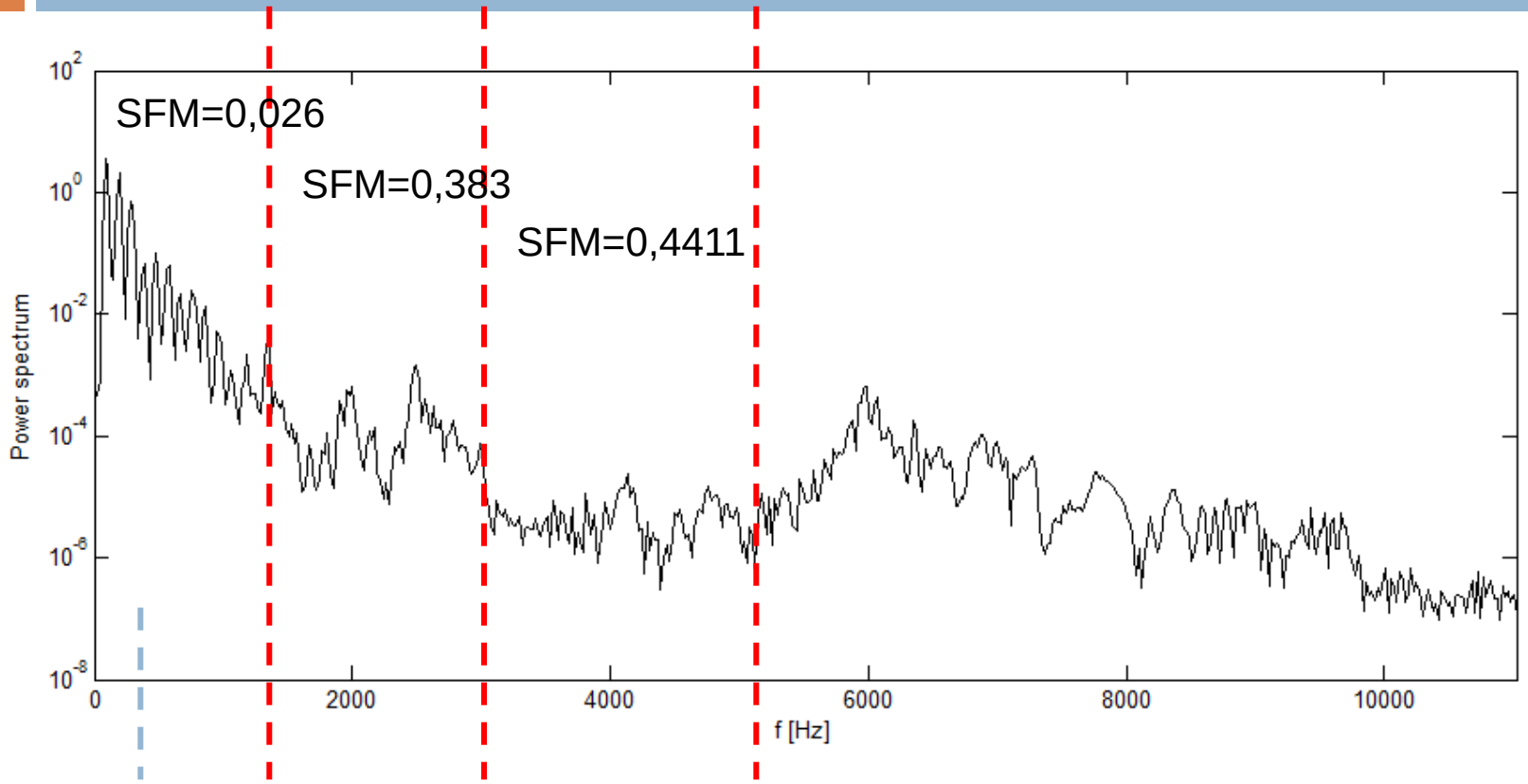
Parametr SFM może być również wyznaczany w pasmach

MPEG-7

- Ogrom parametrów widmowych (i nie tylko) zdefiniowanych jest w tzw. standardzie MPEG-7. Na przykład:
- Audio Spectrum Envelope
- Audio Spectrum Spread
- Audio Spectrum Centroid
- Harmonic Spectral Centroid
- Harmonic Spectral Spread
- Audio Spectrum Flatness
- ...

