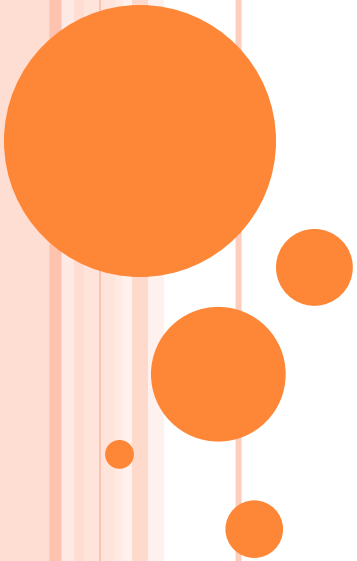


Akustyka mowy

Podstawy rozpoznawania mowy – część I

Plan wykładu

Część I

- Podstawowe pojęcia z dziedziny rozpoznawania mowy
- Algorytmy, parametry i podejścia do rozpoznawania mowy
- Przykłady istniejących bibliotek i systemów

Podstawowe pojęcia

Podstawowe pojęcia

- ASR – Automatic Speech Recognition – inaczej Speech-To-Text – zamiana informacji w formie mowy ludzkiej na tekst
- Speaker Recognition – rozpoznawanie mówcy (niezależne od treści lub na hasło)
- Voice Recognition – rozpoznawanie komend głosowych konkretnego mówcy (trening)
- Speech Recognition – rozpoznawanie treści wypowiedzi niezależnie od mówcy
- LVCSR – Large Vocabulary Continuous Speech Recognition – rozpoznawanie mowy ciągłej (np. dyktowanie tekstu)

