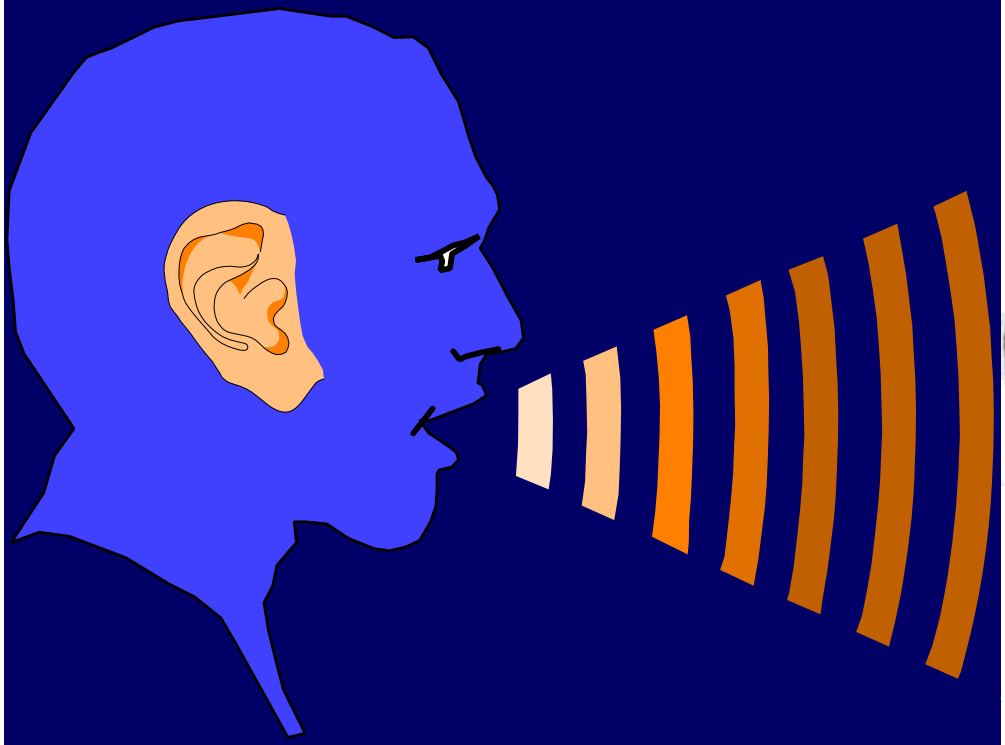
Komputerowe przetwarzanie mowy

Plan wykładu

1. Wprowadzenie – zagadnienia podstawowe
2. Podział systemów rozpoznawania mowy
3. Charakterystyka metod rozpoznawania mowy
4. Model fizyczny traktu głosowego
5. Ekstrakcja parametrów sygnału mowy
6. Przykładowe algorytmy rozpoznawania mowy
7. Podsumowanie

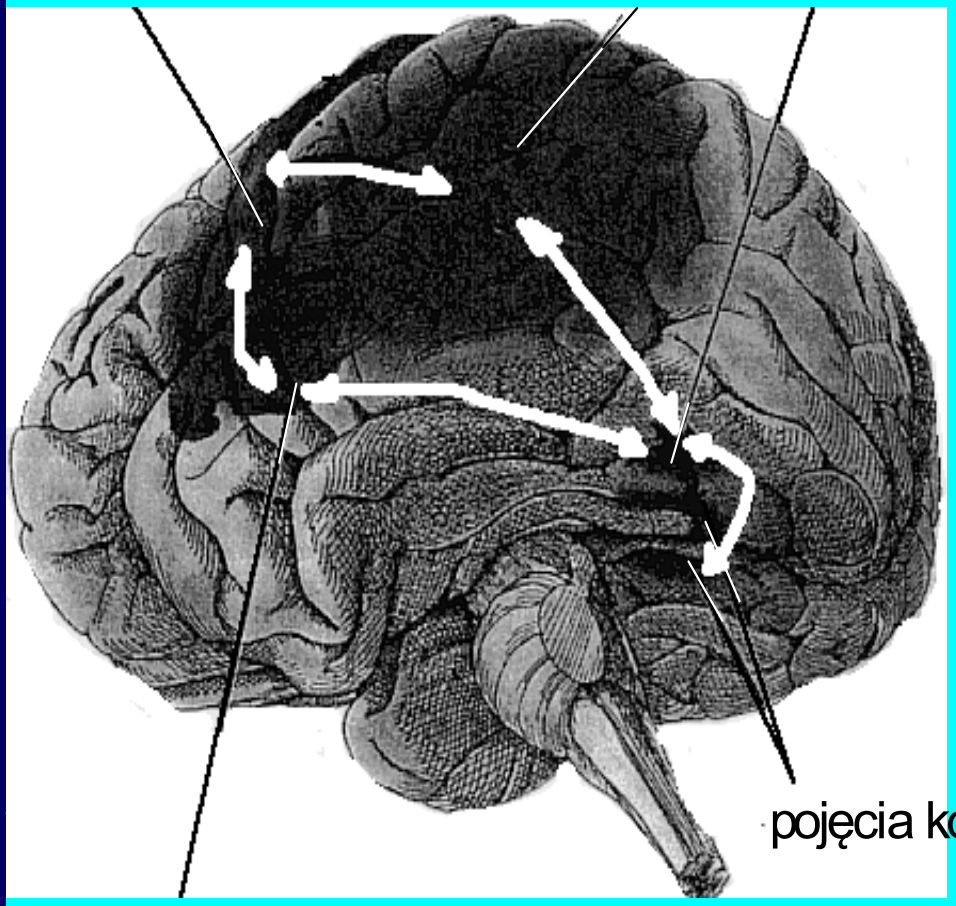
Komputerowe przetwarzanie mowy

Wprowadzenie – komunikacja werbalna



kora ruchowa

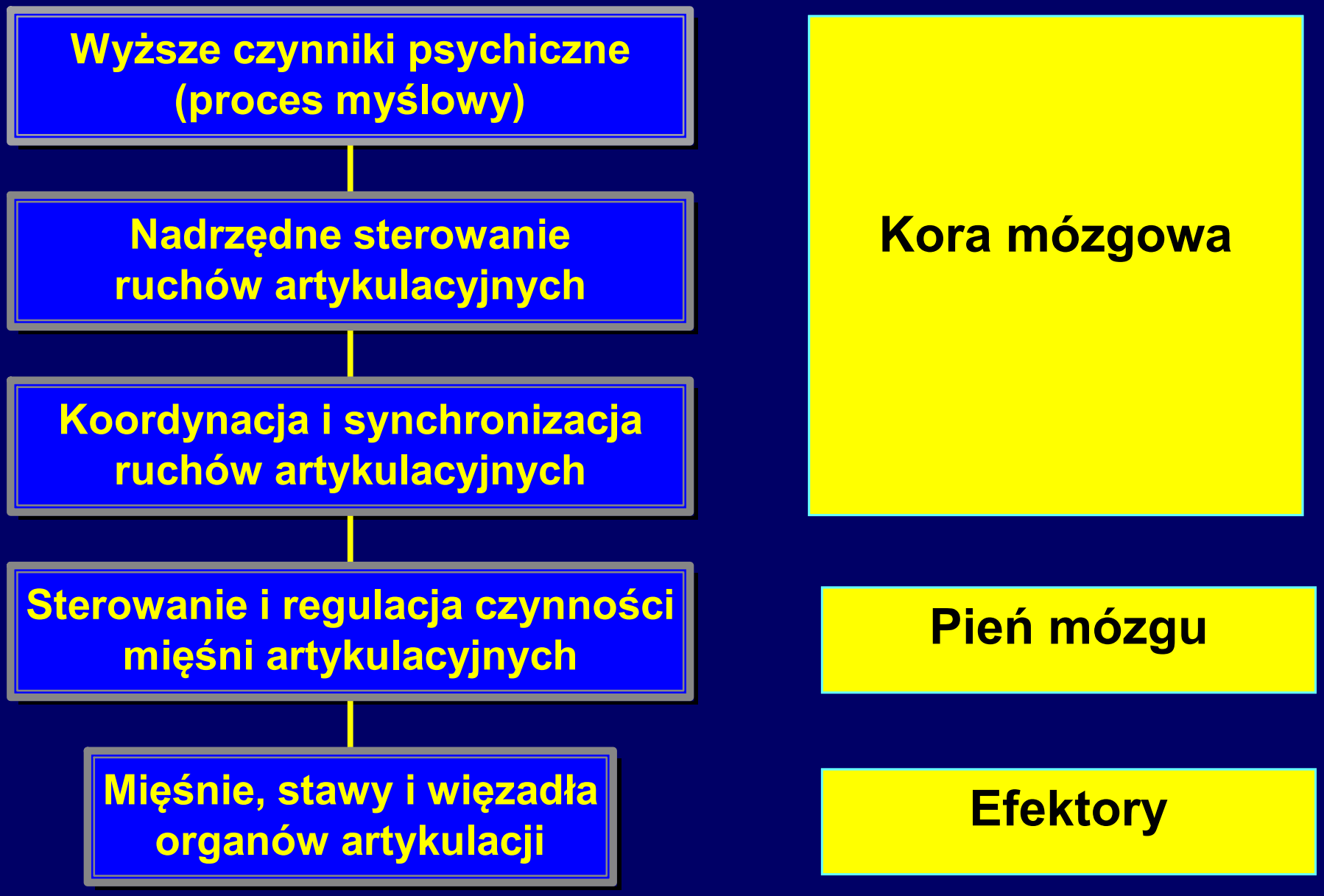
tworzenie form wyrazowych i zdań



pojęcia kolorów

zwoje podstawne lewej półkuli

Struktura systemu wytwarzania mowy



Wytwarzanie mowy

APARAT ARTYKULACYJNY

Składa się z narządów, które modyfikują strumień powietrza.

Na styku jamy gardłowej, ustnej i nosowej powstają głoski ustne i nosowe.

Położenie języka w jamie ustnej decyduje o wytwarzaniu głosek twardych i miękkich.

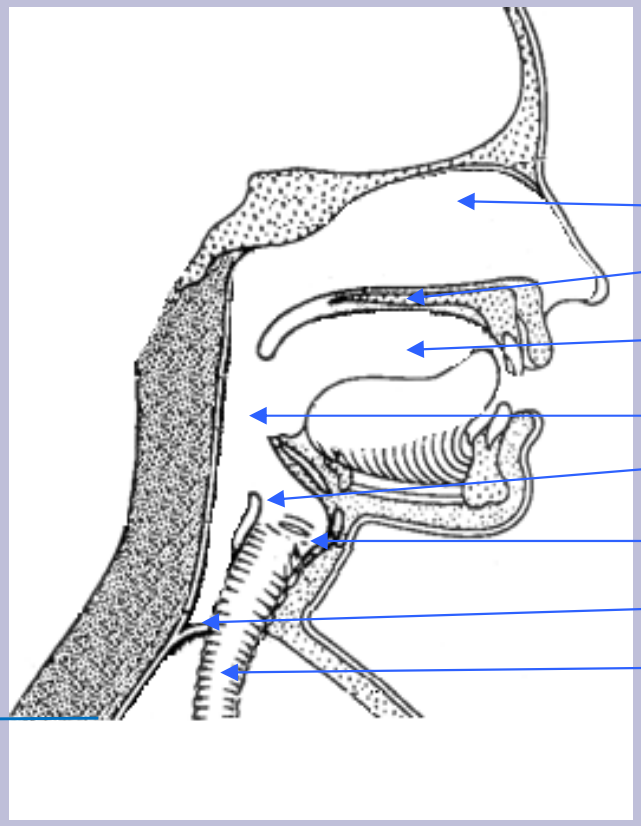
W artykulacji biorą też udział wargi, zęby, dziąsła, podniebienie twarde.

APARAT FONACYJNY

Przy udziale krtani powstają głoski dźwięczne i bezdźwięczne, a położenie wiązadeł głosowych decyduje o ich dźwięczności

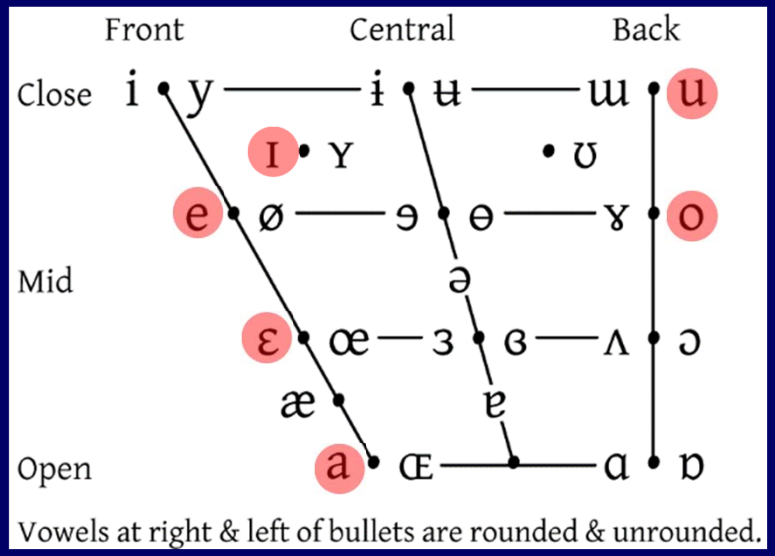
APARAT ODDECHOWY

Dostarczają energię, generującą falę dźwiękową



Międzynarodowy alfabet fonetyczny (angl. International Phonetic Alphabet)

Samogłoski



Jest to standardowy zapis fonetyczny dla wszystkich języków świata.

Alfabet IPA przyjęty w 1886 roku przez Międzynarodowe Towarzystwo Fonetyczne.

Najnowsza wersja alfabetu opublikowana w roku 2005 roku.