Najczęściej są one stosowane do sygnałów muzycznych

Parametry widmowe



Mu1=178,34 - środek ciężkości widma

Muc3 = 1,8 * 10^8 – skośność

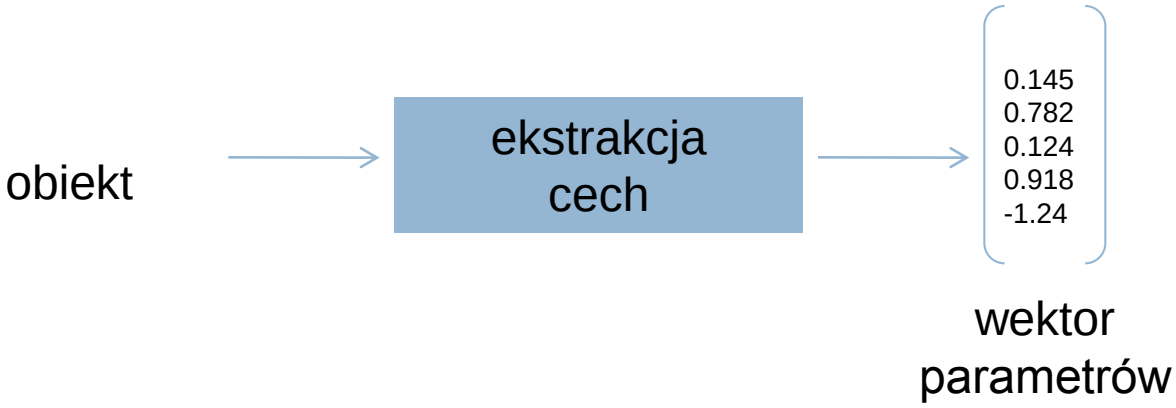
Kurtoza – 506.4653

Ekstrakcja parametrów

Proces obliczania parametrów nazywa się często **ekstrakcją cech obiektu** (lub *cech sygnału*), ang. *feature extraction*.

Ekstrakcja cech powinna być procesem powtarzalnym, deterministycznym i sformalizowanym matematycznie.

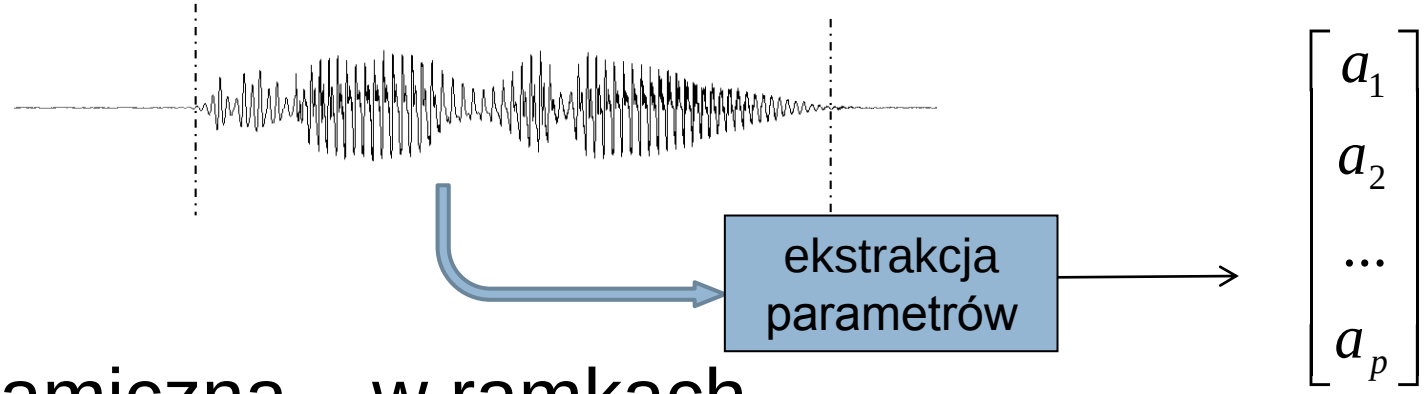
Wynikiem ekstrakcji parametrów jest **wektor cech** związany z obiektem.



