Automatyczne rozpoznawanie mowy

Autor: mgr inż. Piotr Bratoszewski

Rys historyczny

- 1930-1950 – pierwsze systemy Automatycznego rozpoznawania mowy (ang. *Automatic Speech Recognition* – ASR), metody holistyczne; „ad-hoc”; izolowane słowa; małe słowniki; Bell Laboratories
- 1950-1960 – pierwsze systemy ASR oparte na zależnościach fonetycznych; małe słowniki;

Rys historyczny

- 1960-1980 – systemy oparte o rozpoznawanie wzorca (ang. *pattern recognition*); wykorzystanie parametrów kodowania predykcyjnego (LPC); sekwencje izolowanych lub połączonych słów; małe i średnie słowniki
- 1980-2000 – wprowadzenie statystycznego modelowania zależności dynamicznych i statycznych w **mowie ciągłej**; zastosowanie ukrytych modeli Markowa (ang. *Hidden Markov Models* - HMM)

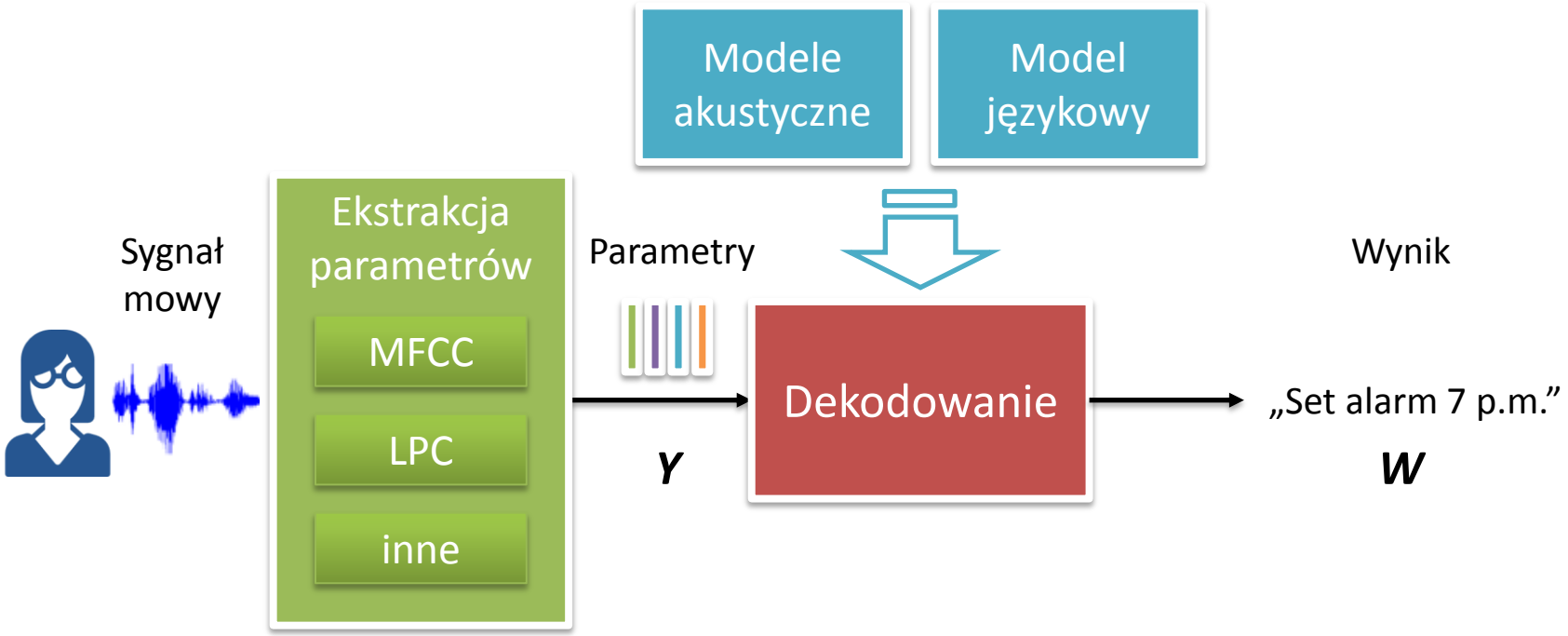
Rys historyczny

- 2000-teraz – kombinacje modeli HMM oraz zależności akustyczno fonetycznych w celu znajdowania i korekcji nieregularności językowych, deep learning, systemy pracujące w chmurze; zwiększanie odporności systemu na pracę w środowisku szumowym; rozpoznawanie wielomodalne