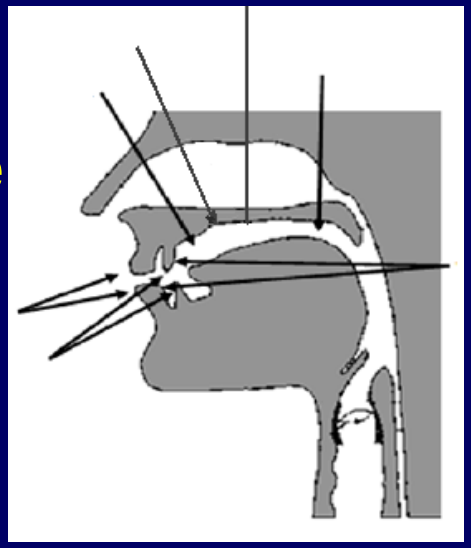
Międzynarodowy alfabet fonetyczny

Spółgłoski	Dwuwargowe (Bilabial)	Wargowo-zębowe (Labiodental)	Zębowe (Dental)	Dziąsłowe (Alveolar)	Zadziąsłowe (Postalveolar)	Retrofleksyjne (Retroflex)	Podniebienne (Palatal)	Miękko-podniebienne (Velar)	Języczkowe (Uvular)	Gardłowe (Pharyngeal)	Krtaniowe (Glottal)
Zwarto-wybuchowe (Plosive)						ʈ ɖ	c ɟ		q ɢ		ʔ
Nosowe (Nasal)	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Drżące (Trill)	ʙ			r					ʀ		
Uderzeniowe (Tap or Flap)				ɾ		ɽ					
Szczelinowe (Fricative)	ɸ β		θ ð			ʂ ʐ	ç ʝ		χ ʁ	ħ ʕ	h ɦ
Boczne szczelinowe (Lateral fricative)				ɬ ɮ							
Aproksymanty (Approximant)		ʋ		ɹ		ɻ	j	ɰ			
Boczne aproksymanty (Lateral approximant)				l		ɭ	ʎ	ʟ			

Podział według miejsca artykulacji

Podniebienne

Zadziąsłowe



Miękko-podniebienne

Dziąsłowe

Wargowo-zębowe

Dwu wargowe

Zębowe

Spółgłoski	IPA	Przykłady
Dwuwargowe	p, p' b, b' m, m'	póvas, petelìškė brolis, labiáu ãmatas, smẽgenys
Wargowo-zębowe	f, f'	fãbrikas, figūrá
Zębowe	t, t' d, d'	tãkas, šaltẽkšnis dárbas, liūdesỹs
Dziąsłowe	s, s' z, z, n, n' l, l'	sáulė, vaĩsius zýlė, zirzė́ti nãmas, nèšti válsas, valià
Zadziąsłowe	ʃ, ʃ' ʒ, ʒ, r, r'	šakà, šiáudas žvãkė, žiógas rãtas, kriáušė
Podniebienne	j	áidas
Miękko-podniebienne	k, k' g, g' x, x' ɣ, ɣ'	kãtinas, kiaũlė gañdras, gérvė chòras, chèmija harmònija, hiacìntas

Podział według sposobu artykulacji

- **Spółgłoski zwarto-wybuchowe**

Zwarcie w jamie ustnej zakańcza się wybuchem

- **Spółgłoski nosowe:**

W jamie ustnej powstaje zwarcie, natomiast w jamie nosowej następuje przepływ powietrza.

- **Spółgłoski drżące:**

między językiem a dziąsłami powstaje zwarcie, przez które w przechodzi powietrze

- **Spółgłoski boczne aproksymanty:**

język zwiera się z zębami. Powietrze przechodzi przez boczną powierzchnią języka a zębami.

- **Spółgłoski szczelinowe**

Powstaje nieduża szczelina, przez którą dostarczane jest powietrze.

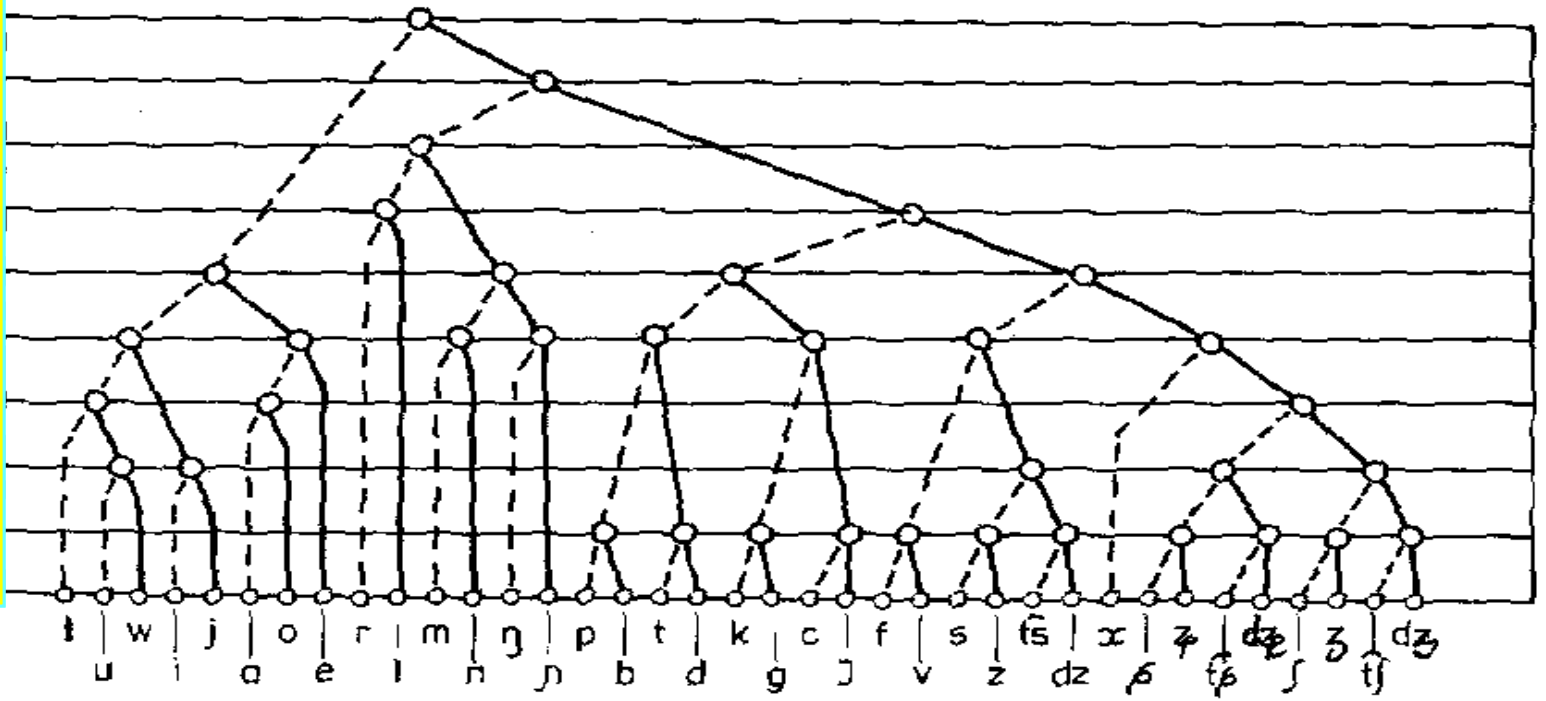
Spółgłoski	IPA	Przykłady
Zwarto-wybuchowe	p, p' b, b' t, t' d, d' k, k' g, g'	póvas, petelìškė brolis, labiáu tãkas, šaltẽkšnis dárbas, liūdesỹs kãtinas, kiaũlė gañdras, gérvė
nosowe	m, m' n, n'	matas, smẽgenys nãmas, nèšti
drżące	r, r'	rãtas, kriáušė
boczne aproksymanty	l, l'	válsas, valià
Szczelinowe	f, f', s, s' z, z' ʃ, ʃ' ʒ, ʒ' x, x' ɣ, ɣ'	fãbrikas, figūrá sáulė, vaĩsius zýlė, zirzė́ti šaka, šiáudas žvãkė, žiógas chòras, chèmija harmònija, hiacìntas

Cechy widma mowy - przykład

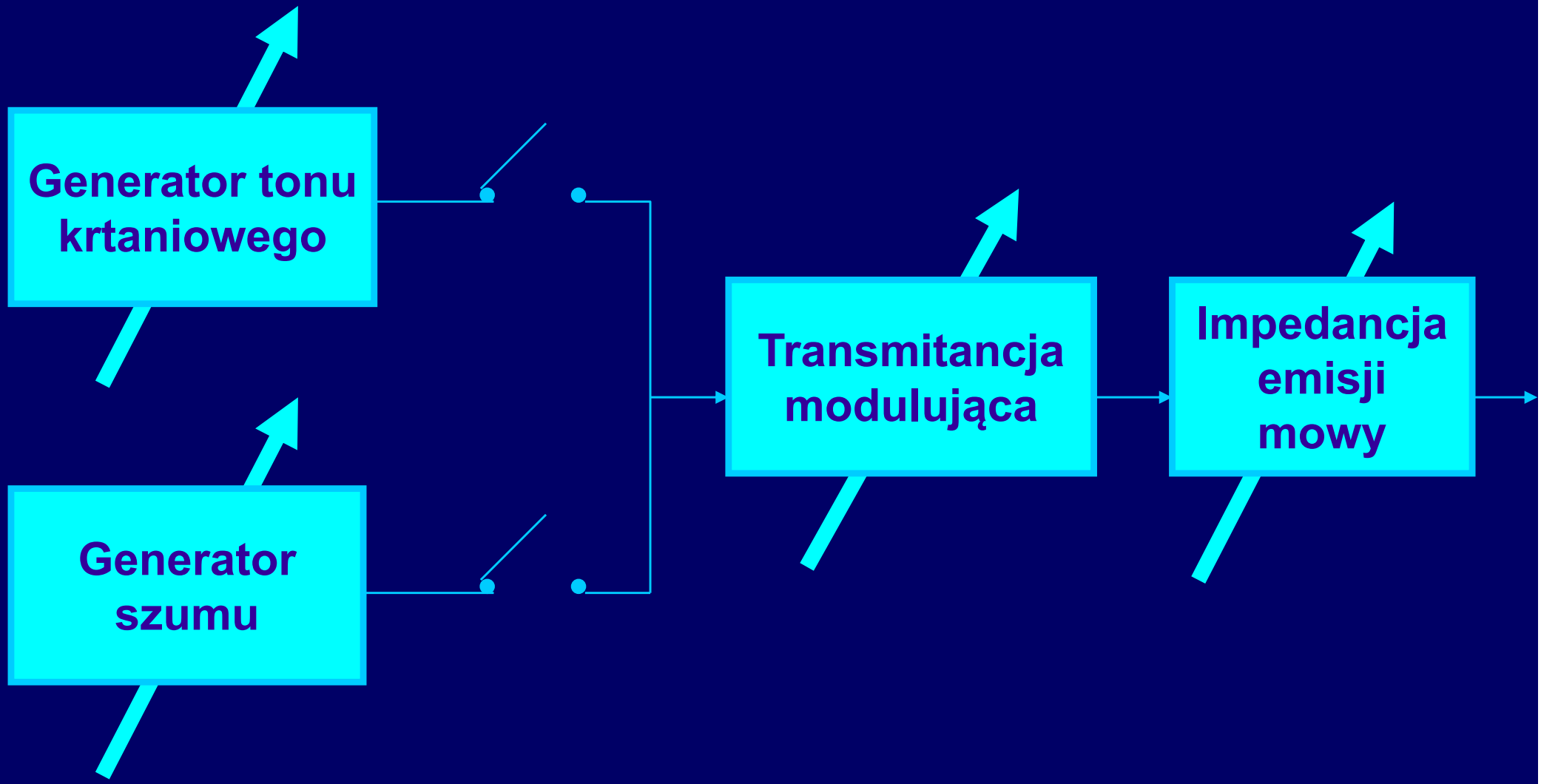
- nosowy-ustny - jeżeli widmo mowy wykazuje więcej niż dwa formanty poniżej 2 kHz, to jest to fonem nosowy. W przeciwnym przypadku fonem jest ustny
- dźwięczny-bezdźwięczny – fonemy dźwięczne charakteryzuje obecność składowej periodycznej, której z kolei brak w fonemach bezdźwięcznych

Najprostszy system rozpoznawania fonemów

Spółgłoskowe
Ponadkrtaniowe
Nosowe
Łagodne
Skupione
Jasne
Niskotonowe
Krótkie
Dźwięczne



Wytwarzanie mowy



Schemat zastępczy systemu artykulacyjnego

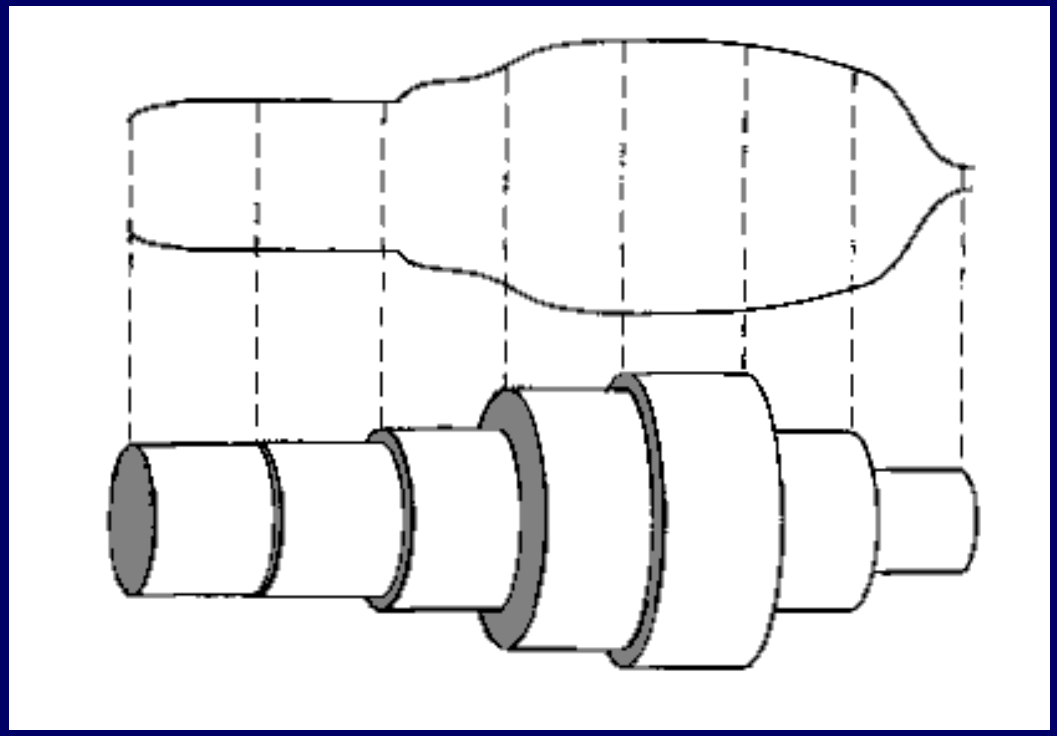
Modelowanie fizyczne - model falowodowy

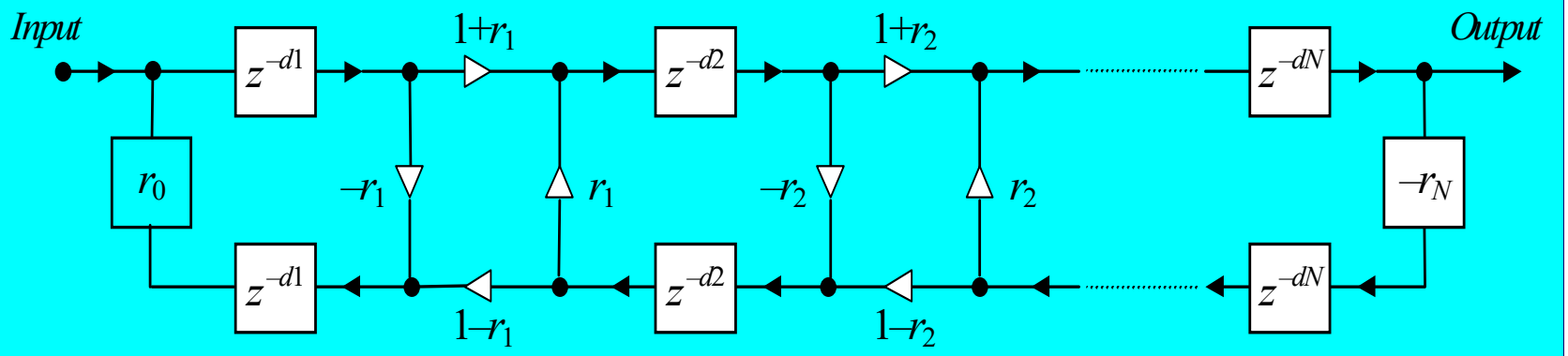
System złożony z N cylindrów o długości L_i i powierzchni A_i (i = 1, 2, ..., N)