Ekstrakcja parametrów

- statyczna – na całym sygnale

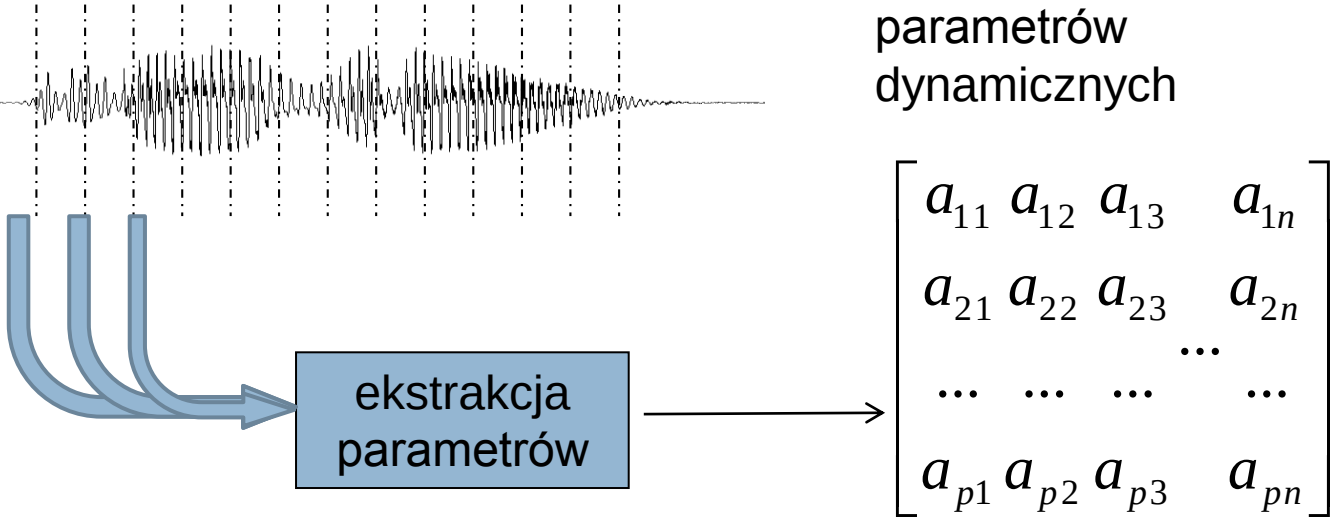
wektor parametrów



- dynamiczna – w ramkach

ramki np. o długości 25ms

macierz parametrów dynamicznych



Klasyfikacja sygnałów

Zadaniem **klasyfikatora** jest przyporządkowanie obiektów nieznanego typu do jednej ze znanych klas.

Aby algorytm był zdolny do takiego przyporządkowania, potrzebny jest **trening**, podczas którego tworzy się **model.**

W ujęciu matematycznym klasyfikator jest układem, który na wejściu przyjmuje wektor cech, a na wyjściu daje wynik klasyfikacji.

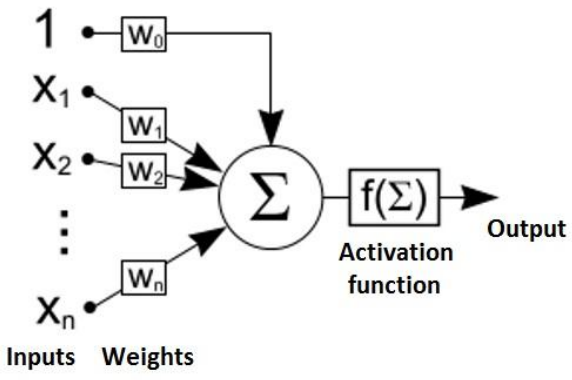
Klasyfikacja

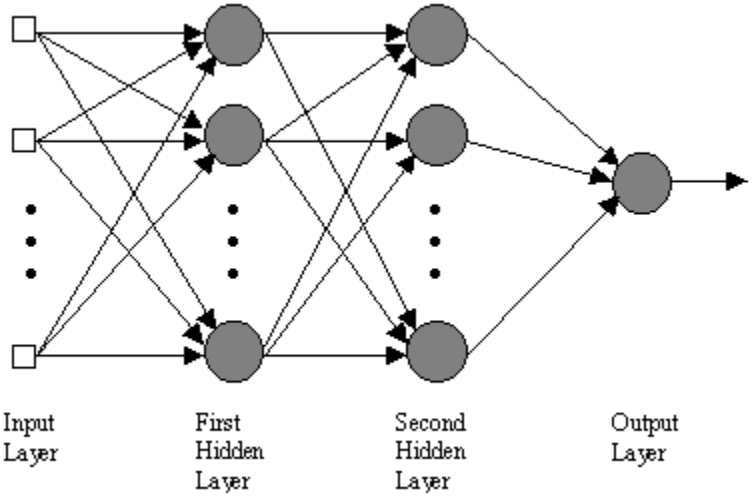
Metody klasyfikacji

- sztuczne sieci neuronowe (ANN)
- maszyny wektorów wspierających (SVM)
- ukryte modele Markowa (HMM)
- mieszane modele Gaussowskie (GMM)
- inne algorytmy statystyczne...