Istotne terminy

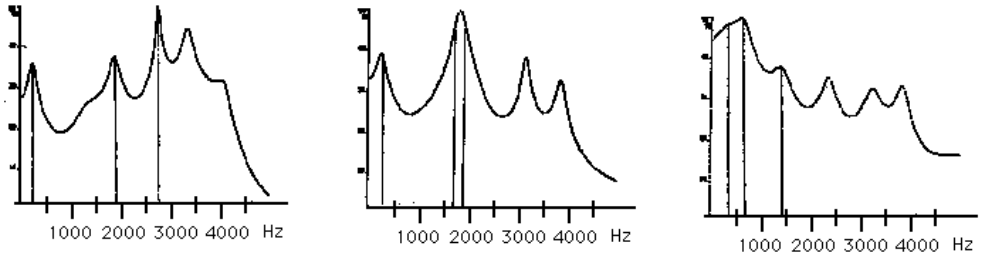
- Słownik – ilość wyrazów które system jest w stanie rozpoznać:
  - Mały słownik: 2 – 100 wyrazów
  - Średni słownik: 100 – 1000 wyrazów
  - Duży słownik: ponad 1000 wyrazów (w tej chwili 50 tys. słów)
- System „zależny/niezależny od mówcy”
- Rozpoznawanie mowy ciągłej/izolowanej
- Składnia (ang. *syntax*) – mowa naturalna/wydawanie poleceń/rozpoznawanie cyfr

Schemat systemu ASR

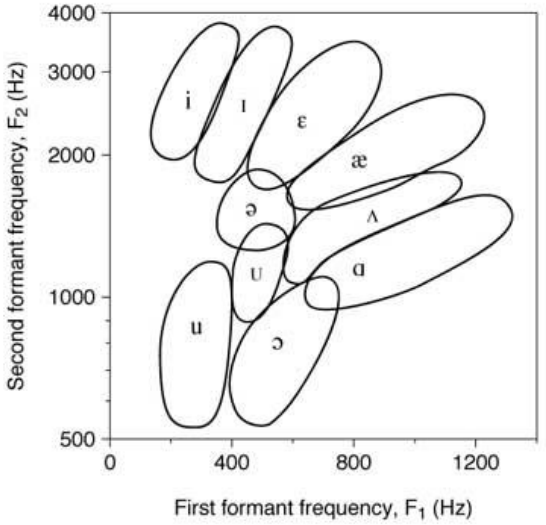


Parametry widmowe

- Podstawowymi parametrami mowy są parametry widmowe uzyskiwane poprzez analizę transformaty Fouriera sygnału mowy
- Analiza rozkładu formantów pozwala na rozpoznawanie samogłosek