Model fizyczny

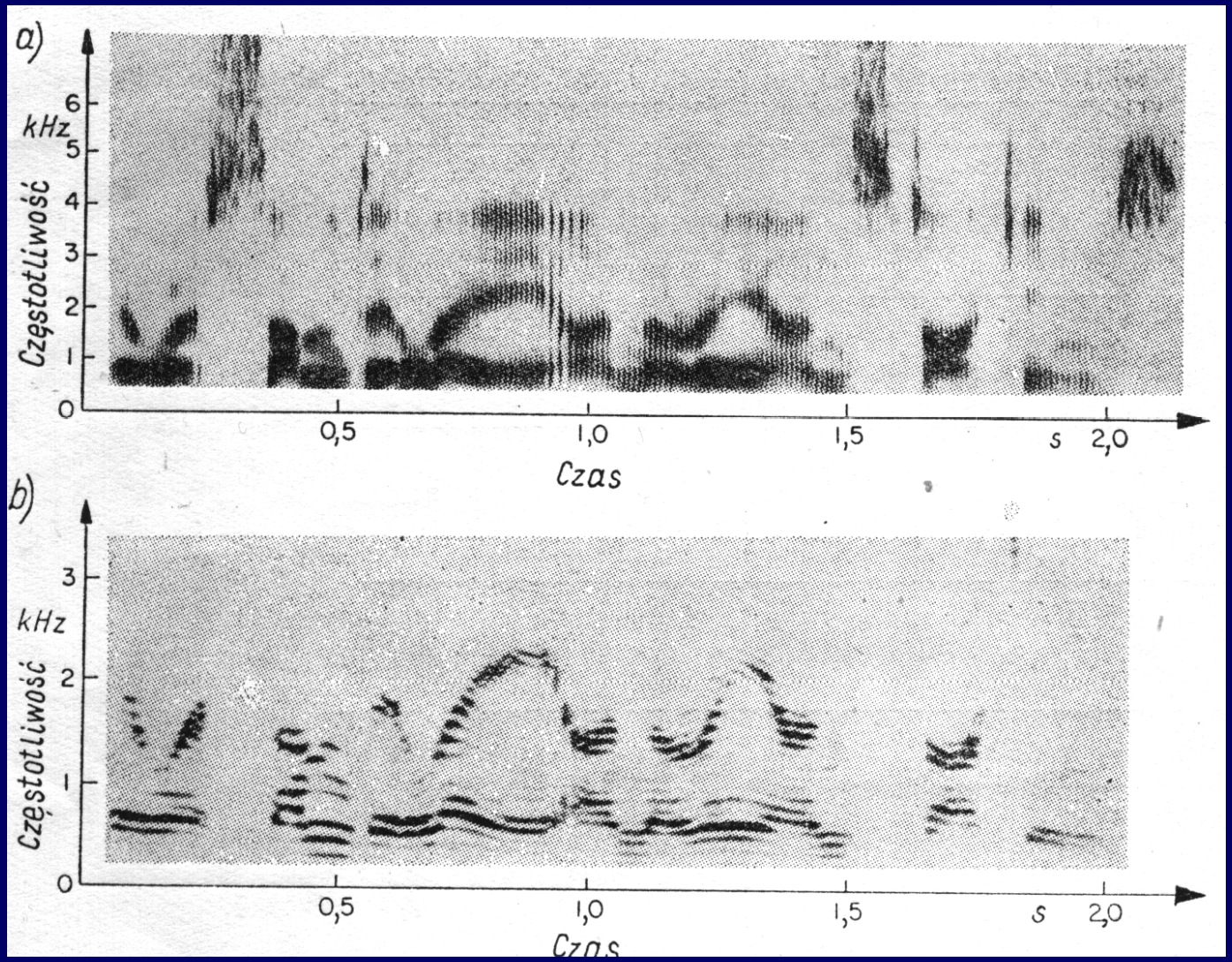
Zbiór rezonatorów cylindrycznych

„Cyfrowy” model falowodowy





Cechy widma mowy



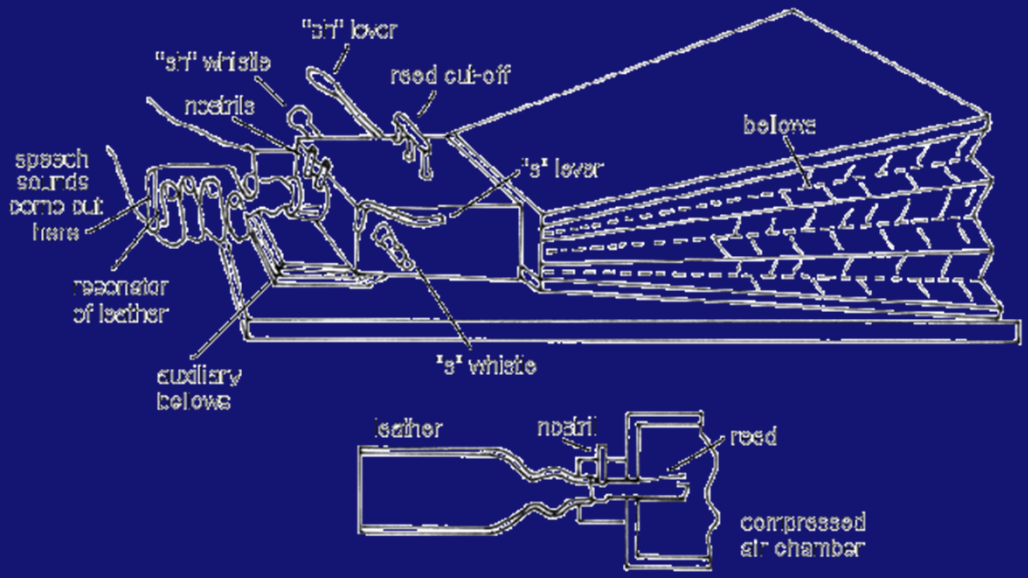
Przykład analizy sonograficznej

Historia syntezy mowy

1773 r. pierwsze badania nad syntezą mowy (profesor Ch.G. Kratzenstein, Kopenhaga)

1846 r. Joseph Faber zaprezentował urządzenie nazwane jako "Euphonia", które generowało nie tylko mowę ludzką, ale także śpiew.

1939 r. pierwszy elektryczny syntezator mowy wykonany przez Homera Dudley'a ("VODER„)



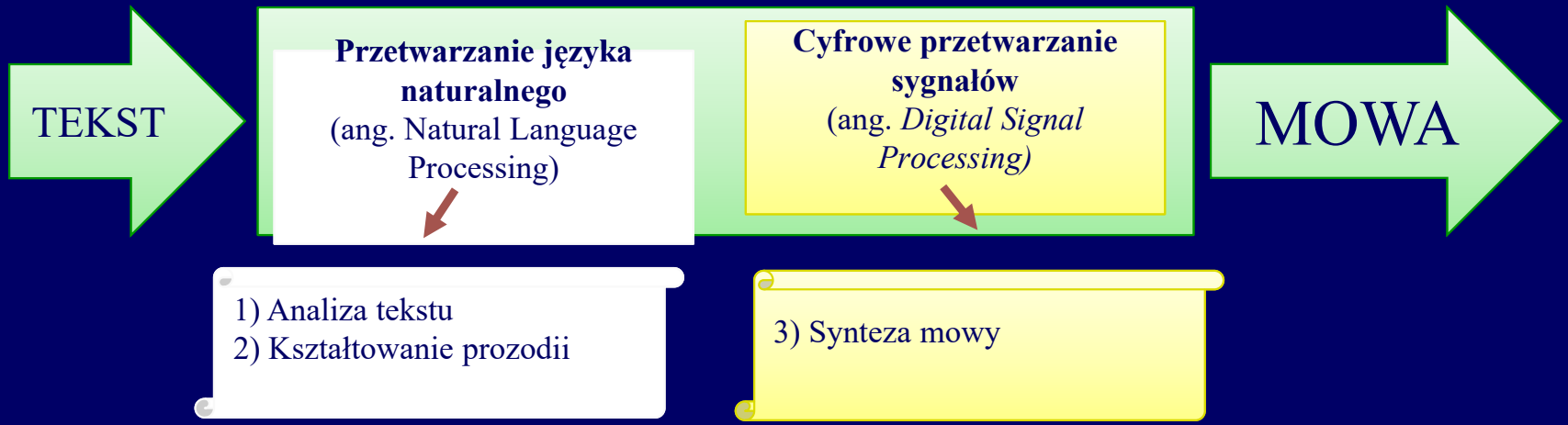
Pierwszy mechaniczny syntezator (von Kempelen, 1791)

Synteza mowy (ang. Text-To-Speech)

Zmiana tekstu na sygnał akustyczny

Podstawowe cele:

- Zrozumiałość wypowiedzi
- Naturalny dźwięk



Przetwarzanie języka naturalnego
Analiza tekstu

Cel analizy:

Przekształcenie tekstu na zapis fonetyczny

Analiza morfologiczna tekstu

Przydzielenie formy podstawowej i wartości cech gramatycznych dla każdego ze słów.

Przykłady:

szafy	szafa, *l. poj., dopełniacz* *l. mnoga, mianownik*
domem	dom, *l. poj., narzędnik*
mówiła	*mówić, czas przeszły, 3osoba l. poj., rodzaj żeński*

Analiza kontekstowa

Zadaniem analizatora kontekstowego jest ograniczenie znaczenia poszczególnych słów. W tym celu badane są części mowy słów znajdujących się w sąsiedztwie.

Analiza kontekstowa obejmuje

- Analizę syntaktyczną (rozpoznanie fraz i ich powiązań składniowych)
- Analizę semantyczną (rozpoznanie obiektów, relacji między nimi)
- Analizę pragmatyczną (interpretacja wypowiedzi w konkretnym kontekście, związki logiczne)

Na danym etapie analizy stosowane są

- Metody n-gramów
- Modele Markowa
- Sieci neuronowe

Analiza prozodyczna

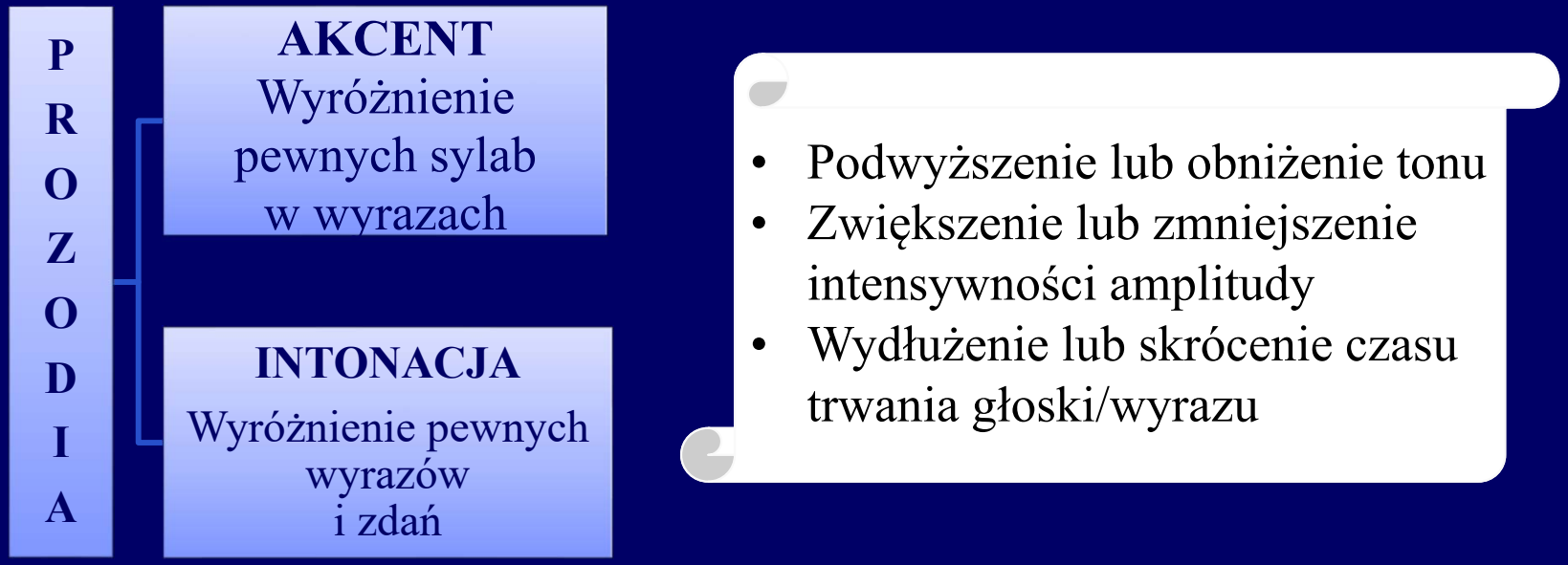
Analizowane są brzmieniowe właściwości mowy nakładające się na głoskowy, sylabiczny i wyrazowy ciąg wypowiedzi.

Prozodie odzwierciedlają:

- Osobiste cechy mówcy
- Stan emocjonalny mówcy
- Cechy wypowiedzi (ironiczny lub sarkastyczny)
- Nacisk, kontrast i ostrość

Kształtowanie prozodii

Kształtowanie prozodii jest niezbędnym procesem dla każdego systemu mowy. Bez zaprogramowania cech emocjonalnych synteza brzmi sztucznie (jak „głos robota”)



Jak stany emocjonalne znajdują swoje odbicie w mowie ?