Sztuczne sieci neuronowe

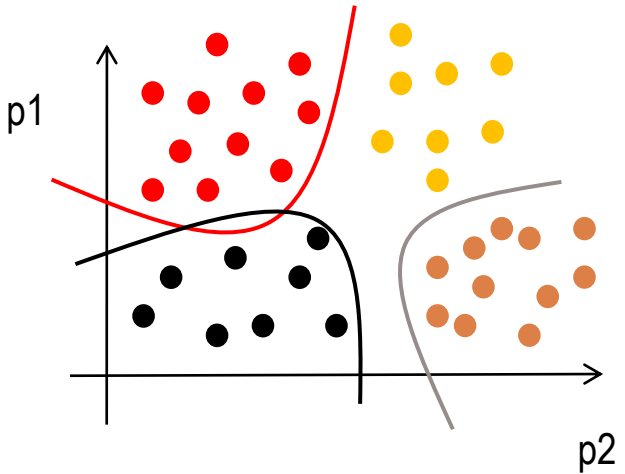
- Struktura warstwowa zbudowana ze sztucznych neuronów
- Trening – dostosowanie wag (najczęściej algorytm gradientowy)

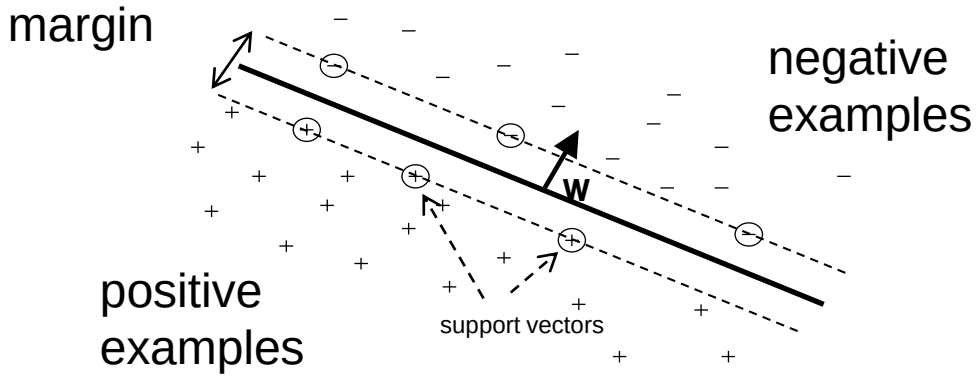




Maszyny wektorów wspierających

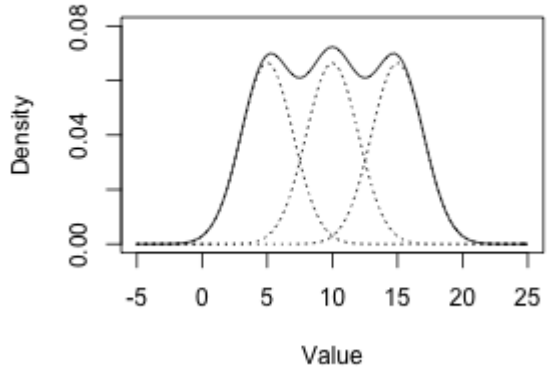
- Znajdowanie optymalnej hiperpłaszczyzny (w N-wymiarowej przestrzeni) separującej **dwie** klasy
- Dla rozpoznania więcej niż dwóch klas konieczne jest stworzenie większej liczby modeli