MIARY DOKŁADNOŚCI

- WER – Word Error Rate
- SER – Sentence Error Rate

$$WER = (S + D + I)/N$$

S – liczba pomylonych słów

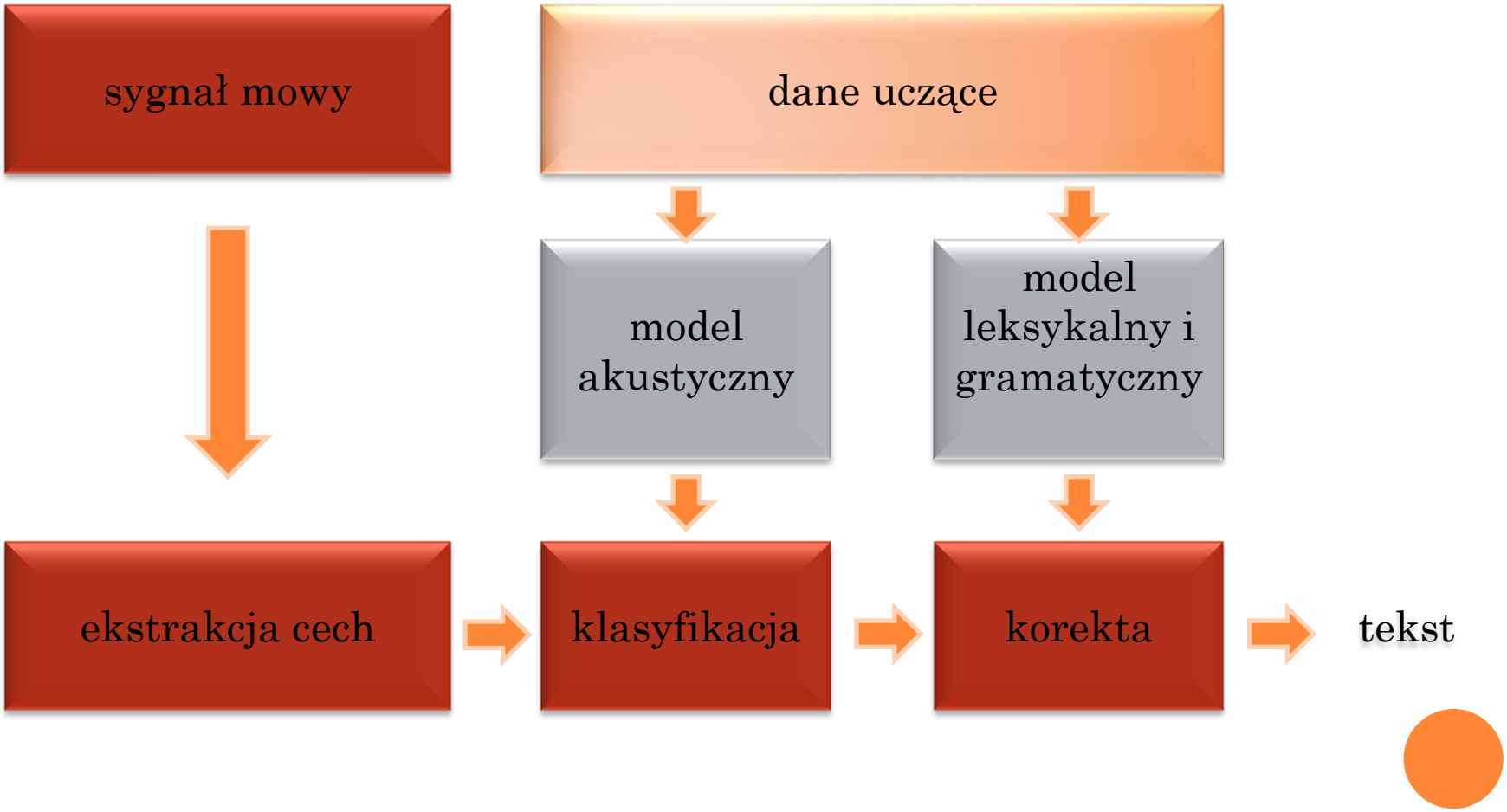
D – liczba pominiętych słów

I – liczba wstawionych słów

N – liczba wszystkich słów

Parametry i Algorytmy

Ogólny schemat systemu ASR

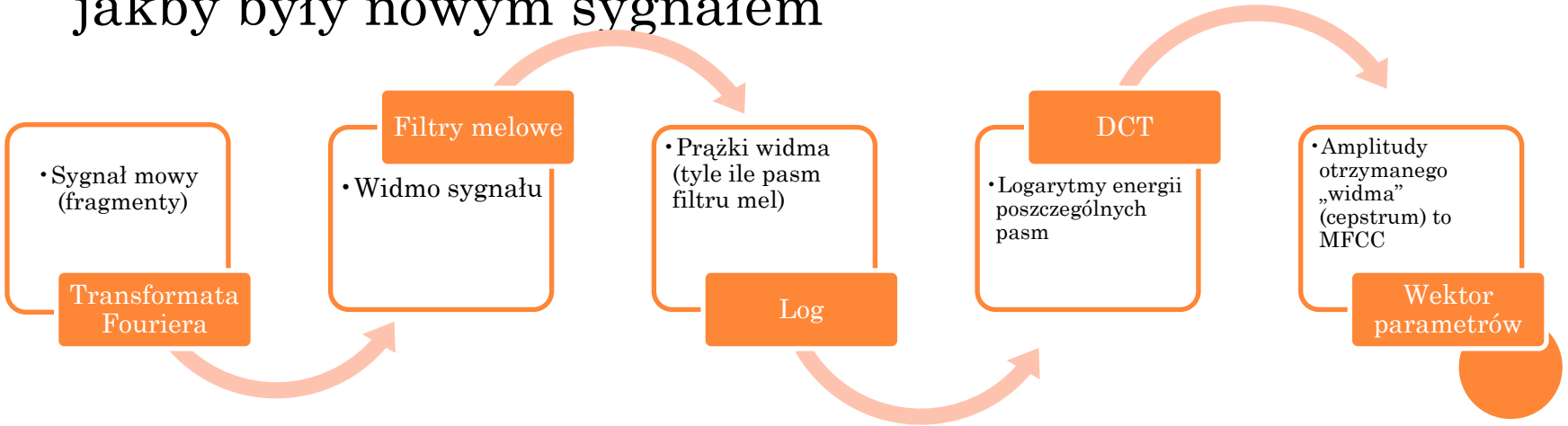


Parametry mel-cepstralne

Parametry mel-cepstralne (ang. MFCC – Mel-Frequency Cepstral Coefficients) to parametry szeroko stosowane w akustyce mowy oraz w kompresji sygnałów fonicznych. Powstają z cepstrum sygnału przedstawionego w skali melowej (mel-cepstrum).

Parametry mel-cepstralne

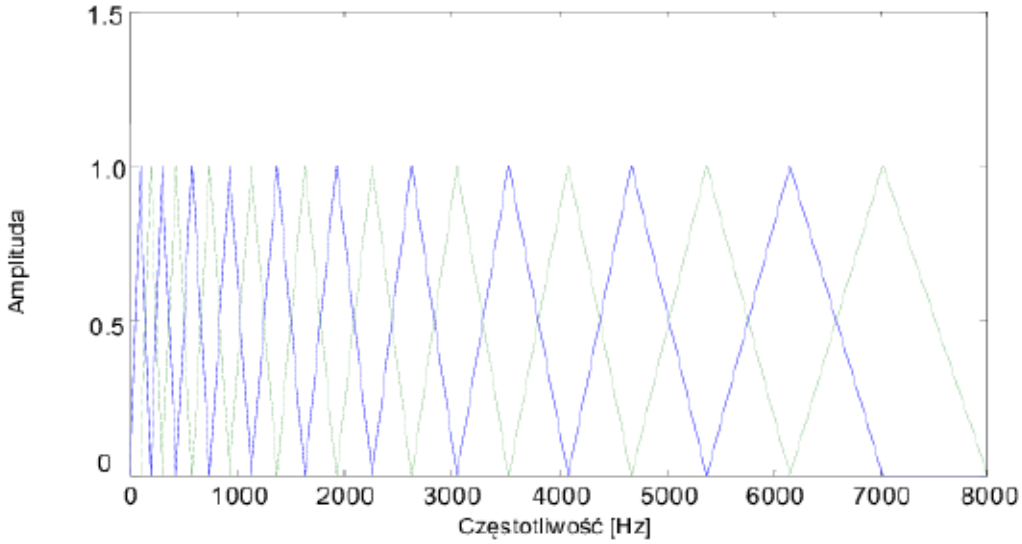
- Sygnał dzielony na okna wg zadanego algorytmu (fonemy, głoski, wg energii itd.)
- Jeden filtr melowy – jeden prążek – jeden współczynnik – jeden parametr mel-cepstralny
- Transformata kosinusowa logarytmów współczynników uzyskanych z filtracji sygnału, tak jakby były nowym sygnałem



Parametry mel-cepstralne

Skalę melową uzyskuje się poprzez filtrację sygnału bankiem filtrów o charakterystyce trójkątnej.

wg Beranka: $F_{mel}(f_{kHz})=1127 \ln(1+f_{kHz}/0.7)$



K- ty współczynnik mel-cepstralny odpowiada zawartości k-tego pasma. Zazwyczaj liczba pasm wynosi od 12 do 20.

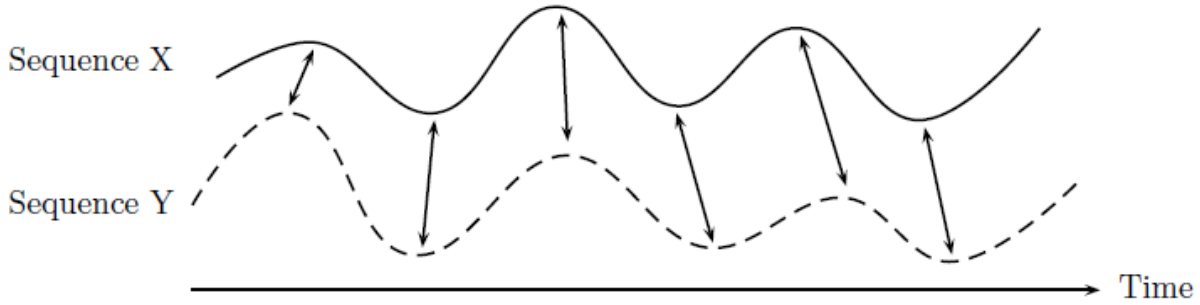
Parametry mel-cepstralne

- Wektor parametrów mel-cepstralnych to wektor współczynników cepstrum w odpowiednich pasmach melowych
- Mają za zadanie odzwierciedlać naturalną odpowiedź układu słuchowego na pobudzenie dźwiękami mowy
- Parametry mel-cepstralne cechuje mała wrażliwość na szum
- Są często wykorzystywane przy rozpoznawaniu mowy

$$\begin{bmatrix} MFCC_1 \\ MFCC_2 \\ MFCC_3 \\ \dots \\ MFCC_k \\ \dots \\ MFCC_K \end{bmatrix}$$