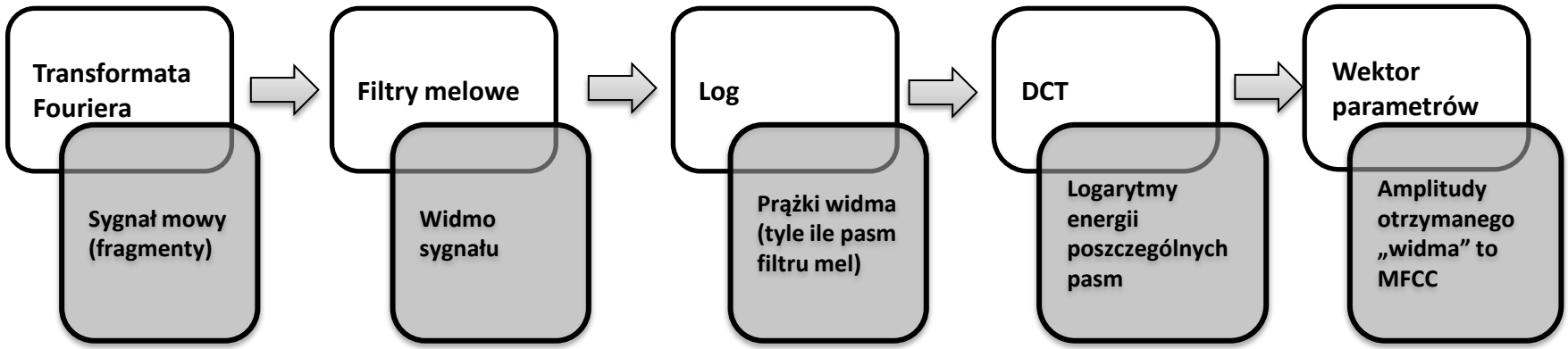
Formanty



Metody parametryzacji mowy

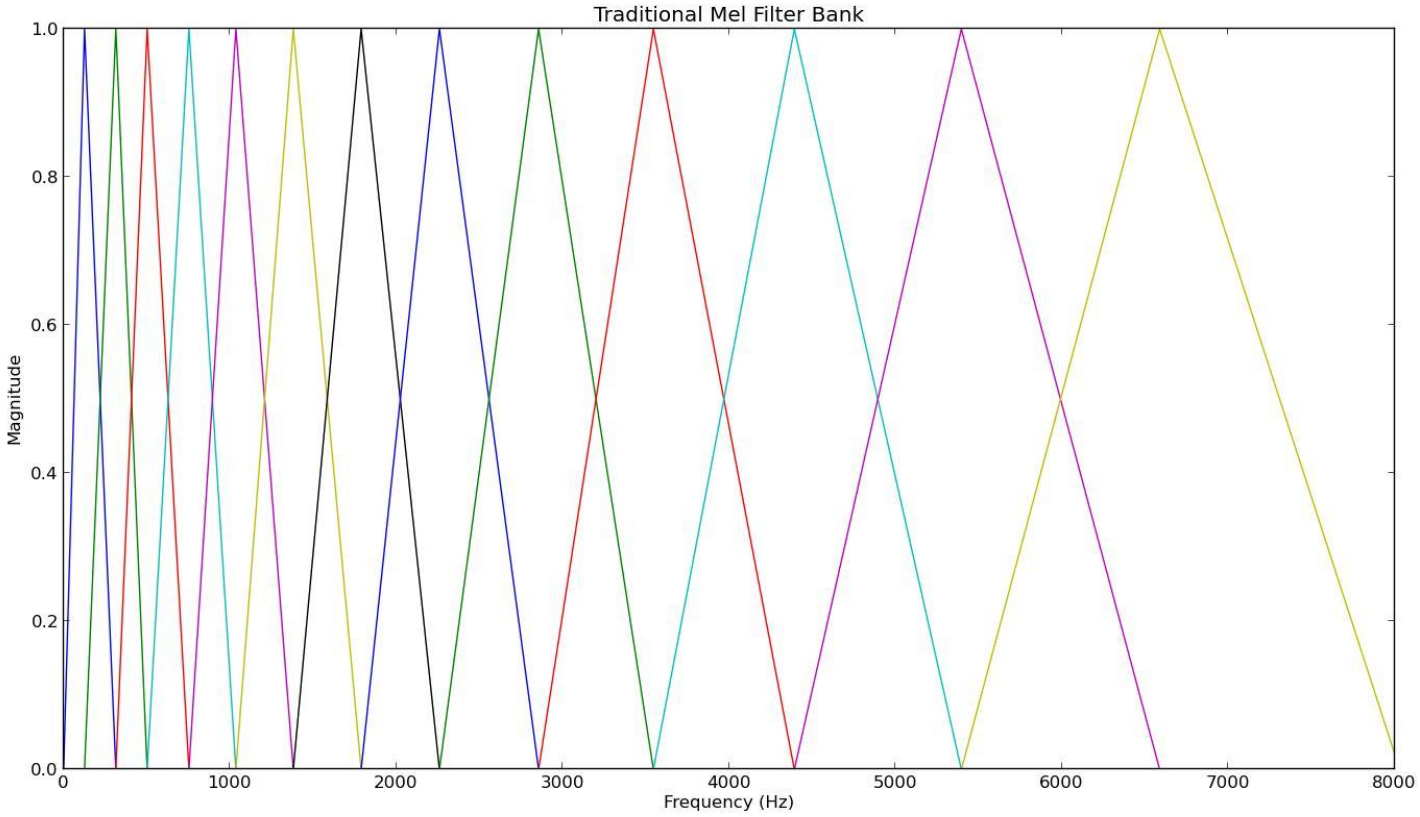
Cepstralne (np. MFCC)