Mieszane modele Gaussowskie

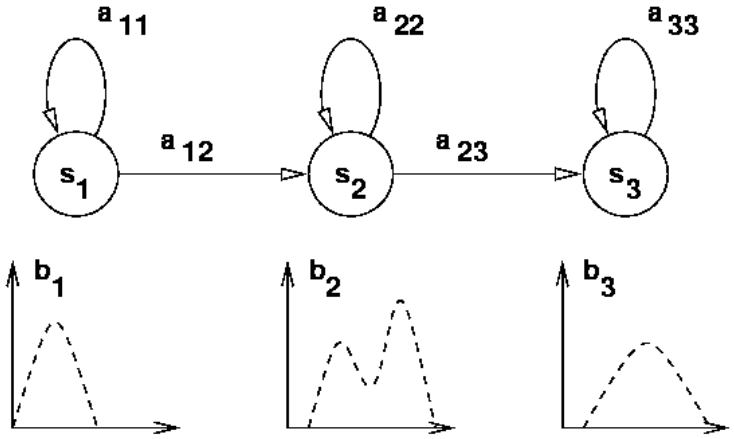
- Opisywanie rozkładów prawdopodobieństwa wartości parametrów danych klas za pomocą sumy rozkładów Gaussa
- obiekt przyporządkowany do klasy o największym prawdopodobieństwie *a posteriori*



$$p(\mathbf{x}|\lambda) = \sum_{i=1}^{M} w_i \; g(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i),$$

Ukryte modele Markowa

- Ukryte modele Markowa (Hidden Markov Model) zawierają dynamiczny model, w którym zdefiniowane są stany i rozkłady prawdopodobieństwa przejść między stanami
- Estymowane prawdopodobieństwo wystąpienia zaobserwowanej sekwencji



Klasyfikacja

Najprostszy przypadek – dwie klasy

Wektory uczące w postaci par (x_i, y_i)

gdzie x_i to wektor parametrów y_i – klasa {1;2}

Funkcja klasyfikacji

$$y = f(x) = \begin{cases} 1, & warunek\,klasy1 \\ 2, & warunek\,klasy2 \end{cases}$$

Klasyfikacja

- Macierz pomyłek (confusion matrix)

klasa 1	klasa 2	← sklasyfikowano jako
TP1	FP2	klasa 1
FP1	TP2	klasa 2

- TP – True Positive – wartość prawdziwie pozytywna
- FP – False Positive – wartość fałszywie pozytywna
- precision – pewność w klasie 1 = TP1/(TP1+FP1)
- recall – czułość w klasie 2 = TP1/(TP2+FP2)

Klasyfikacja

- Przykład

klasa 1	klasa 2	← sklasyfikowano jako
127	0	klasa 1
54	66	klasa 2

precision – pewność w klasie 1 = 127/(127+54) = 0,7
recall – czułość w klasie 1 = 127/(127+0) = 1
średnia skuteczność = (127+66)/(127+66+54)=78%

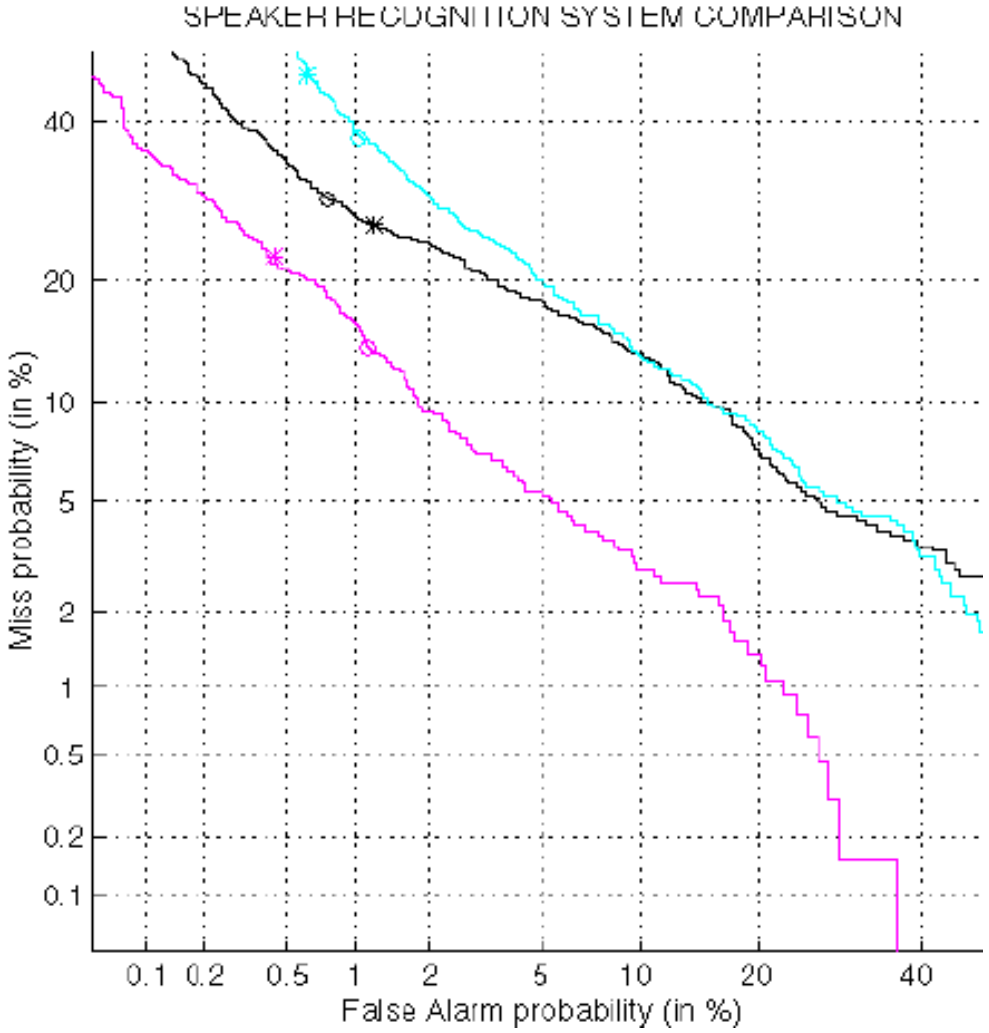
klasa 1	klasa 2	← sklasyfikowano jako
127	28	klasa 1
26	66	klasa 2