Ryszard Gubrynowicz

Interpretacja aktorska

happy

sad

angry

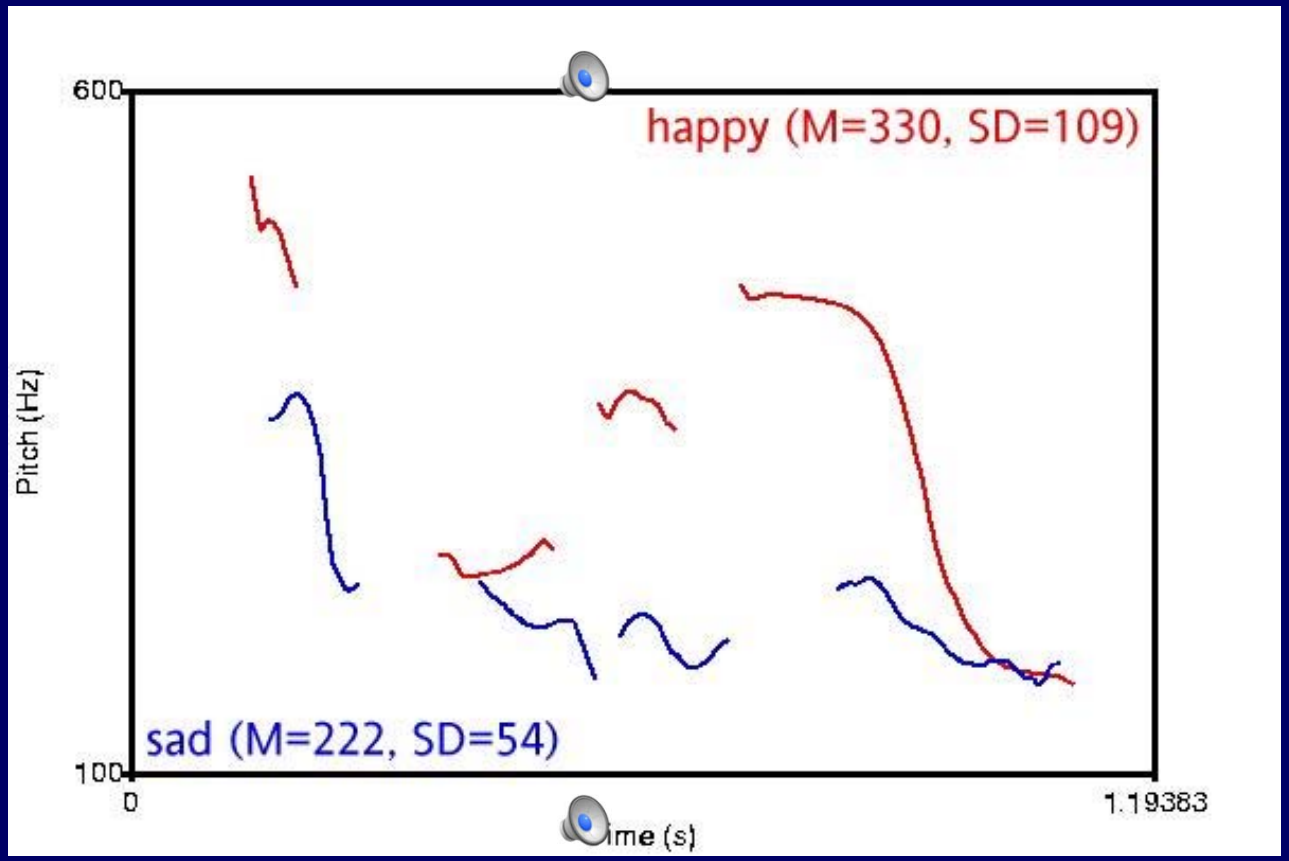
interested

bored

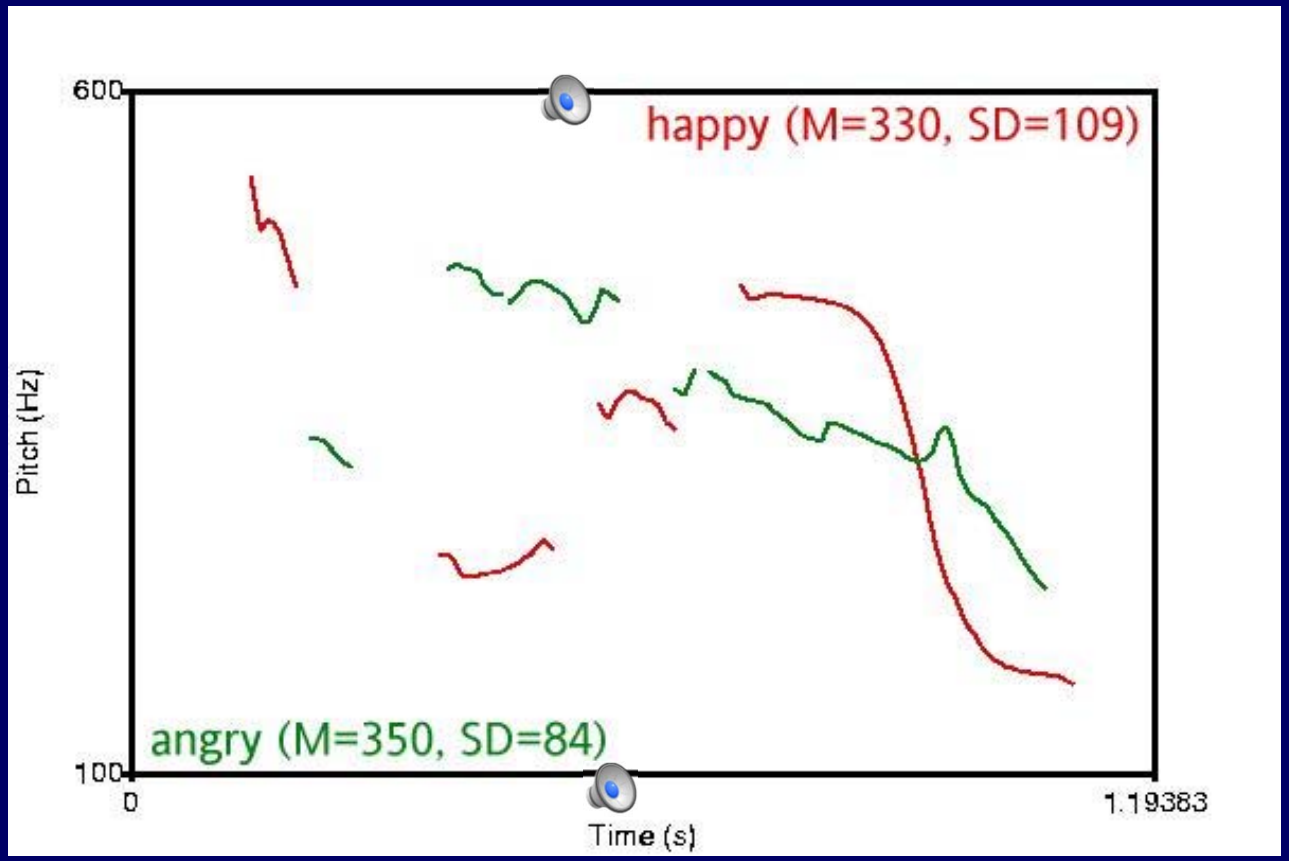
Funkcje emocjonalne cech prozodycznych

Słuchacz na ogół kontroluje w wypowiedzi swój stan emocjonalny. W jego wyrażeniu posługuje się przede wszystkim tempem mówienia, głośnością, wprowadzaniem dodatkowych pauz, przedłużaniem niektórych dźwięków, a także modulowaniem melodii. W wypowiedziach nacechowanych emocjonalnie wahania melodii są znacznie większe, niż w wypowiedziach o charakterze neutralnym. Neutralne – 3-4 tony, z dużym ładunkiem emocjonalnym - > 1 oktawy.

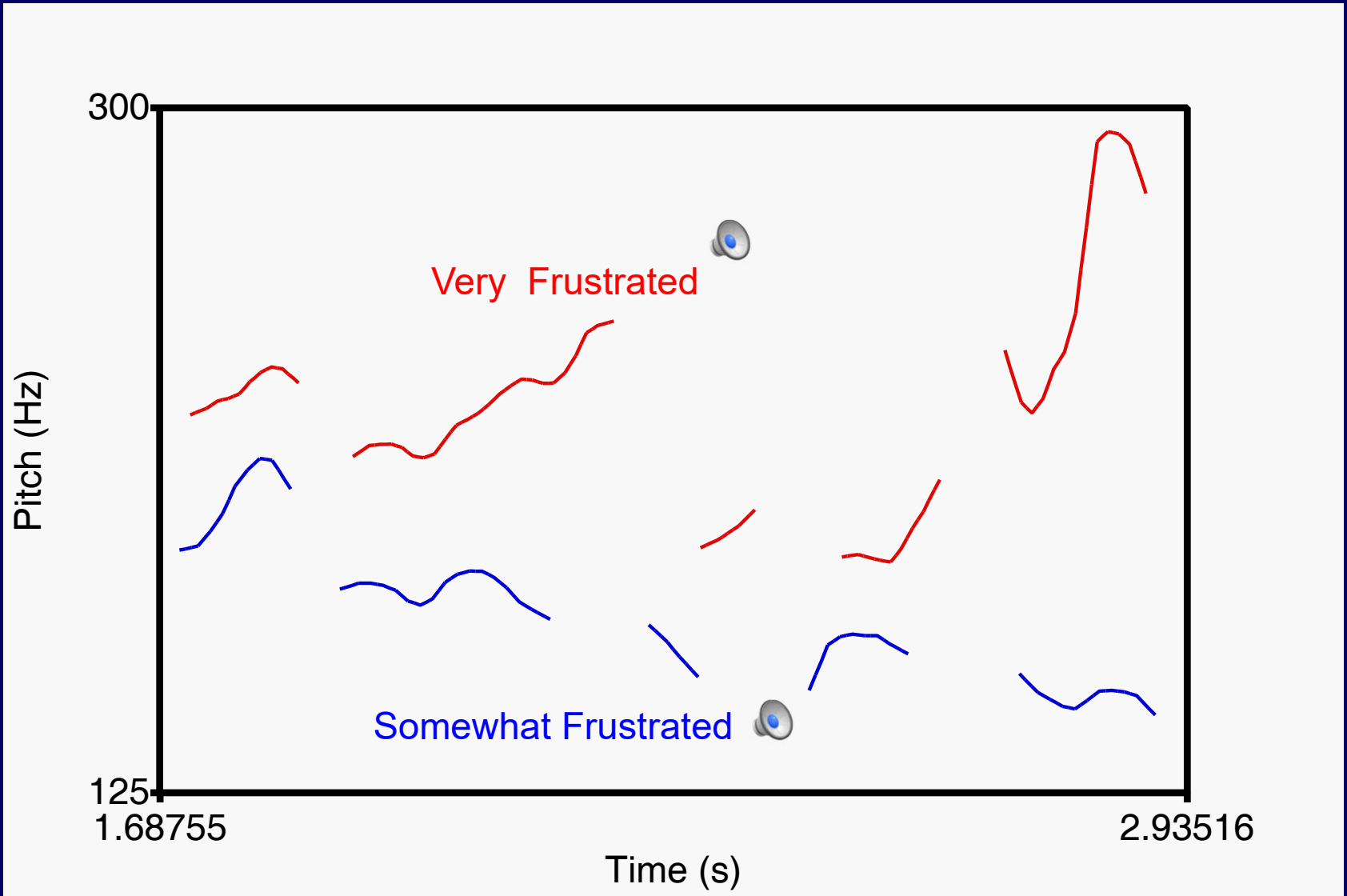
Przykład opozycji szczęśliwy – smutny w konturze melodycznym



Przykład opozycji szczęśliwy – gniewny w konturze melodycznym



Przykład z dialogu typu HMIHY – How may I help you ?

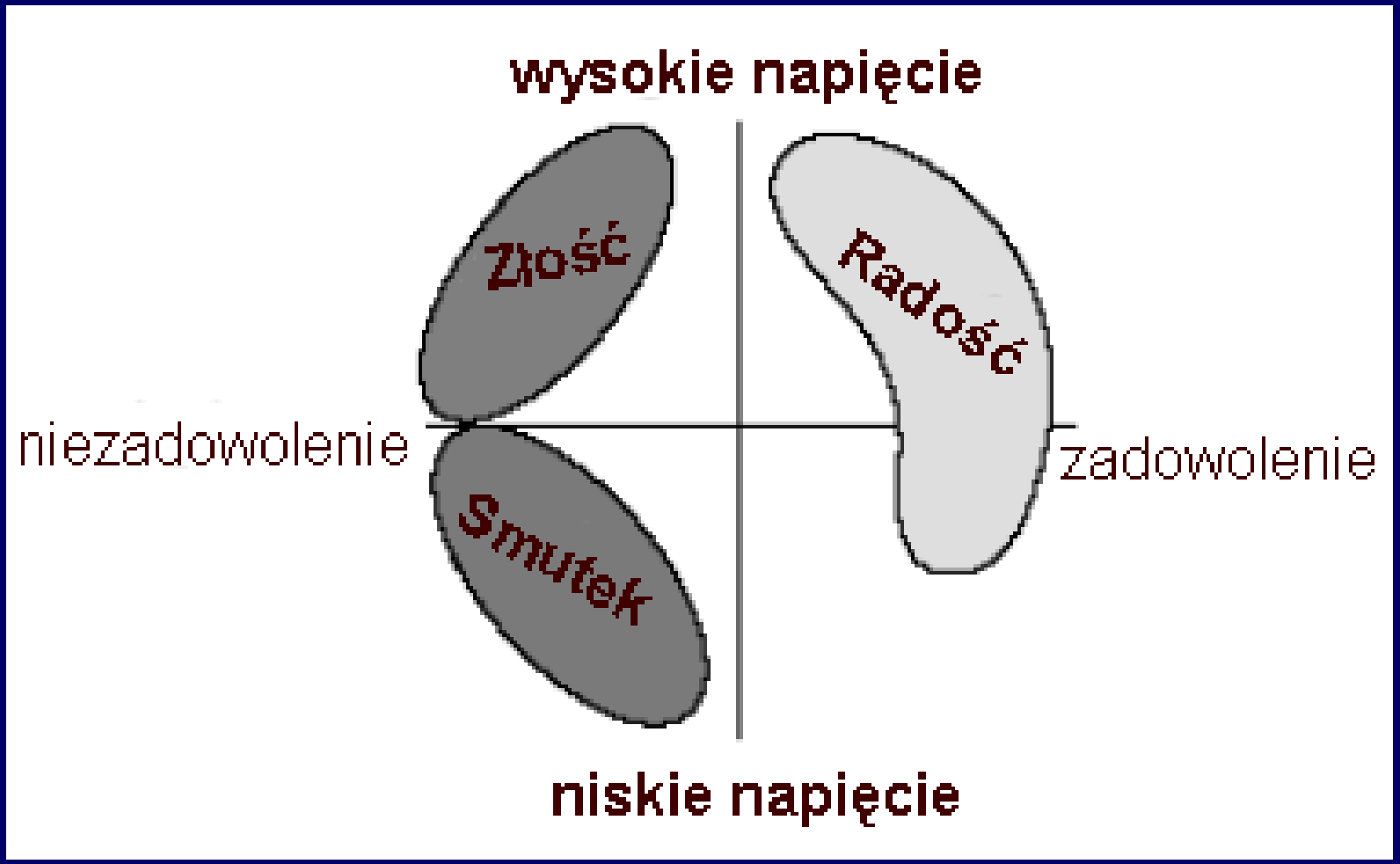


Trudności w określaniu emocji

Nadanie wypowiedzi określonego typu emocji jest zadaniem bardzo złożonym. Osoby określające typ wypowiedzi pod względem emocji rzadko są zgodne w swych ocenach, z wyjątkiem krańcowych, lub wyraźnie kontrastowych typów emocji

Słuchacze w swojej ocenie głównie opierają się na cechach prozodycznych, zwłaszcza na iloczasach i stylizowanym przebiegu F0.