– Metody efektywne i łatwe w implementacji



Metody parametryzacji mowy

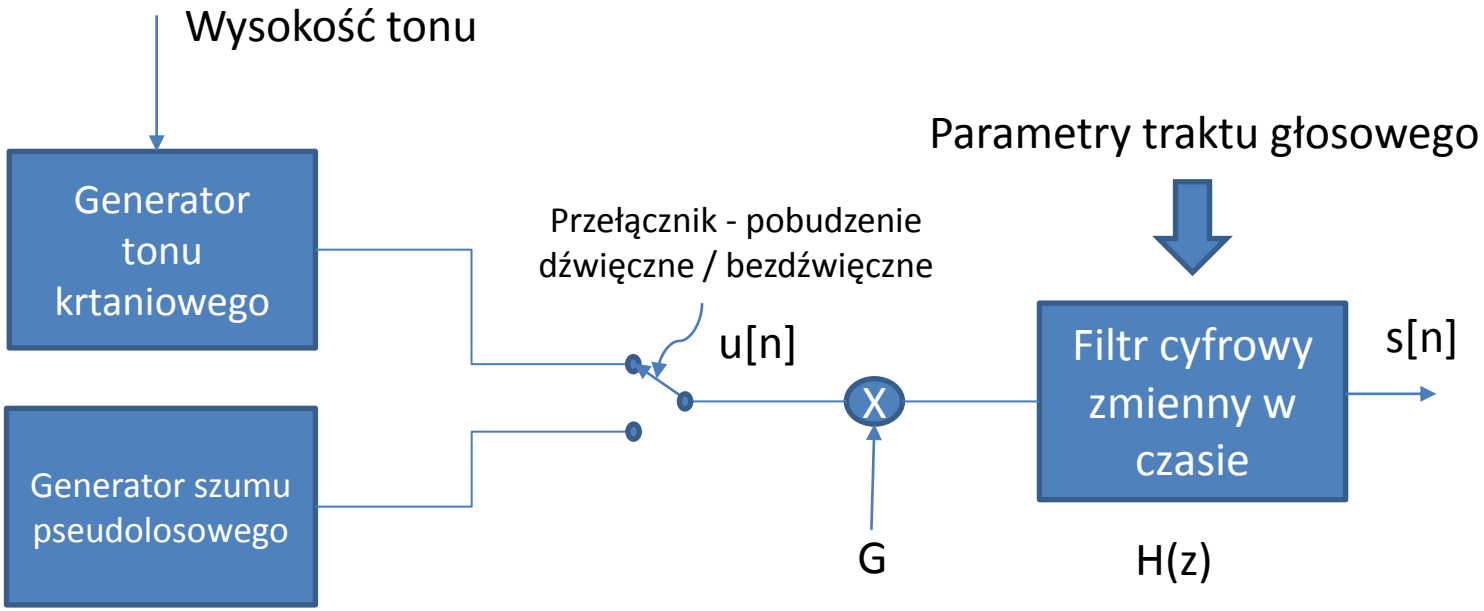
- Filtry w skali melowej



Metody parametryzacji mowy

Predykcyjne (LPC, PLP)