precision – pewność w klasie 1 = 127/(127+26) = 0,83
recall – czułość w klasie 1 = 127(127+28) = 0,82
średnia skuteczność = (127+66)/(127+66+54)=78%

Wykres DET

- FAR – False Acceptance Rate, np. przyznanie dostępu osobie nieuprawnionej
- FRR – False Rejection Rate, np. nieprzyznanie dostępu osobie uprawnionej
- Wykres DET (Detection Error Tradeoff) obrazuje FRR w funkcji FAR w zależności od czułości

Wykres DET


SPEAKER RECOGNITION SYSTEM COMPARISON
Miss probability (in %)
40
20
10
5
2
1
0.5
0.2
0.1
0.1 0.2 0.5 1 2 5 10 20 40
False Alarm probability (in %)

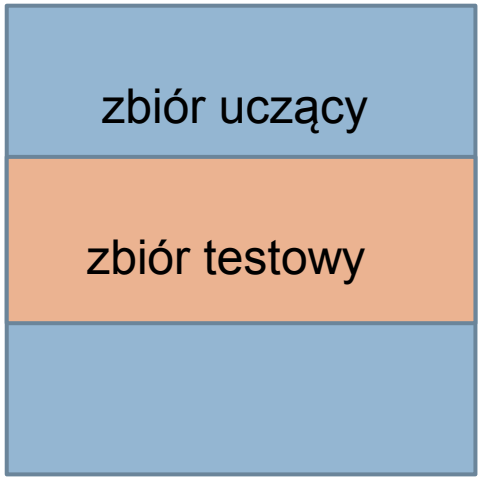
Trening

- Zbiór uczący – wektory parametrów + znane klasy wykorzystywane do treningu klasyfikatora
- Zbiór testowy – wektory parametrów nieznane na etapie treningu sprawdzające działanie klasyfikacji
- Walidacja krzyżowa – wielokrotny podział zbioru treningowego na zbiór uczący i testowy celem przetestowania klasyfikatora na każdym wektorze