Emocje w płaszczyźnie subiektywnej



Emocje kontrastowe w płaszczyźnie akustycznej

Strach/złość

- zwiększona prędkość i głośność wypowiedzi
- podwyższone F0
- zwiększony zakres F0
- zaburzony rytm mowy
- dokładniejsza artykulacja
- zwiększona energia w zakresie wyższych częstotliwości

Smutek/odprężenie

- zmniejszona prędkość i głośność wypowiedzi
- obniżone F0
- zmniejszony zakres F0
- wyrównany rytm mowy, płynna mowa
- niedokładna artykulacja
- obniżona energia w zakresie wyższych częstotliwości

electronics

Artide

Recognition of Emotions in Speech Using Convolutional Neural Networks on Different Datasets

Marta Zielonka †, Artur Piastowski †, Andrzej Czyżewski *, Paweł Nadachowski, Maksymilian Operlejn and Kamil Kaczor

Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology, Gabriela Narutowicza 11/12, 80-233 Gdańsk, Poland
* Correspondence: ac@pg.edu.pl
† These authors contributed equally to this work.

Abstract: Artificial Neural Network (ANN) models, specifically Convolutional Neural Networks (CNN), were applied to extract emotions based on spectrograms and mel-spectrograms. This study uses spectrograms and mel-spectrograms to investigate which feature extraction method better represents emotions and how big the differences in efficiency are in this context. The conducted studies demonstrated that mel-spectrograms are a better-suited data type for training CNN-based speech emotion recognition (SER). The research experiments employed five popular datasets: Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D), Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), Surrey Audio-Visual Expressed Emotion (SAVEE), Toronto Emotional Speech Set (TESS), and The Interactive Emotional Dyadic Motion Capture (IEMOCAP). Six different classes of emotions were used: happiness, anger, sadness, fear, disgust, and neutral. However, some experiments were prepared to recognize just four emotions due to the characteristics of the IEMOCAP dataset. A comparison of classification efficiency on different datasets and an attempt to develop a universal model trained using all datasets were also performed. This approach brought an accuracy of 55.89% when recognizing four emotions. The most accurate model for six emotion recognition was trained and achieved 57.42% accuracy on a combination of four datasets (CREMA-D, RAVDESS, SAVEE, TESS). What is more, another study was developed that demonstrated that improper data division for training and test sets significantly influences the test accuracy of CNNs. Therefore, the problem of inappropriate data division between the training and test sets, which affected the results of studies known from the literature, was addressed extensively. The performed experiments employed the popular ResNet18 architecture to demonstrate the reliability of the research results and to show that these problems are not unique to the custom CNN architecture proposed in experiments. Subsequently, the label correctness of the CREMA-D dataset was studied through the employment of a prepared questionnaire.

Keywords: speech emotion recognition; SER; machine learning; artificial intelligence; classification; convolutional neural networks

check for updates

Citation: Zielonka, M.; Piastowski, A.; Czyżewski, A.; Nadachowski, P.; Operlejn, M.; Kaczor, K. Recognition of Emotions in Speech Using Convolutional Neural Networks on Different Datasets. *Electronics* 2022, *11*, 3831. https://doi.org/10.3390/electronics11223831

Academic Editors: Daniel Hládek, Matúš Pleva, Piotr Szczuko and Andrej Zgank

Received: 9 August 2022
Accepted: 15 November 2022
Published: 21 November 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

The recognition of emotions is a relatively difficult and complex task [1], even for humans. Many people could say that they can perform this task efficiently; however, they often have the opportunity to recognize emotions based on a few different aspects, such as body language, facial expression, and voice timbre or prosody. Meanwhile, speech emotion recognition (SER) is a potentially significant step toward the future as it presents a huge variety of use cases.

SER considers recognizing emotions using only one modality, voice recordings, which makes it more complex. Thus, it uses one additional medium—a microphone that may also capture some noise [2]. Achieving decent results on this type of problem could lead to the

Rodzaje syntezy mowy

- **Metoda formantowa**

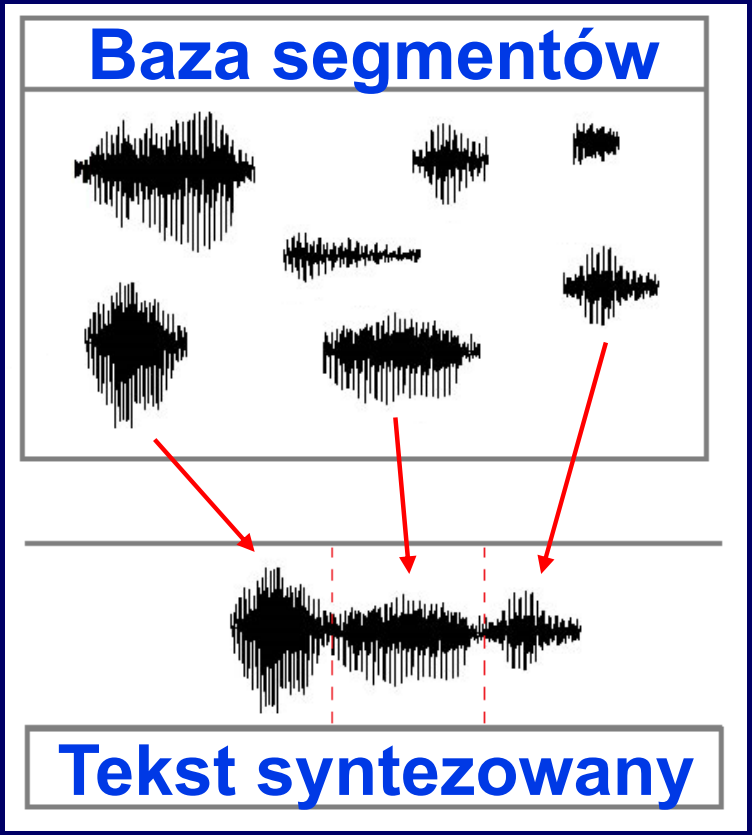
  Odwzorowanie widma sygnału mowy
- **Metoda artykulacyjna**

  Fizyczne odwzorowanie mechanizmów wytwarzania mowy

Metoda konkatenacyjna

Wykorzystanie nagranych próbek sygnału mowy

Konkatenacyjna synteza mowy



Łączenie wypowiedzi z mniejszych jednostek nagranych przez lektora

Wykorzystywane jednostki:

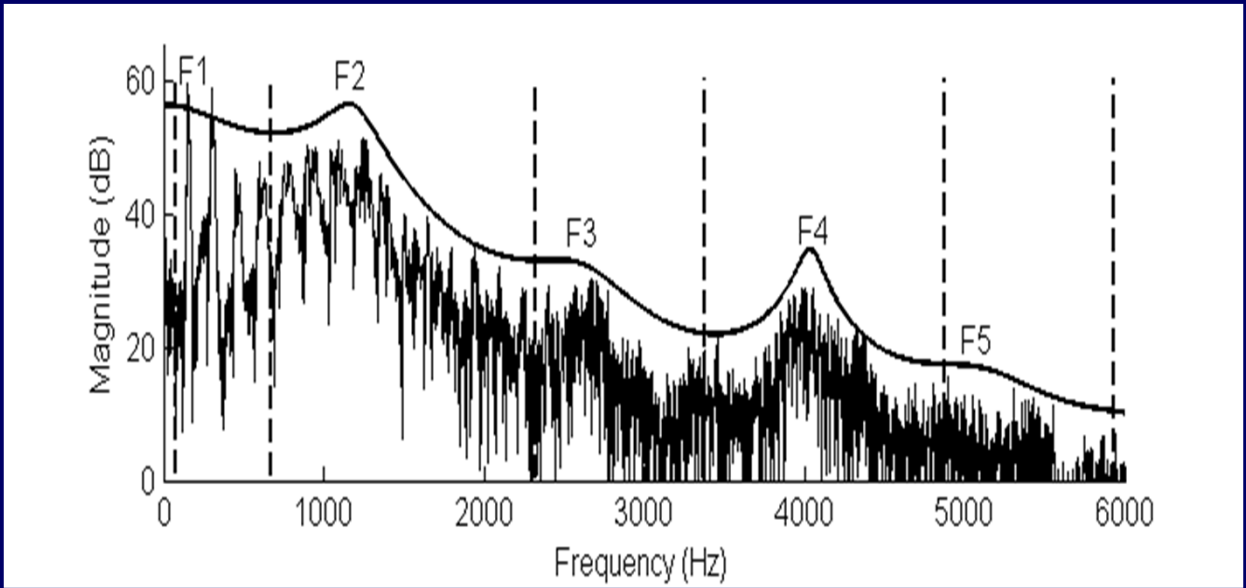
- fonem (głoska)
- difony
- trifony
- sylaby
- całe wyrazy

Jest to najczęściej spotykana metoda syntezy.

Formantowa synteza mowy

Modelowanie traktu głosowego jako połączenie rezonatorów – filtrów elektrycznych lub cyfrowych.

Podejście to ma w założeniu odwzorować formantowy charakter sygnału mowy.



Formant - skupisko energii w widmie sygnału mowy.

Od rozmieszczenia formantów zależy zrozumiałość mowy.

Artykulacyjna synteza mowy

Zakłada się, że głos powstaje w trakcie głosowym (układ filtrów - rezonatorów o zmiennych parametrach) za pomocą sygnału pobudzającego

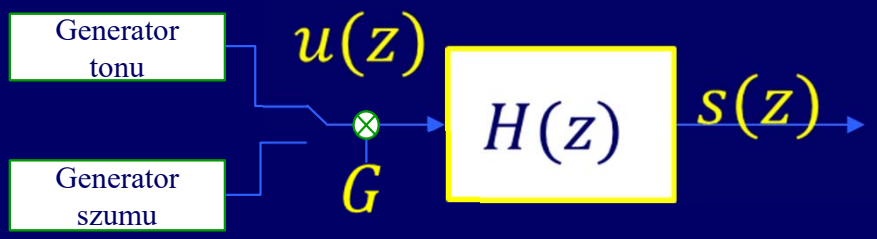
Sygnał pobudzający - struny głosowe (oddziaływanie strumienia powietrza i fałd głosowych lub szumu białego)

Najczęściej używa się kodowania predykcyjnego (*Linear Predictive Coding*).

1. Obliczanie *charakterystyki traktu głosowego*

Analiza LPC

2. Odwzorowanie *charakterystyki traktu głosowego* za pomocą modelu matematycznego.

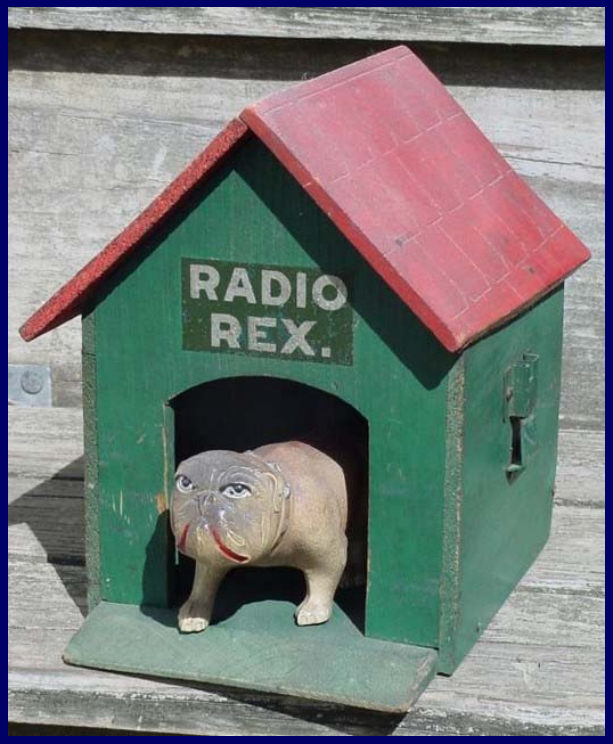


$$\frac{1}{-\sum_{k=1}^{p} a_k}$$

Zastosowania syntezy mowy

- Urządzenia dla osób niewidomych
- Mówiące telefony, komputery, planszety
- Słowniki językowe
- Udźwiękowianie stron internetowych, aplikacji, gier edukacyjnych
- Odczyt poczty elektronicznej

Historia rozpoznawania mowy



Zabawka Radio Rex powstała w 1920 roku

1937 r. Stevens i Newman zdefiniowali melową skalę częstotliwości

1952 r. Naukowcy z Bell Labs wynaleźli system rozpoznawania cyfr izolowanych.

1965 r. Cooley i Tukey opracowali algorytm szybkiej transformacji Fouriera.

Rozpoznawanie mowy

System może być zależny i niezależny od mówcy

Wielkość słownika

Słownik	Ilość wyrazów
Mały	2 – 100 wyrazów
Średni	100 – 1000 wyrazów
Duży	ponad 1000 wyrazów

Obecnie systemy są w stanie rozpoznać 50 tysięcy słów

Rozpoznawanie mowy

Ekstrakcja parametrów

- MFCC
- LPC
- ...