- Skuteczność porównywalna do MFCC w warunkach optymalnych
- Większa skuteczność w przypadku rozbieżności danych testowych z treningowymi



Model akustyczny

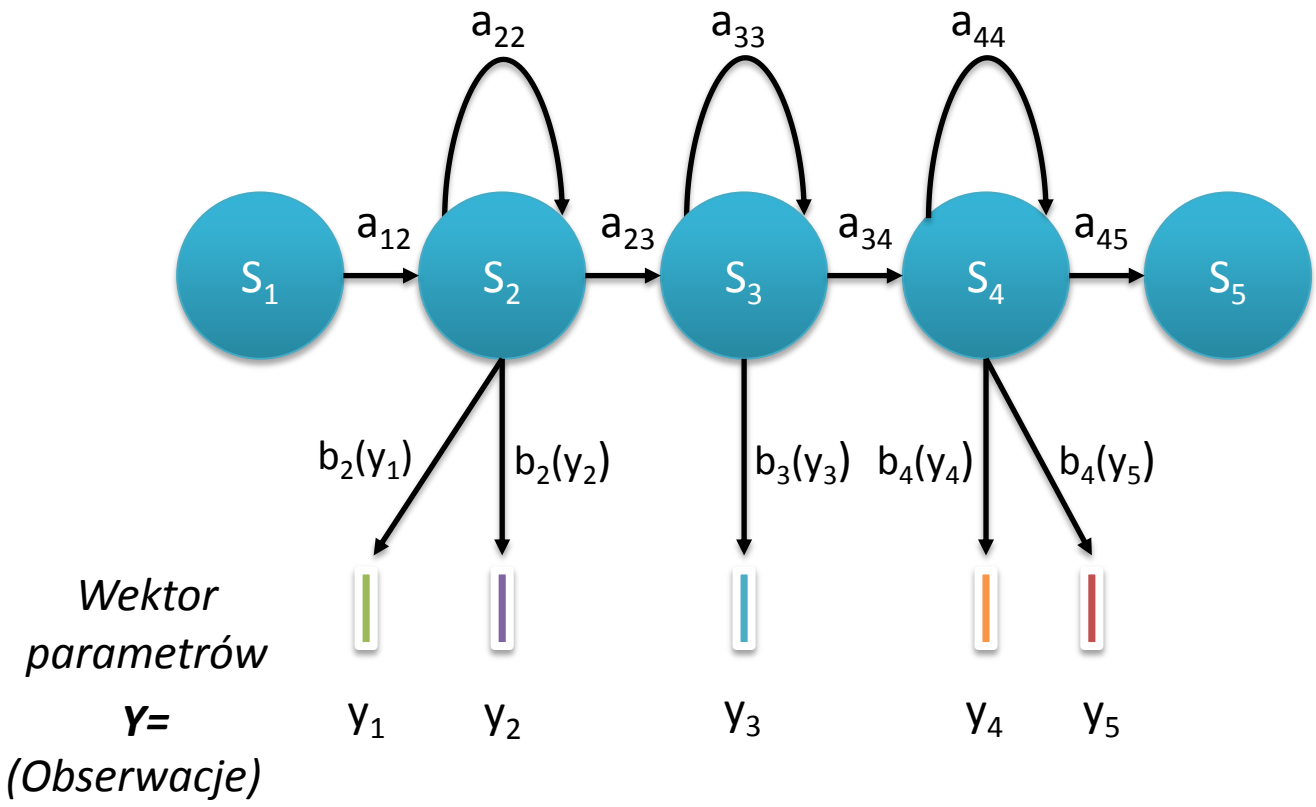
- Wymawiane słowa dzieli się na sekwencję podstawowych jednostek mowy – fonemy
- Model akustyczny reprezentuje zależności pomiędzy sygnałem akustycznym a fonemami

BROWSER	b r aw z ax
BROWSER	b r aw z ax r
CALCULATOR	k ae l k y uh l ey t ax
CALCULATOR	k ae l k y uh l ey t ax r
COMPUTER	k ax m p y uw t ax
COMPUTER	k ax m p y uw t ax r