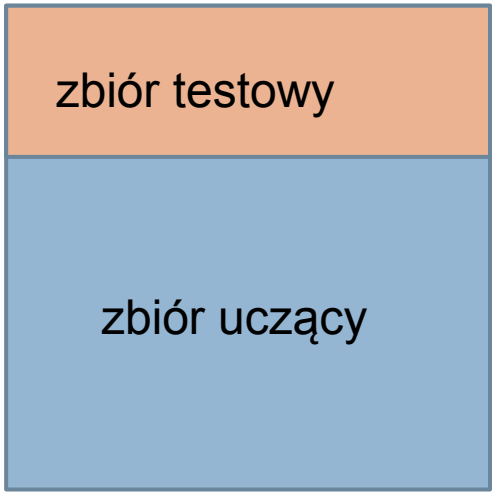
Walidacja krzyżowa

krok 1

krok 2



krok 3



Rozpoznawanie zdarzeń dźwiękowych

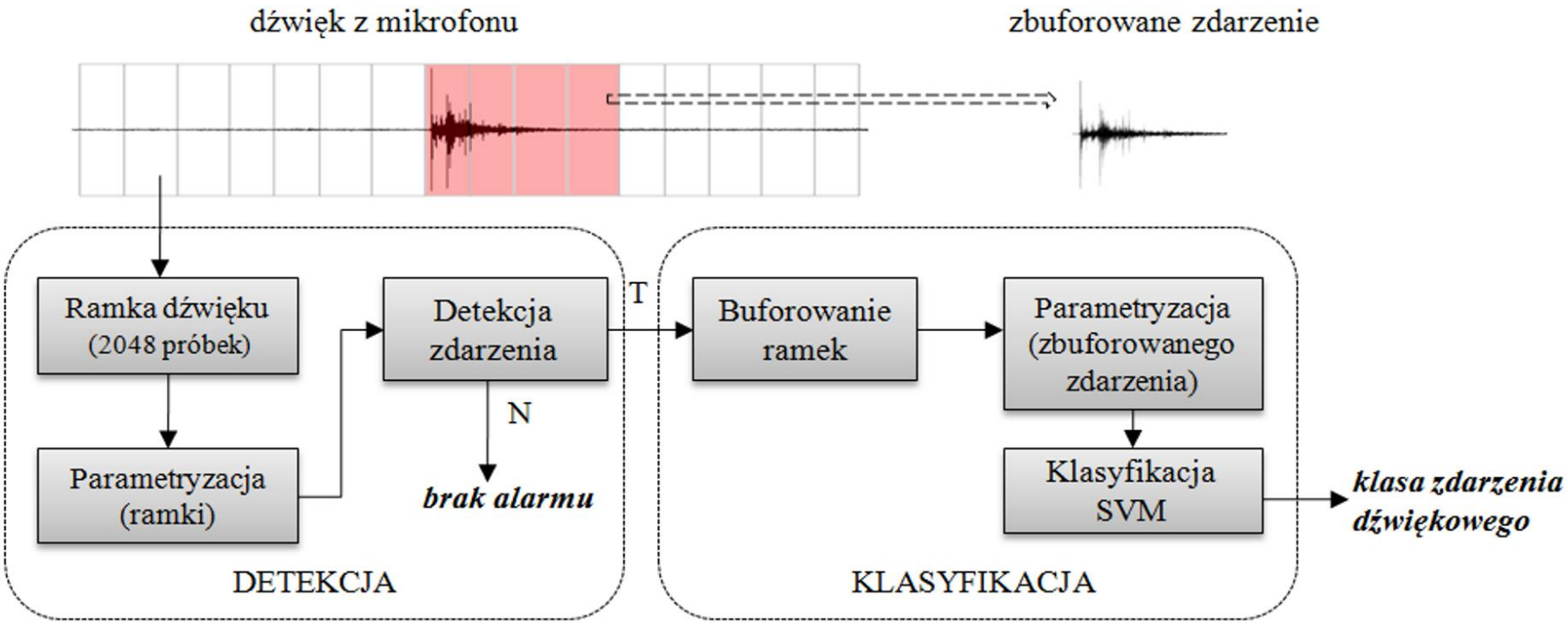
Zdarzenie dźwiękowe – zdarzenie akustyczne – zdarzenie, które występuje w środowisku, trwa skończony czas i można je rozpoznać za pomocą zmysłu słuchu lub analizy próbek dźwięku.