$Y = y_1, y_2, \ldots, y_T$

Modele akustyczne

$P(Y|w_i)$

Słownik

w_i

Dekodowanie

$\max_i \{P(w_i|Y)\}$

- HMM

$P(w_i)$

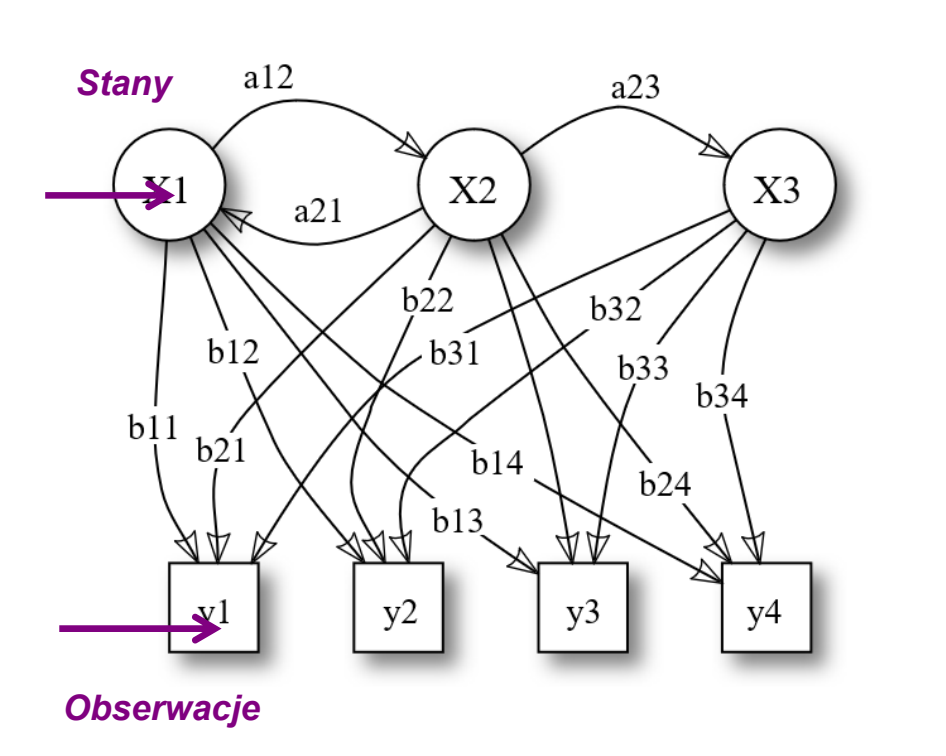
Modele językowe (n-gramy)

Tekst

Twierdzenie Bayesa:

$$P(w_i|Y) = \frac{P(Y|w_i)P(w_i)}{P(O)}$$

Dekodowanie sygnału za pomocą ukrytych modeli Markowa



https://upload.wikimedia.org/wikipedia/commons/8/8a/HiddenMarkovModel.svg

ang. Hidden Markov Models (HMM)

Obliczenie prawdopodobieństwa $P(\mathrm{Y}|w_i)$ sprowadza się do obliczenia sumarycznego prawdopodobieństwa (zdarzeń i przejść).

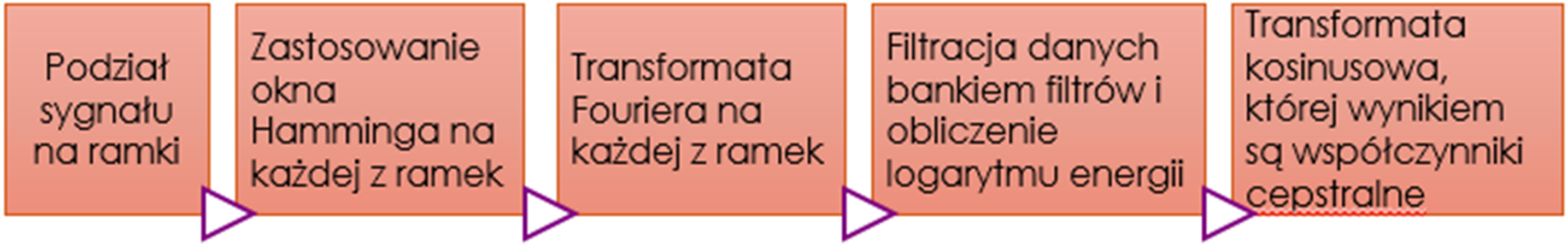
W ukrytym modelu Markowa stan nie jest widoczny, jednak wyjście zależne od niego jest znane.

Do odkrywania ukrytej sekwencji stanów modelu HMM stosuje się algorytm *Viterbiego*

Ekstrakcja parametrów - metody cepstralne

ang. Mel Frequency Cepstral Coefficient (MFCC)

- Cepstrum - to transformata Fouriera logarytmu widma $\hat{X}(T) = F[\ln(X(f))]$.
- Skala cepstrum odpowiada dziedzinie czasu
- Współczynniki cepstralne niosą informacje o trakcie głosowym i o tonie krtaniowym
- Skala melowa, określająca subiektywny odbiór wysokości dźwięku przez ludzkie ucho względem skali w hercach $Fmel = 1127 loge\ 1 + f\ 700$



Analiza mowy – parametryzacja

- współczynniki cepstralne (MFCC) w skali nieliniowej (melowej)

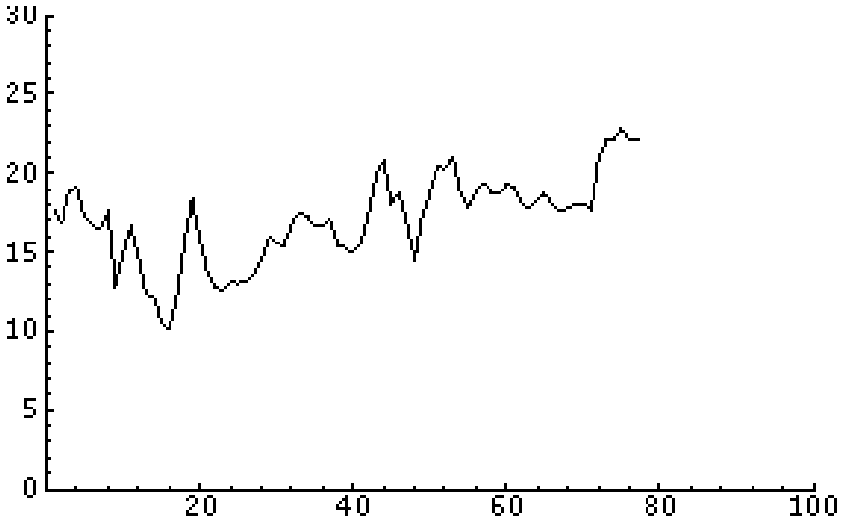
$$M_i = \sum_{k=1}^{20} X_k \cos\left[i(k-0.5)\pi/20\right]$$

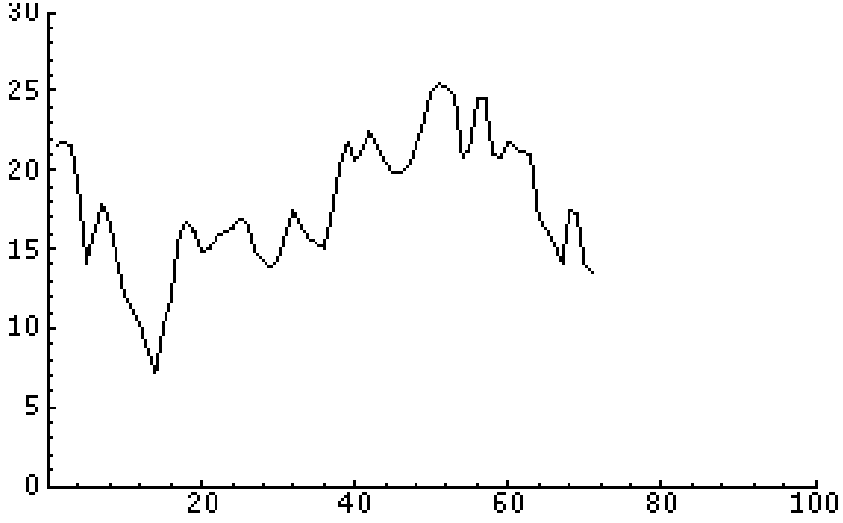
gdzie: *i* - numer współczynnika cepstralnego;
k - liczba pasm częstotliwości
X_k - logarytm energii w danym paśmie częstotliwości *k*

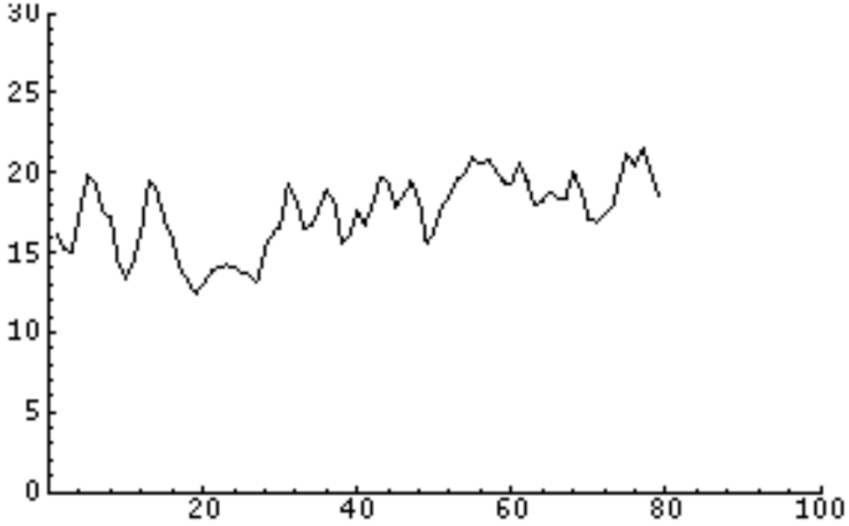
- proste parametry, np. liczba przejść przez zero lub przez inną wartość (w celu ograniczenia wpływu składowej stałej)

- analiza LPC – współczynniki LPC

Współczynniki mel-cepstralne

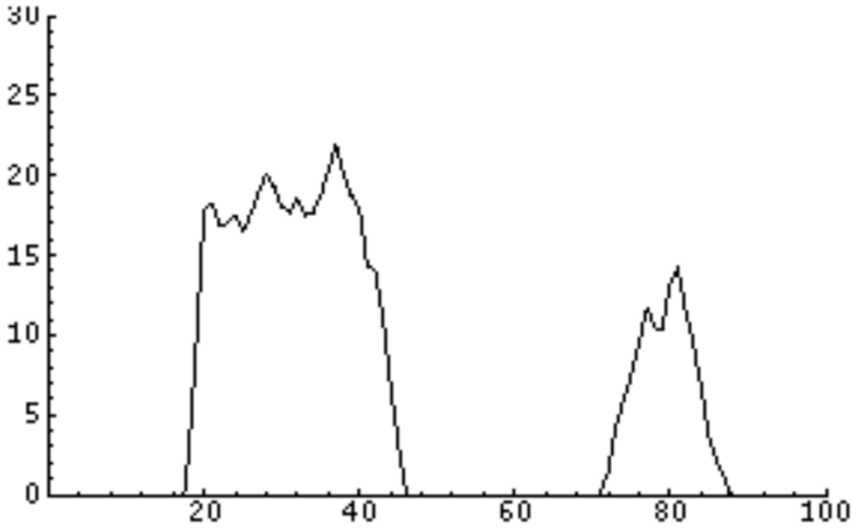


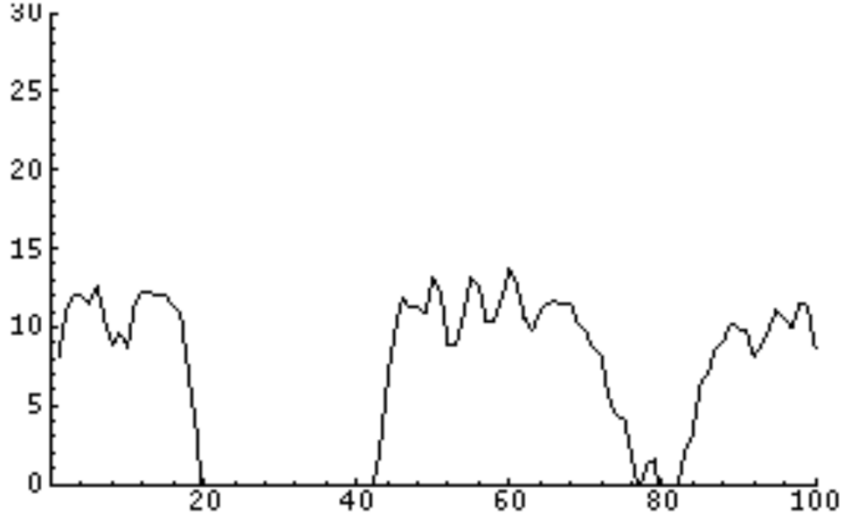


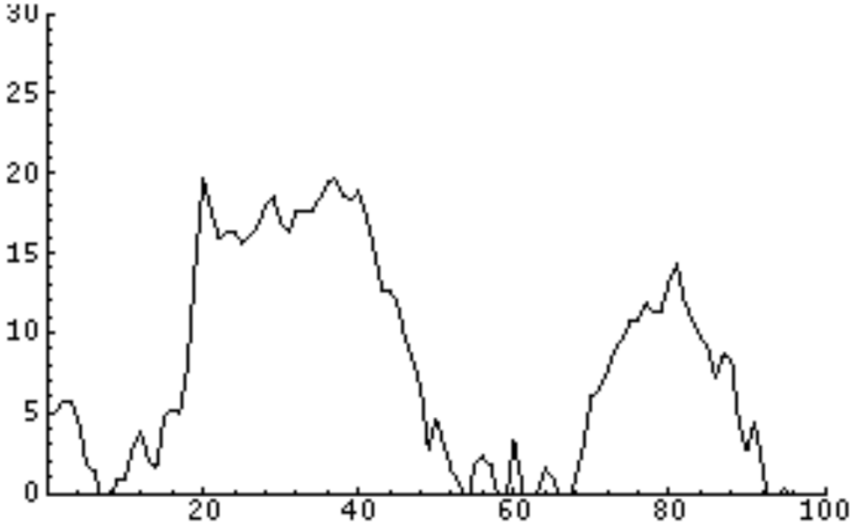


Słowo „zero” - trzech mówców

Współczynniki mel-cepstralne





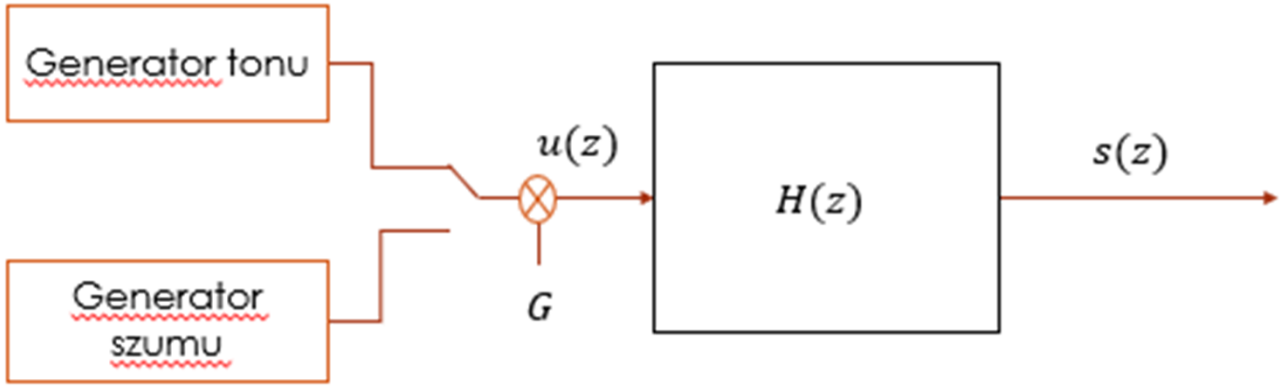


Słowo „sześć” - trzech mówców

Ekstrakcja parametrów - metody predykcyjne

ang. Linear Predictive Coding (LPC)