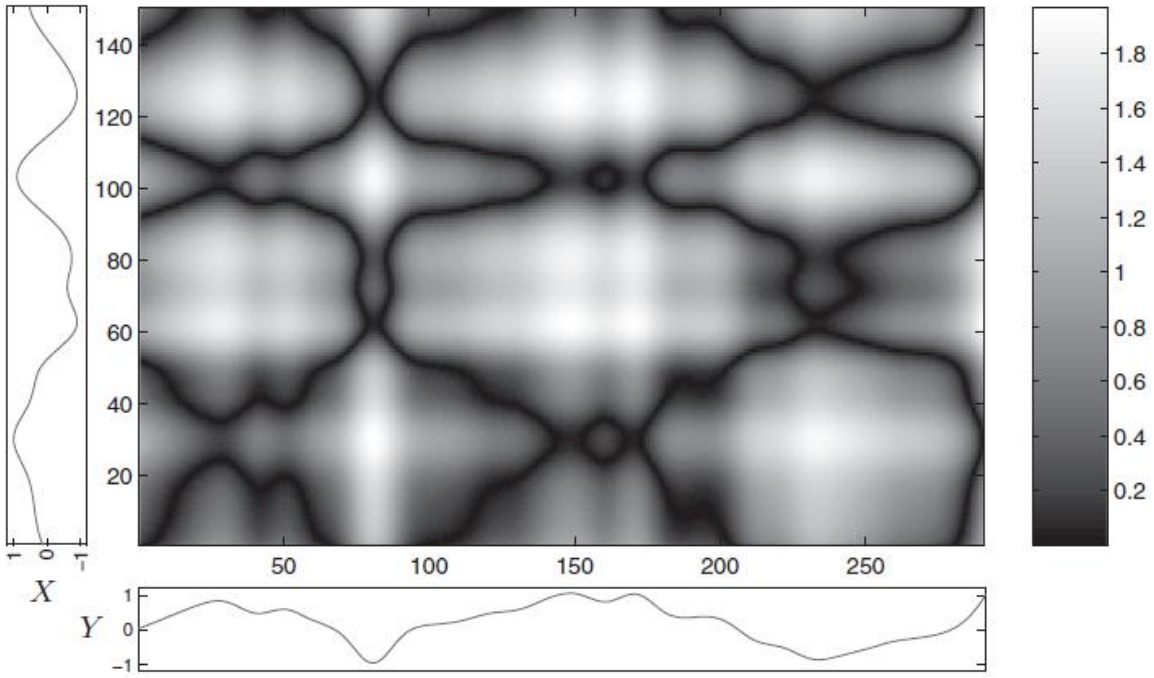
Dynamic Time Warping

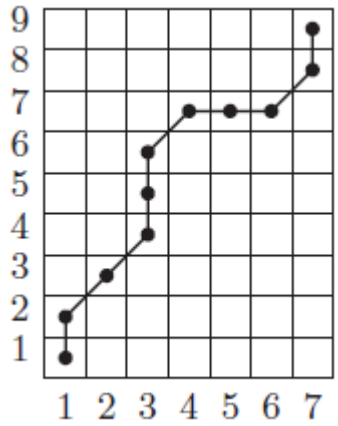
- DTW (zwana też liniową transformacją czasową) to metoda transformacji osi czasu służąca lepszemu dopasowaniu czasowemu dwóch sekwencji czasowych
- Stosowane w celu lepszego dopasowania wypowiedzi do wzorca



Dynamic time warping

- podobieństwo dwóch sekwencji i ścieżka dopasowania



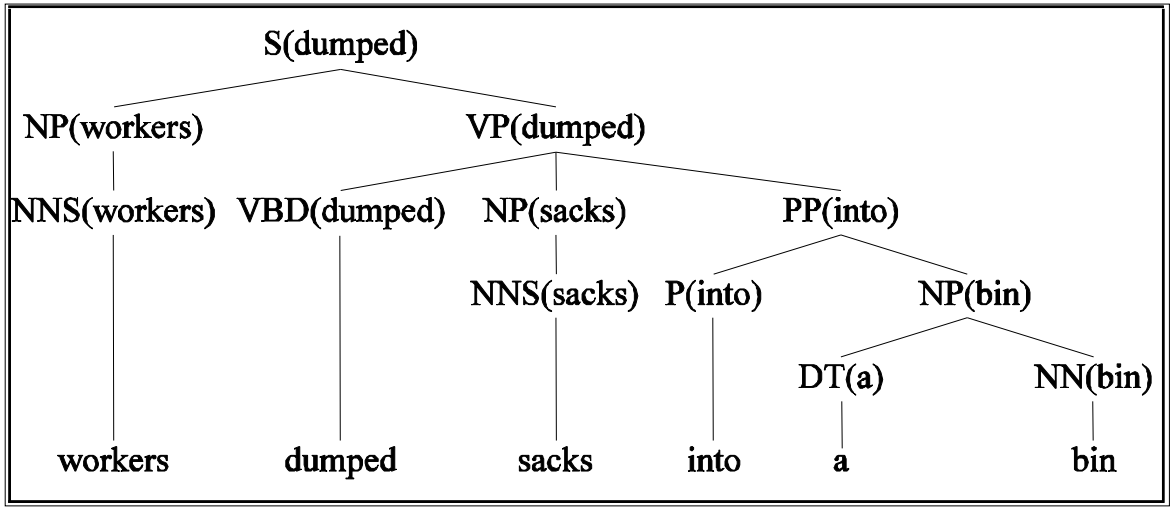


Model gramatyczny

- Finite State Grammar – gramatyka skończonych stanów

```
(Hello | Hi) (John | Sally | Sam)? it's (John | Sally | Sam)
```

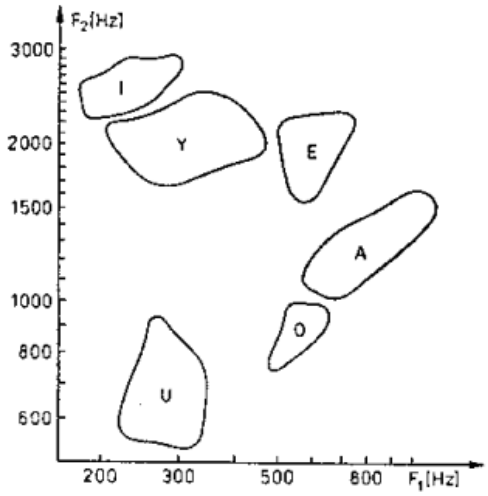
- Gramatyka bezkontekstowa (Context-Free Grammar)