Zastosowanie

- monitoring akustyczny
- pomoc dla niesłyszących
- automatyka

Rozpoznawanie zdarzeń dźwiękowych

Przykładowy algorytm



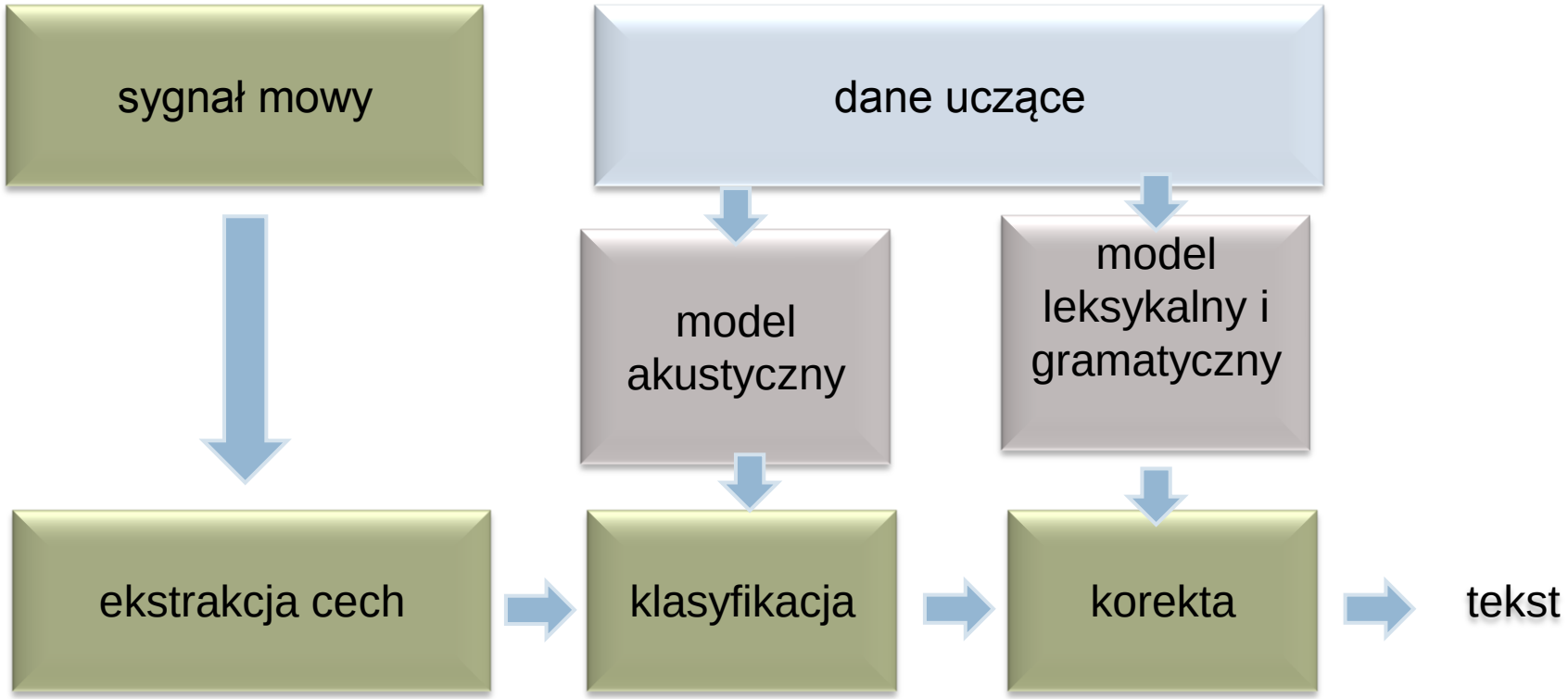
Rozpoznawanie mowy

- Rozpoznawanie mowy (ang. speech to text, Automatic Speech Recognition – ASR) – zamiana mowy na wypowiedź w formie tekstowej.
- Należy rozróżnić pojęcia:

- rozpoznawanie mowy (speech recognition)
- rozpoznawanie mówcy (speaker recognition)
- rozpoznawanie głosu (voice recognition)

Rozpoznawanie mowy

- Wykorzystanie parametrów perceptualnych (odwzorowujących działanie ucha)
- Struktury dynamiczne (HMM)



Rozpoznawanie muzyki

- MIR – Music Information Retrieval
- Automatyczne rozpoznanie gatunku muzycznego
- Rozpoznanie tonacji, tempa utworu
- Automatyczne wyszukiwanie muzyki w Internecie
- Rozpoznawanie melodii – Query by humming, query by example
- Najczęściej wykorzystywane parametry MPEG-7

Rozpoznawanie dźwięków w warunkach rzeczywistych

- W warunkach rzeczywistych obecne są zakłócenia (szum), które wpływają na wartości parametrów i mają wpływ na wynik rozpoznania
- Skuteczność rozpoznawania należy podawać przy zmierzonym stosunku sygnału do szumu (SNR – ang. Signal to Noise Ratio)

$$SNR[dB] = 10\log\left(\frac{\int s^2(t)dt}{\int n^2(t)dt}\right)$$

Dziękuję za uwagę!

Kuba Łopatka
klopatka@sound.eti.pg.gda.pl