Model akustyczny

- W celu zamodelowania najkrócej trwających fonemów (głoski wybuchowe – ang. *plosive phonemes: -p; -t; -k*) należy dobrać odpowiednie okno analizy – typowo o długości 10 ms
- Przy tworzeniu modelu akustycznego korzysta się najczęściej z parametrów mel-cepstralnych (MFCC) lub parametrów LPC
- Typowo stosuje się modele trifonowe
- Istostne – osobny model ciszy

Ukryte Modele Markowa



Model *Bakisa,* (ang. *left-right HMM*)

Ukryte Modele Markowa

- Modelowanie procesu na podstawie skończonej liczby stanów S
- Opisywane są przez:
  - N, liczba stanów
  - M, liczba obserwacji
  - Prawdopodobieństwo przejścia pomiędzy stanami $A=\{a_{ij}\}$
  - Prawdopodobieństwo wygenerowania danej obserwacji w stanie j: $B=\{b_j(Y)\}$
  - Rozkład początkowy prawdopodobieństwa $\pi=\{\pi_i\}$

Dekodowanie

- Zadaniem dekodera jest rozpoznanie wymawianego słowa
- Posiadając na wejściu wektory $Y=y_1, \ldots, y_T$ dekoder ma za zadanie rozpoznać sekwencję odpowiadających im wyrazów $W= w_1, \ldots w_K$ zgodnie z:

$$\hat{W} = arg\ \max_{W}[p(W|Y)]$$

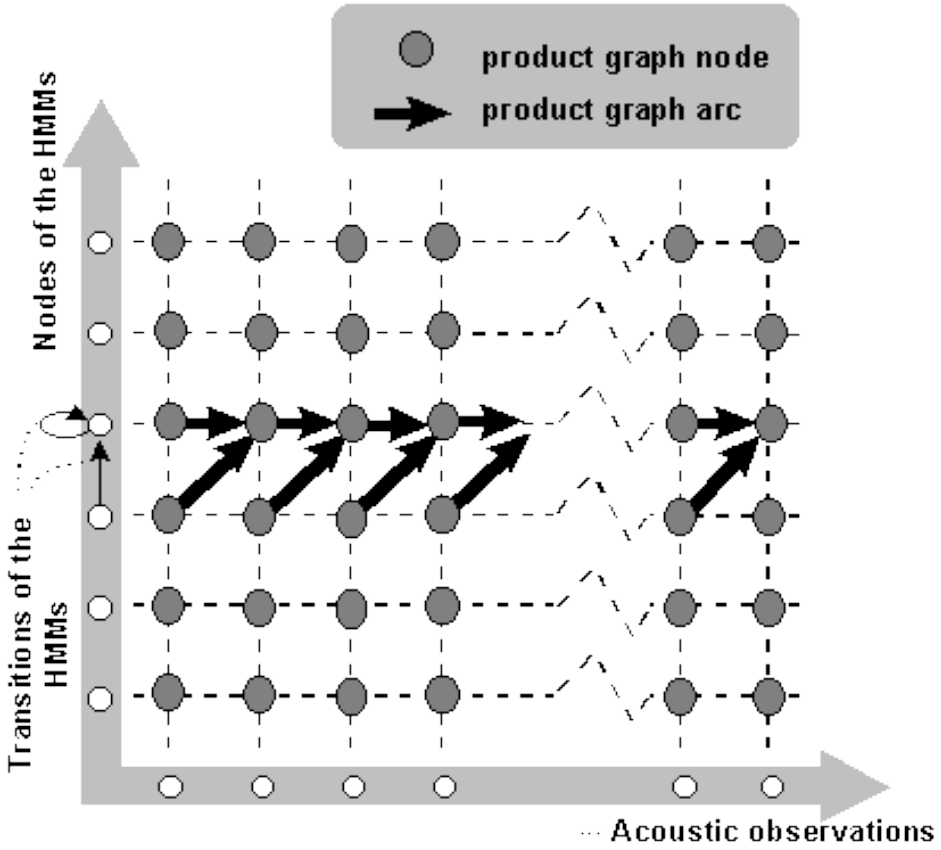
Korzystając z twierdzenia Beyesa:

$$\hat{W} = arg\ \max_{W}[p(Y|W)\cdot p(W)]$$

Model akustyczny

Model językowy

Dekodowanie



W fazie dekodowania (odkrywania ukrytej sekwencji stanów modelu HMM) często stosuje się algorytm *Viterbiego*