Inne podejścia

- Segmentacja na fonemy + statystyczna klasyfikacja
- Rozpoznawanie samogłosek na podstawie formantów, momentów widmowych

Fonem	częstotliwości [Hz]			poziomy względne [dB]			
i	210	2750	3500	4200	0	-15	-15	-27
e	380	2640	3000	3600	0	-12	-16	-20
a	780	1150	2700	3500	0	-7	-25	-25
y	240	1550	2400	3300	0	-12	-20	-30
o	400	730	2300	3200	0	-3	-30	-35
u	270	615	2200	3150	0	-13	-40	-50
w	600	1700	2900	4100	-9	0	-2	-10
sz	-	2300	2900	3600	-	-9	-8	0
h	500	1700	2500	4200	-12	0	-10	-17
z	-	1750	2950	4300	-	-6	-10	0

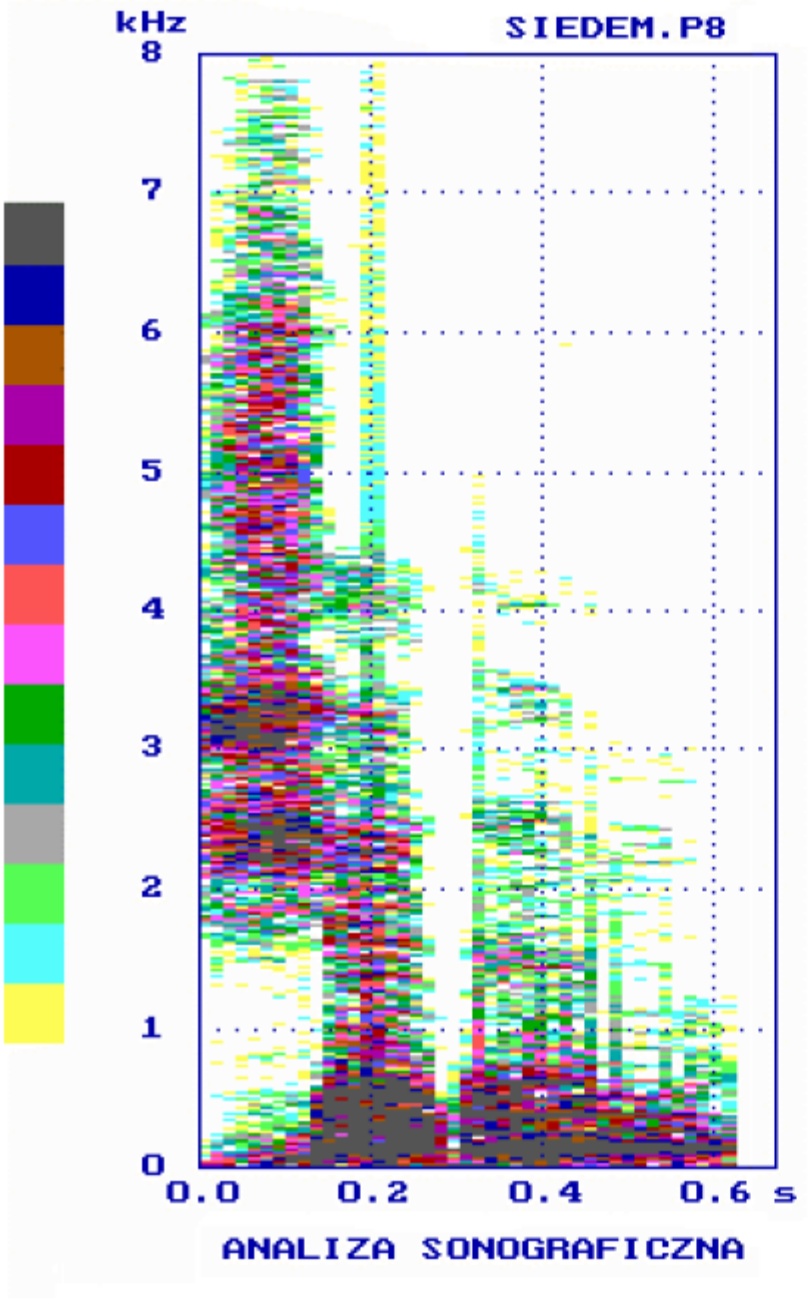


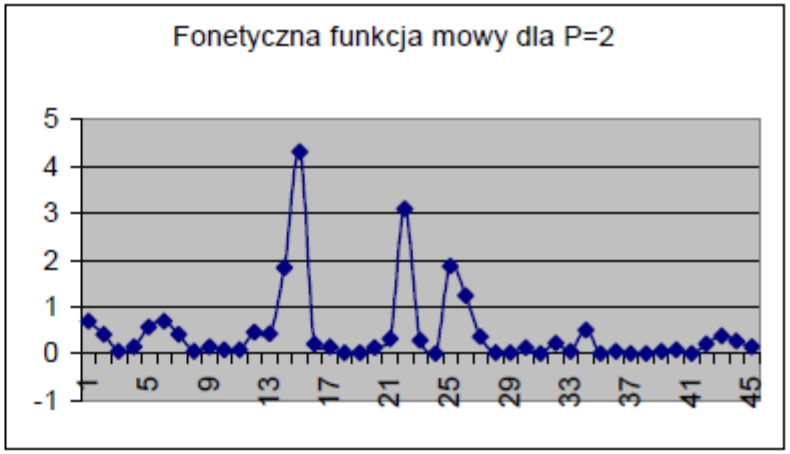
Segmentacja na fonemy

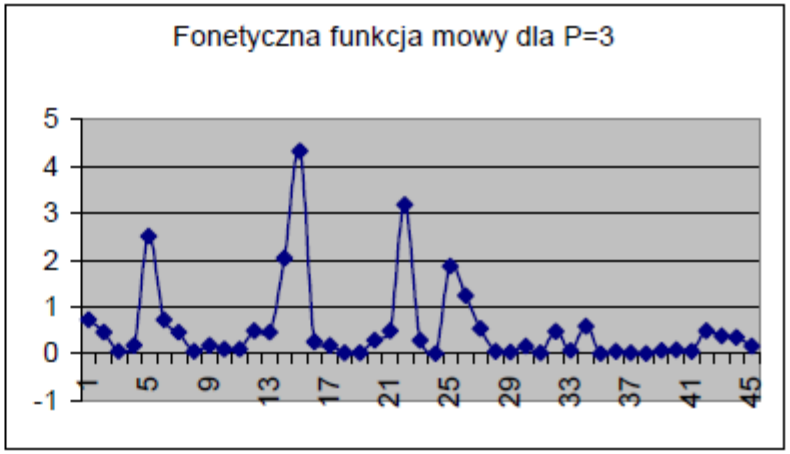
- fonetyczna funkcja mowy