Odwzorowuje rezonansową strukturę traktu głosowego

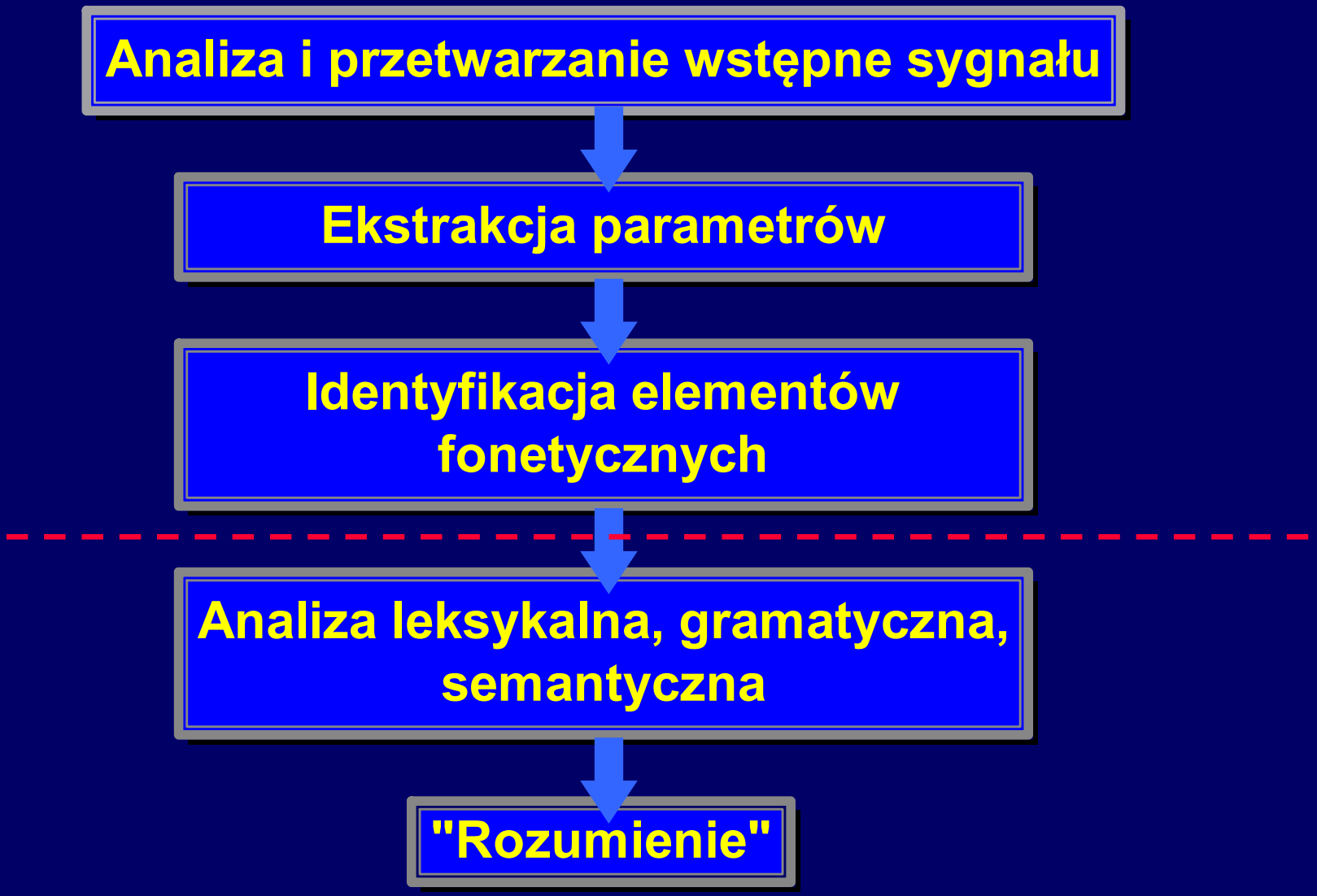


$$H(z) = G\frac{1}{1-\sum_{k=1}^{p} a_k \cdot z^{-k}}$$

Sygnał mowy - odpowiedź filtru na pobudzenie
Filtr - rezonansowa charakterystyka traktu głosowego
Pobudzenie - sygnał tonu krtaniowego

Podział systemów ARM

Proces rozpoznawania sygnału mowy



Analiza mowy – przetwarzanie wstępne

- **Normalizacja energetyczna**
- **Segmentacja sygnału (detekcja granic wyrazów)**

Przykładowo:

- Segmentacja poprzez analizę obwiedni amplitudowej

$$p_i - p_{i-1} > k \vee p_i - p_{i+1} > k$$

gdzie:

p_i - i-ta próbka sygnału

k - arbitralnie przyjęta wartość progowa

$$c = \frac{\int_{t_1}^{t_2} ts(t)dt}{\int_{t_1}^{t_2} s(t)dt}$$

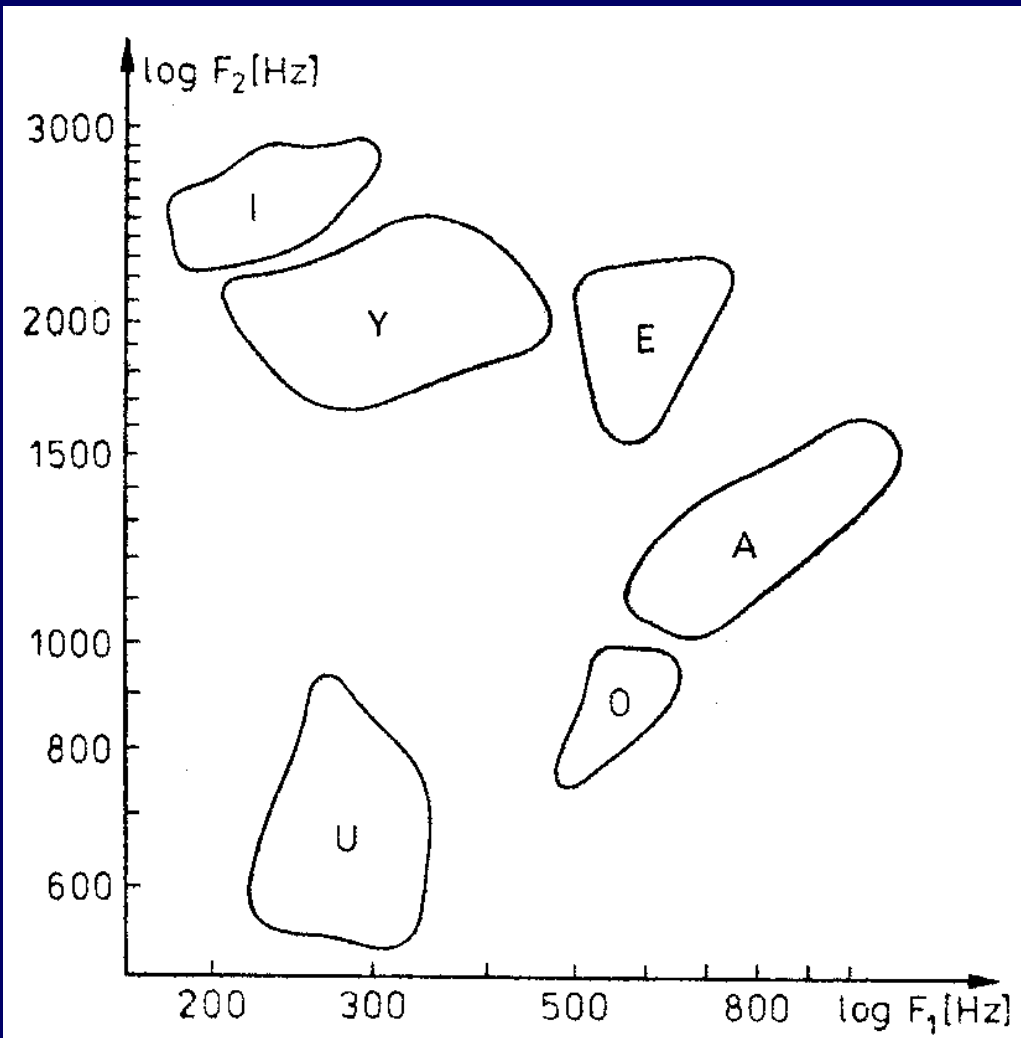
$$d = \sqrt{\frac{\int_{t_1}^{t_2} (t-c)^2 s(t)dt}{\int_{t_1}^{t_2} s(t)dt}}$$

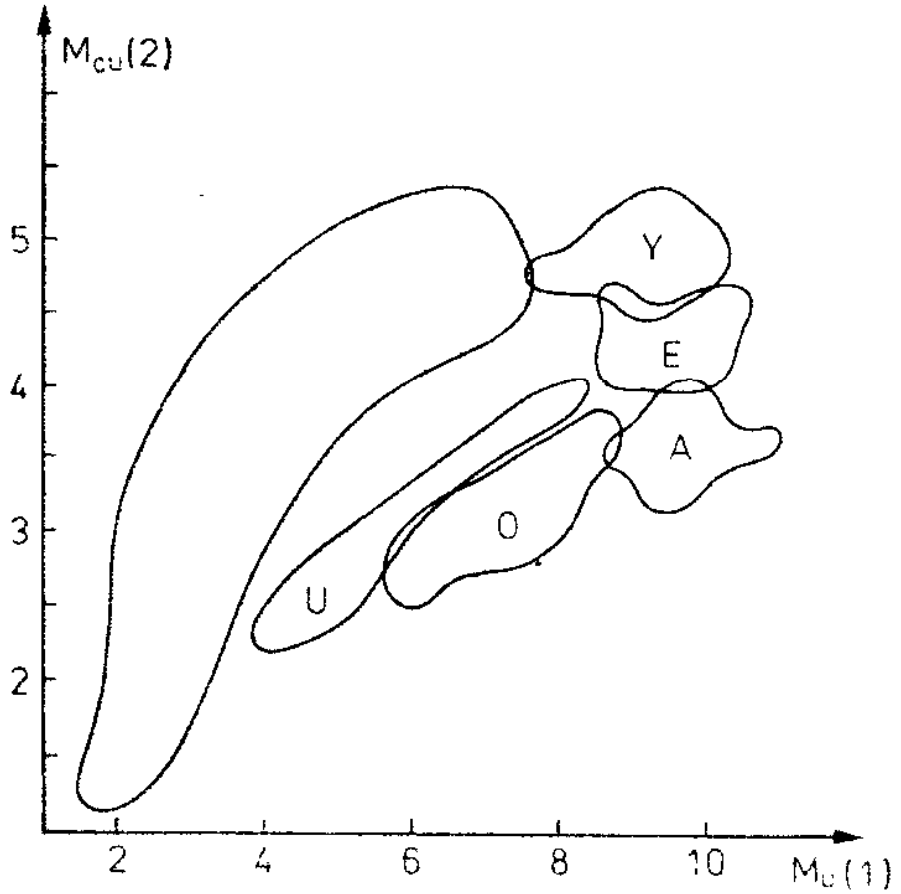
$$k_1 \cong c - wd$$

$$k_2 \cong c + wd$$

gdzie: c - środek ciężkości, d – dyspersja, t_1, t_2 – dowolna próbka „przed” i „za” wyrazem, $s(t)$ – rozkład czasowy funkcji gęstości p, k_1, k_2 – granice wyrazu (numer próbki),