Model językowy

$$p(W)=\prod_{k=1}^{K} p(w_k|w_{k-1},\ldots,w_{k-N+1})$$

Model N-gramowy – biorący pod uwagę N wyrazów „wstecz”

Bigram:

P(ala ma małego kota)=P(ala|<start>)·P(ma|ala)
·P(małego|ma)·P(kota|małego)·P(<end>|kota)

Unigram:

P(ala ma małego kota)=P(ala)·P(ma)·P(małego)·P(kota)

Skuteczność rozpoznawania

- Do oceny skuteczności systemu ASR stosowana jest miara „wyrazowej stopy błędu” (ang. *Word Error Rate* – WER)

$$WER = \frac{D+S+I}{H+D+S} \cdot 100\%$$

H – liczba poprawnie rozpoznanych słów
D – liczba nie rozpoznanych słów (ang. *deletions*)
S – liczba błędnie rozpoznanych słów (ang. *substitutions*)
I – liczba wstawionych słów (ang. *insertions*)

Skuteczność rozpoznawania

- Wynik rzeczywistego rozpoznawania

"'*'/SPEAKER03_C1_AUD1_13.rec“
999999 5889999 TUESDAY
7889999 14239999 WEDNESDAY
16239999 22729999 THURSDAY
24729999 31399999 FRIDAY
33399999 41129999 SATURDAY
43129999 49709999 SUNDAY
51709999 57859999 MOUSE
59859999 64749999 MOVE
66749999 71429999 MUSIC
73429999 78809999 MUTE

"'*'/SPEAKER03_C1_AUD1_13.rec"
600000 7600000 TUESDAY
7600000 16000000 WEDNESDAY
16000000 24700000 THURSDAY
24700000 33400000 FRIDAY
33400000 43200000 SATURDAY
43200000 51900000 SUNDAY
51900000 55200000 MY
55200000 60200000 AS
60200000 67200000 MOVE
67200000 74000000 MUSIC
74000000 80500000 MUTE

Dynamic Time Warping

- Algorytm DTW – dynamiczne marszczenie czasu

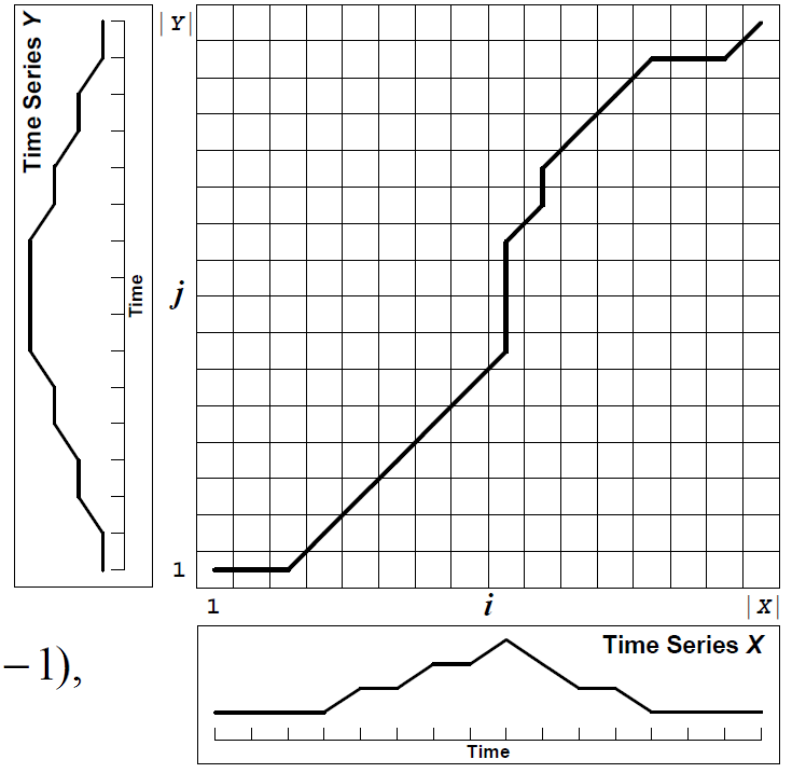
Dwa przebiegi czasowe:

$$X = x_1, x_2, \ldots, x_i, \ldots, x_{|X|}$$

$$Y = y_1, y_2, \ldots, y_j, \ldots, y_{|Y|}$$

Tworzenie macierzy kosztów

Obliczanie ścieżki o najmniejszym koszcie



$$D(i,j) = Dist(i,j) + \min[D(i-1,j), D(i,j-1), D(i-1,j-1)]$$

Poprawa rozpoznawania w szumie

- Stosowanie filtracji widmowej (ang. *Spectral Subtraction*)
- Stosowanie filtracji adaptacyjnej
- Stosowanie filtracji przestrzennej – *Beamforming*
- Stosowanie różnych modeli akustycznych

Bimodalne rozpoznawanie mowy

- Dołączenie do wektora parametrów akustycznego parametrów wizyjnych
- Fonemy = wizemy
- Widowiskowe podejście – czytanie z ruchu warg
- Teoretycznie wzrost skuteczności w warunkach szumowych
- Wiele problemów do rozwiązania (detekcja ust, framerate, cechy osobnicze)

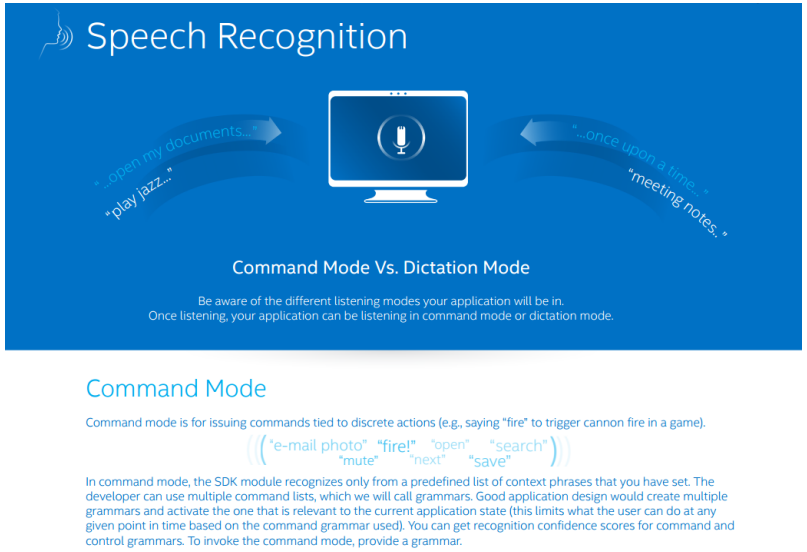
Przykładowe systemy ASR

Apple Siri

Google Now

Przykładowe systemy ASR

Intel RealSense



NUANCE

Dragon

Dragon Anywhere

Complete work on the go with professional-grade mobile dictation

Dragon Professional Individual

Put your voice to work for greater efficiency on the job

Dragon for Mac

Get the all-in-one Mac solution that let's you dictate, edit and transcribe by voice

Nuance® Dragon® Professional Group

Dragon Professional Group

Improve documentation productivity across the enterprise

Przykładowe otwarte systemy ASR

CMU Sphinx

Project by Carnegie Mellon University

HTK Toolkit

htk3

Bibliografia

- HTK Book: speech.ee.ntu.edu.tw/homework/DSP_HW2-1/htkbook.pdf
- Rabiner L., A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition
- Benesty, Springer Handbook of Speech Processing