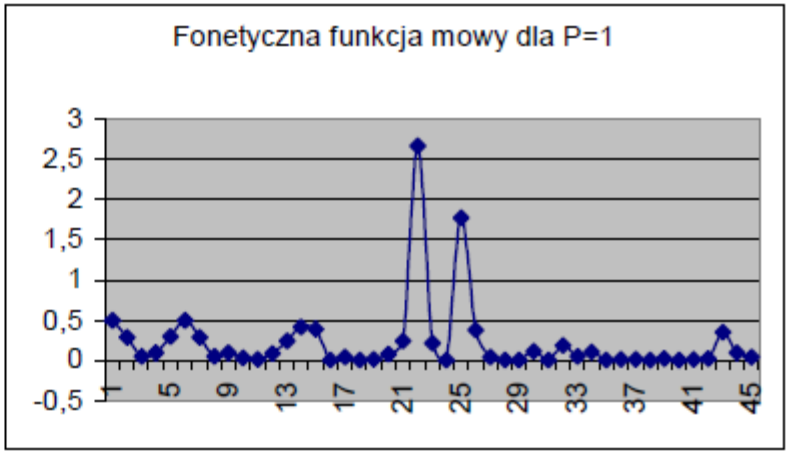
$$P(t) = \frac{1}{P} \cdot \sum_{p=1}^{P} \alpha_p \left[\ln \frac{R(t+\tau, p)}{R(t, p)} \right]^2$$

gdzie: $R(t,p)$ – wektor parametrów w oknie czasowym $(t,\ t+\Delta t)$,
Δt – długość okna czasowego,
a_p – waga p-tego parametru,
P – liczba parametrów,
τ – przesunięcie czasowe, krok analizy .


kHz
SIEDEM.P8
8
7
6
5
4
3
2
1
0
0.0
0.2
0.4
0.6
s
ANALIZA SONOGRAFICZNA

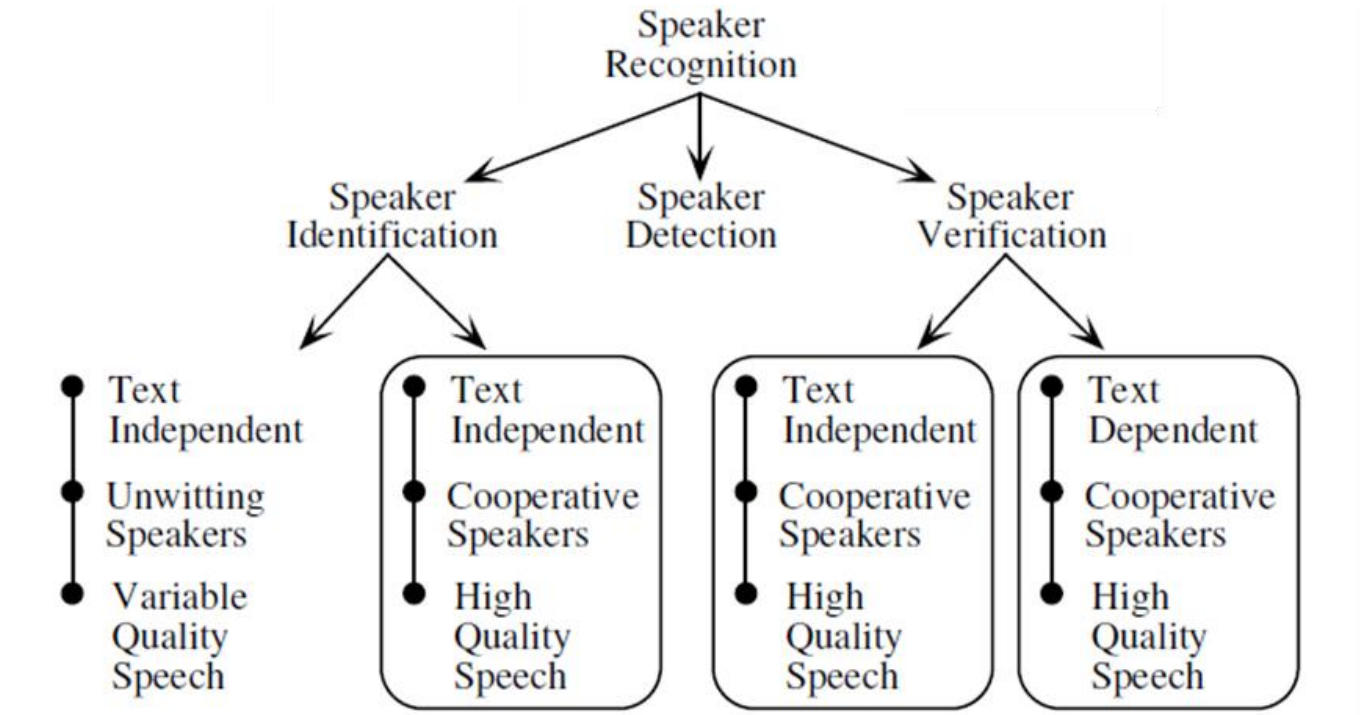

Fonetyczna funkcja mowy dla P=2
5
4
3
2
1
0
-1
1
5
9
13
17
21
25
29
33
37
41
45


Fonetyczna funkcja mowy dla P=3
5
4
3
2
1
0
-1
1
5
9
13
17
21
25
29
33
37
41
45


Fonetyczna funkcja mowy dla P=1
3
2,5
2
1,5
1
0,5
0
-0,5
1
5
9
13
17
21
25
29
33
37
41
45

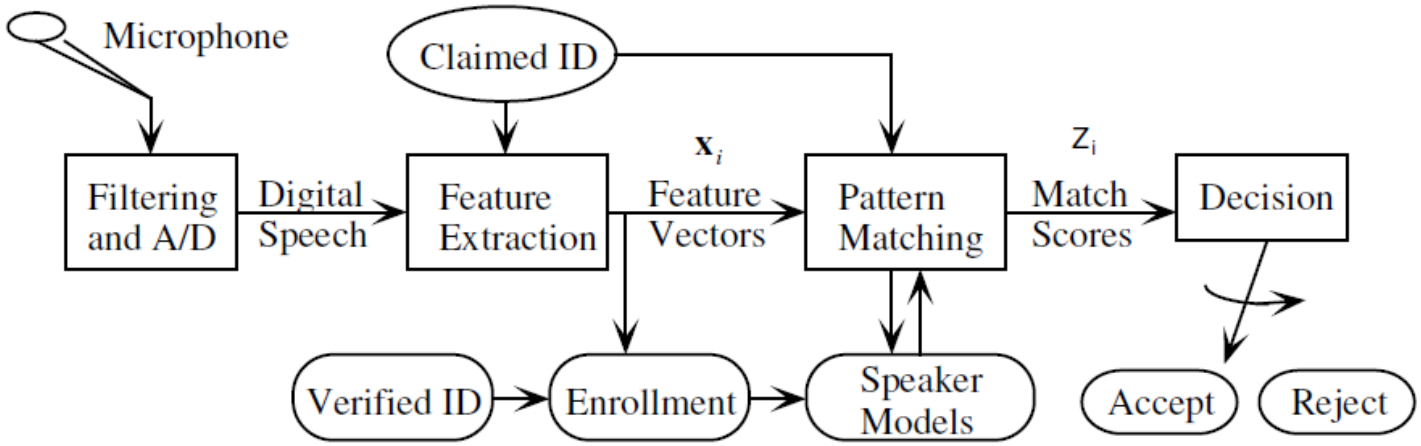
Rozpoznawanie mówcy

- Identyfikacja osoby
- Potwierdzenie tożsamości (speaker verification)
- zależne / niezależne od treści
- mówca świadomy / nieświadomy



Weryfikacja mówcy

- Mówca przedstawia swoją tożsamość
- Cechy sygnału porównywane są z wzorcem w bazie
- Wzorzec jest rejestrowany na podstawie potwierdzonej tożsamości



Rozpoznawanie mówcy – cechy sygnału

Cechy sygnału wykorzystywane w procesie rozpoznawania mówcy

- Mel-Frequency Cepstral Coefficients (MFCC)
- Cepstral features
- Linear Predictive Coding features (LPC)
- Filterbank features
- Autocorrelation features

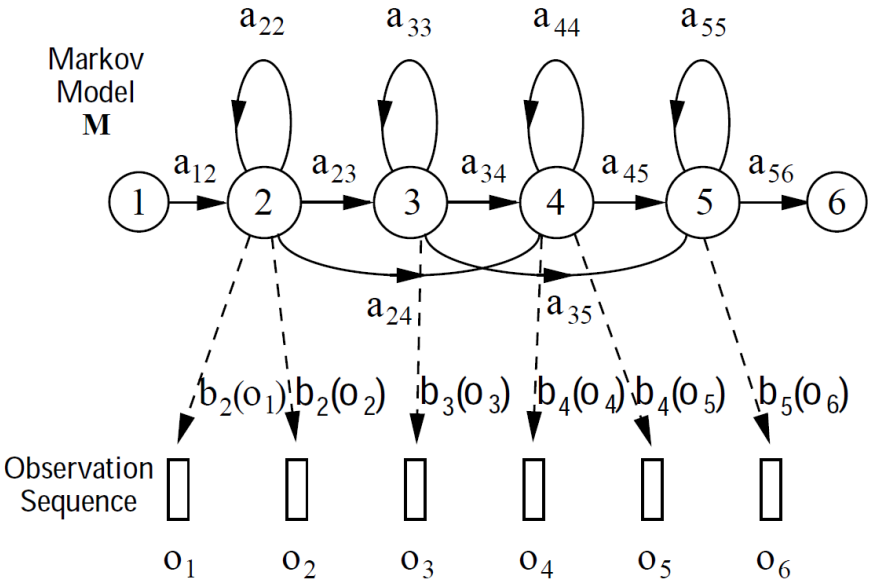
Rozpoznawanie mówcy - klasyfikatory

Metody dopasowania wzorców (pattern matching)

- Dynamic Time Warping (DTW)
- Vector Quantization (VQ)
- Hidden Markov Models (HMM)
- Artificial Neural Networks (ANN)
- (k-)Nearest Neighbour ((k-)NN)
- ...

Ukryte modele Markowa

- HMM (Hidden Markov Model) modeluje proces na podstawie skończonej liczby stanów
- obserwacja **o** (wektor parametrów sygnału mowy) może należeć do jednego ze stanów
- określony jest rozkład prawdopodobieństwa wartości parametrów w każdym ze stanów oraz prawdopodobieństwa przejść pomiędzy stanami



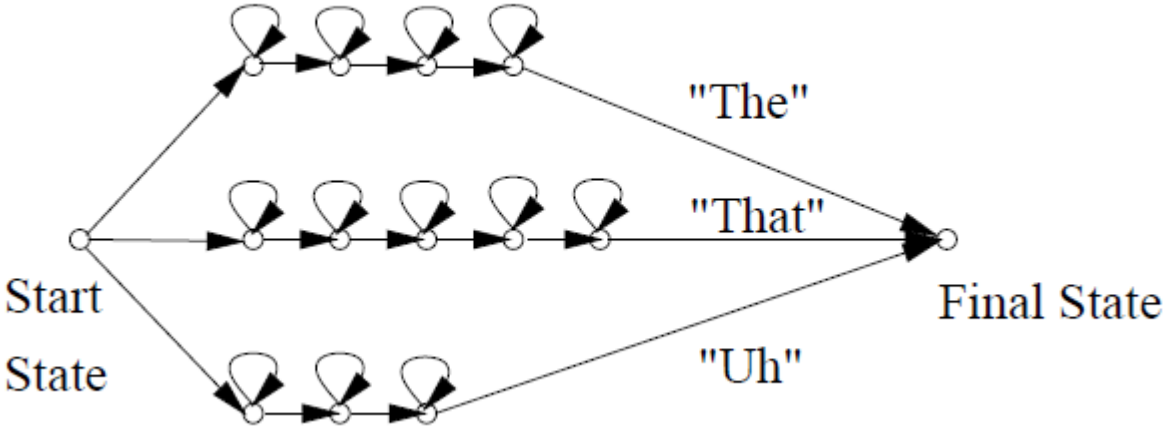
Ukryte modele Markowa

- Prawdopodobieństwo tego, że dana sekwencja obserwacji pasuje do danego modelu określone jest wzorem:

$$P(\mathbf{a}|\mathbf{w}) = \sum_{\mathbf{s}} P(\mathbf{a}|\mathbf{s})P(\mathbf{s}|\mathbf{w}) = \sum_{\mathbf{s}} \prod_{t=1\ldots n} b_{s_t}(o_t)a_{s_t s_{t-1}}$$

Ukryte modele Markowa

- Do celów rozpoznawania mowy konieczne jest wytrenowanie osobnego modelu dla każdego słowa i głosowanie na podstawie prawdopodobieństw *a posteriori*



Istniejące oprogramowanie

Dragon Naturally Speaking

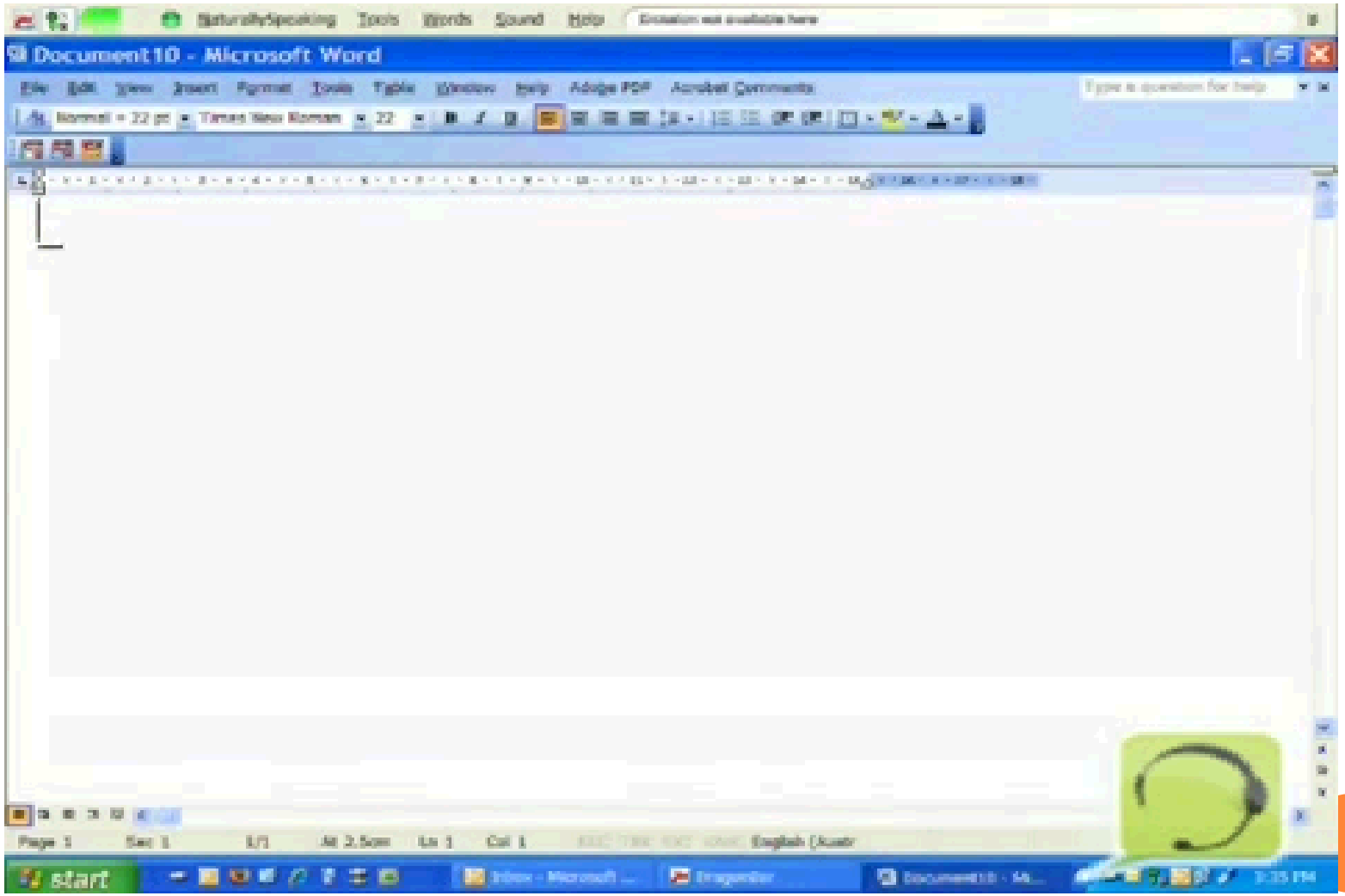
http://www.nuance.com/naturallyspeaking/products/sdk/sdk_client.asp

Wiodące oprogramowanie i SDK służące do rozpoznawania mowy ciągłej (LVCSR):

- możliwość transkrypcji mowy ciągłej w komputerze do 160 słów na minutę
- rozpoznawanie niezależne od barwy głosu mówcy, akcentu, tempa wypowiedzi
- języki: angielski, francuski, włoski, niemiecki, hiszpański, holenderski
- skuteczność do 99%
- możliwość tworzenia profili użytkowników i personalizacji rozpoznawania
- możliwość dodawania nowego słownictwa do bazy
- cena – ok. 25 000$
- dostępne SDK umożliwiające wykorzystanie silnika w aplikacji Windowsowej

NUANCE

Dragon NaturallySpeaking 10 SDK



Loquendo ASR

- oprogramowanie firmy Loquendo, przejętej przez Nuance do rozpoznawania mowy ciągłej (LVCSR)
- połączenie sieci neuronowych i modeli Markowa
- rozpoznawanie niezależne od mówcy; możliwa adaptacja do charakterystyki mowy danego użytkownika
- narzędzia treningu fonetycznego i adaptacji modelu akustycznego
- słownik ponad 10 000 wyrazów
- algorytm odszumiania zwiększający odporność na zakłócenia
- wsparcie rozpoznawania statystycznymi modelami językowymi
- Wsparcie dla ponad 20 języków, w tym polskiego
- Dostępne API w C/C++/C#/.NET/Java

Loquendo
We Speak. We Listen. We Understand.

SPHINX TOOLKIT

Wiodąca biblioteka oferująca funkcje przydatne przy budowie aplikacji i systemów rozpoznawania mowy. Dostępne następujące pakiety:

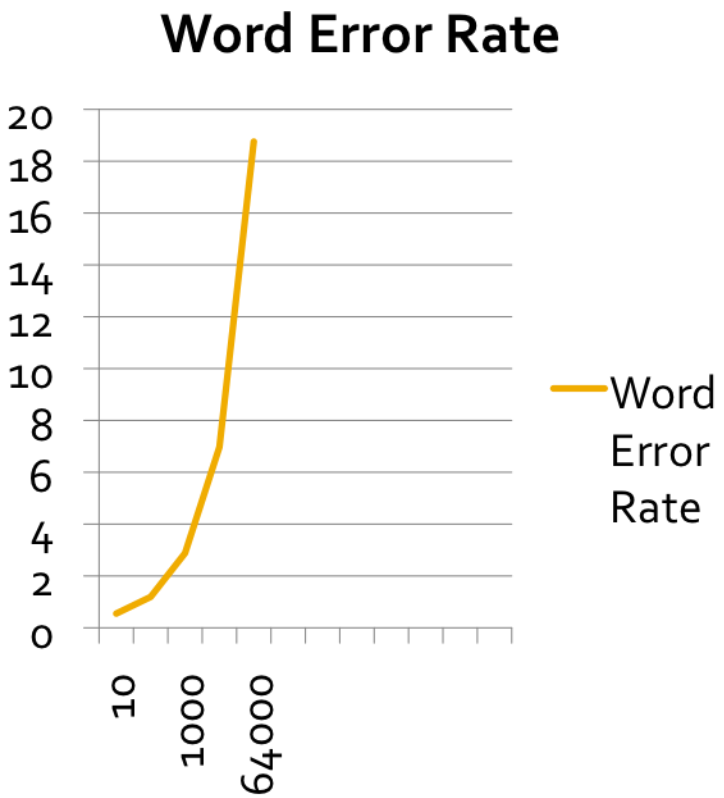
- Pocketsphinx — „lekka” wersja biblioteki na urządzenia mobilne - C
- Sphinxbase — biblioteka wspierająca i wymagana przez PocketSphinx
- Sphinx4 — silnik rozpoznawania z możliwością adaptacji i rozszerzania słownictwa – Jave
- CMUclmtk — narzędzia do tworzenia i zarządzania modelami językowymi - **Java Speech Grammar Format (JSGF)**
- Sphinxtrain —narzędzie do trenowania modeli akustycznych

Biblioteka typu open source

SPHINX TOOLKIT

Vocabulary	Sphinx4 WER
Digits 0-9	.549%
100 Word	1.192%
1,000 Word	2.88%
5,000 Word	6.97%
64,000 Word	18.756%

**If you have noisy audio input multiply expected error rate x 2*

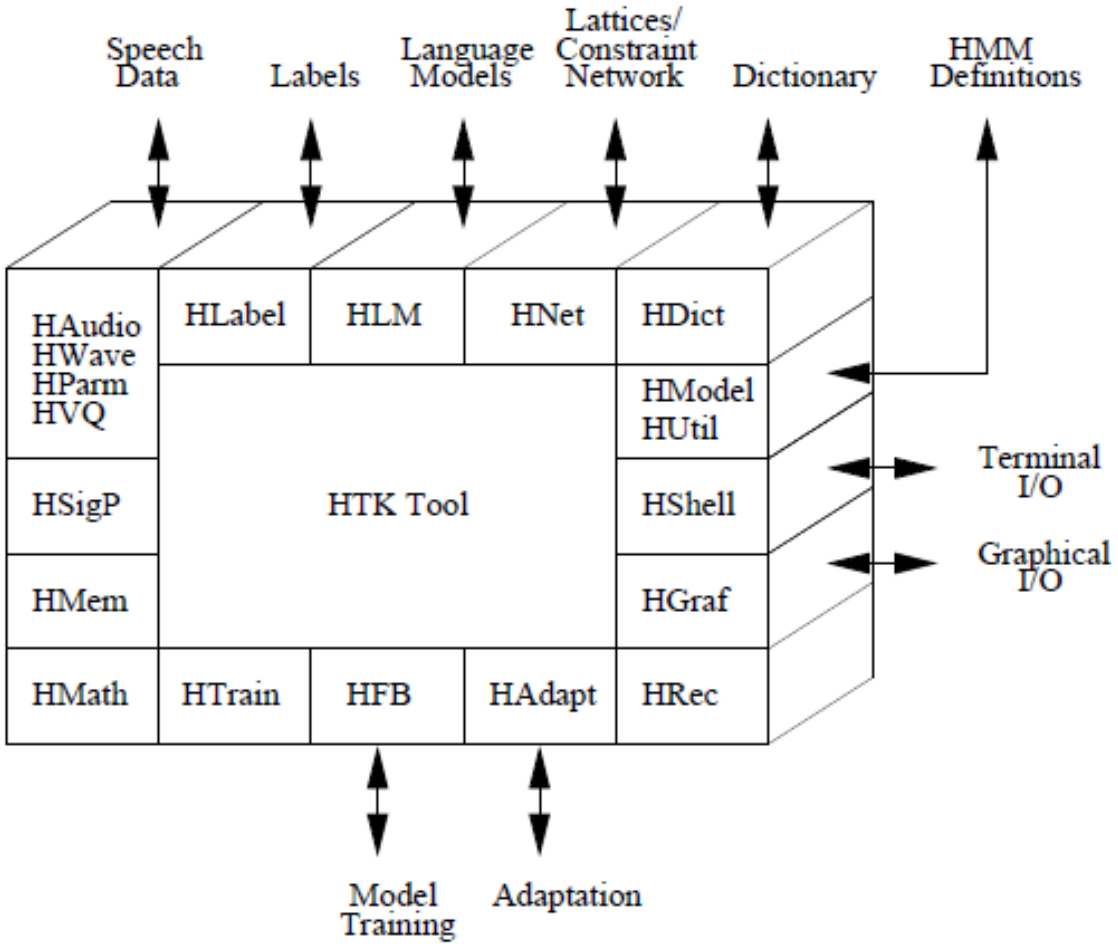


HTK Speech Recognition Toolkit

HTK – Hidden Markov Toolkit – narzędzie popularne wśród badaczy pracujących nad rozpoznawaniem mowy

- interfejs command-line'owy
- dostępne źródła w języku C
- dostępne API w C++ (ATK Real-time API)
- Projekt nierozwijany od 2007 roku
- narzędzia do przygotowania danych, treningu, adaptacji, testowania

HTK Speech Recognition Toolkit
Architektura



Dziękuję za uwagę!