Ekstrakcja parametrów - fonemy samogłoskowe

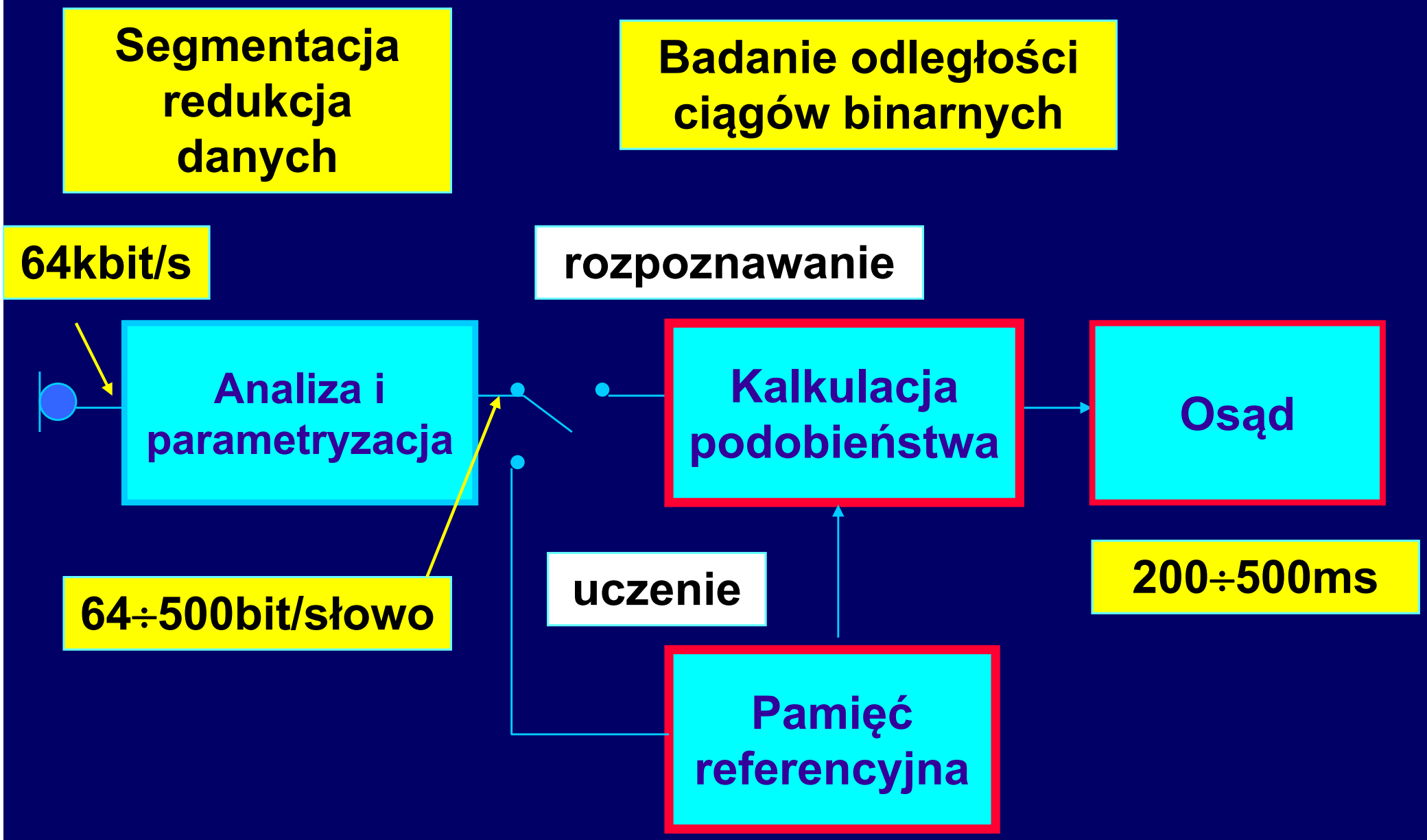




Formanty F_1 i F_2

Momenty centralne M_{c1} i M_{cu2}

Automatyczna klasyfikacja

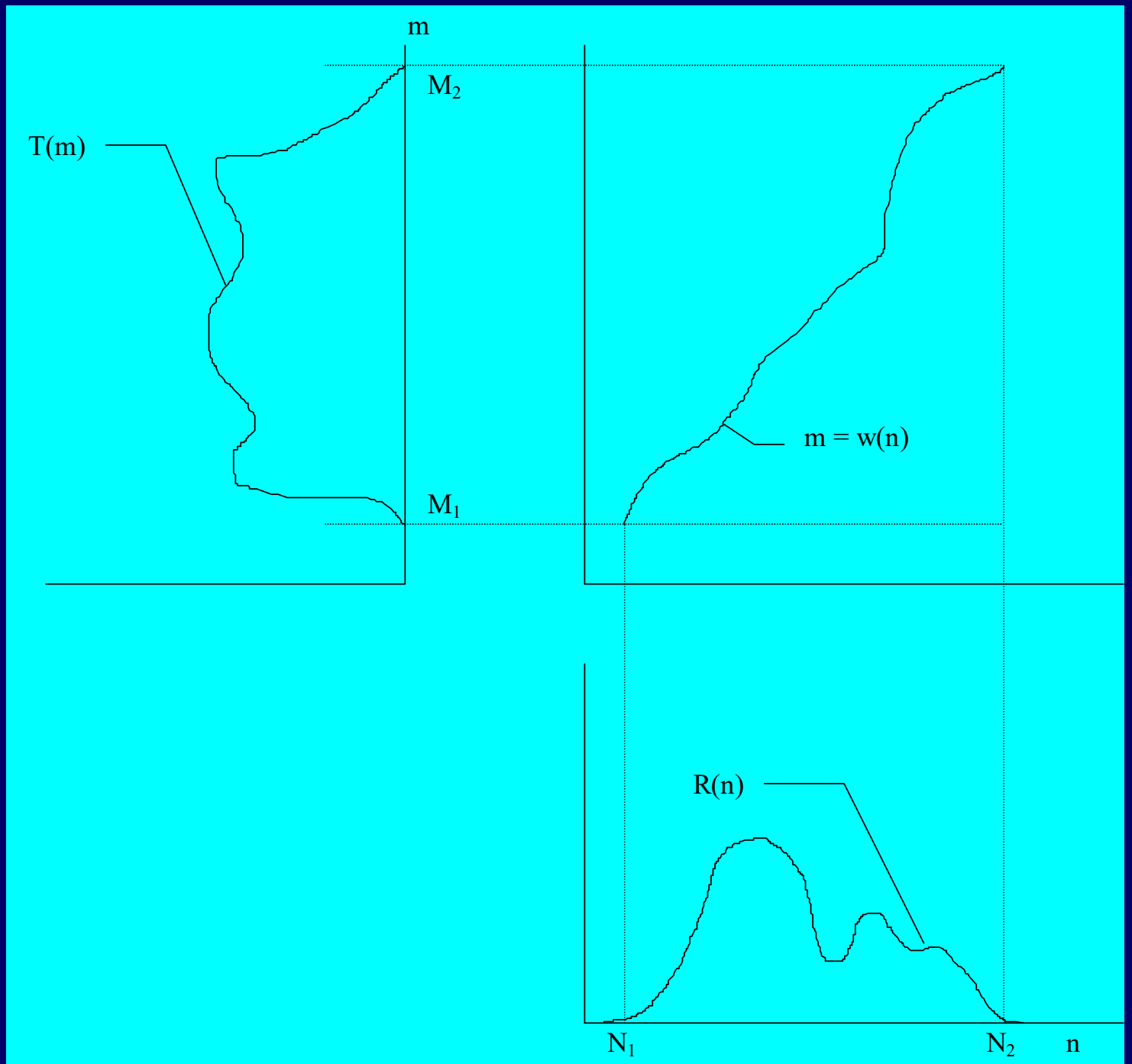


ARM – systemy decyzyjne

Metody rozpoznawania izolowanych wyrazów:

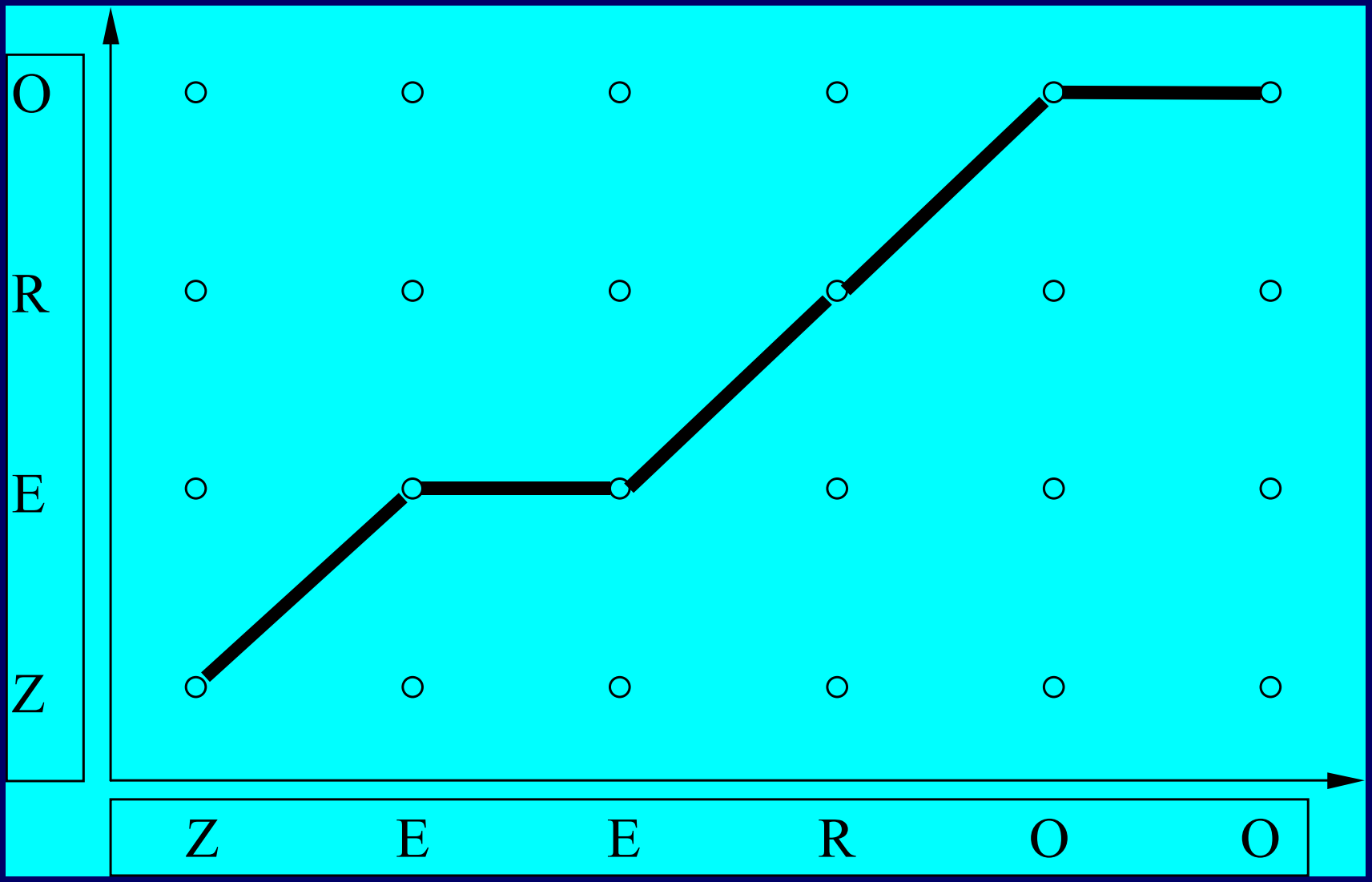
- nieliniowa normalizacja czasowa
- niejawne modele Markowa (HMM)
- sztuczne sieci neuronowe
- metoda zbiorów przybliżonych

Algorytm nieliniowego dopasowania czasowego



Proces nieliniowego dopasowania czasowego

Algorytm nieliniowego dopasowania czasowego



Ilustracja procesu nieliniowego dopasowania czasowego w przypadku izolowanych wyrazów

Algorytm nieliniowego dopasowania czasowego

- Dopasowanie można przedstawić jako funkcję $m = w(n)$

przy spełnionych warunkach brzegowych: $M_1 = w(N_1)$ $M_2 = w(N_2)$

oraz warunków ciągłości (następstwo segmentów)

$$w(n+1) - w(n) = 0,1,2 \quad (w(n) \neq w(n-1))$$

$$w(n+1) - w(n) = 1,2 \qquad (w(n) = w(n-1))$$

- Dystans skumulowany jest miarą wskazującą na podobieństwo obiektu do wzorca:

$$D_T = \min_{\{w(n)\}} \sum_{n=1}^{N} D(R(n),T(w(n)))$$

$$D_A(n,m) = D(n,m) + \min_{q \le m} D_A(n-1,q)$$

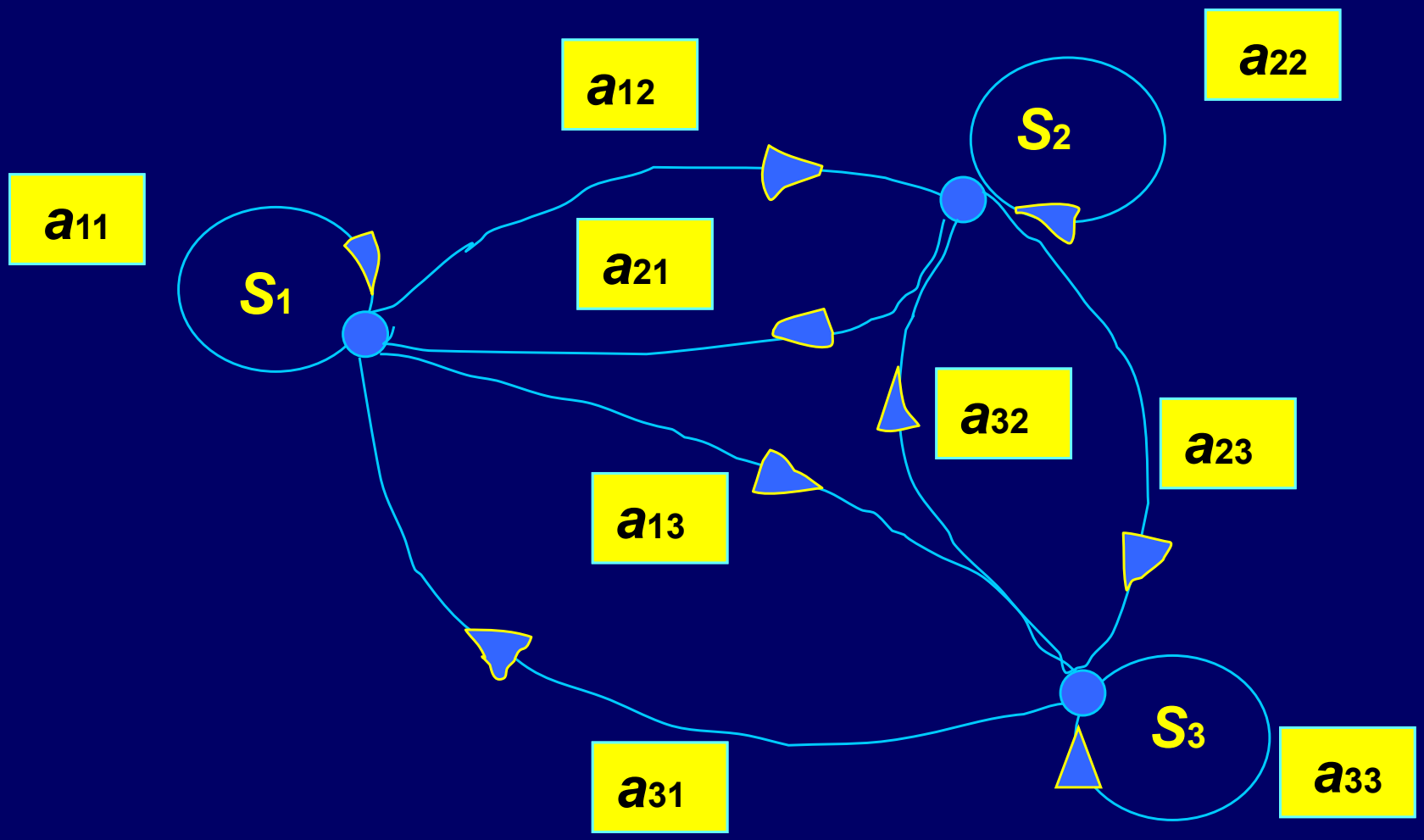
gdzie: $D_A(n,m)$ - minimalny dystans skumulowany względem punktu (n,m) siatki

HMM

- Dane słowo S_m w słowniku M możliwych słów jest reprezentowane ciągiem m zdarzeń O
- Każde słowo w słowniku jest opisane Łańcuchem Markowa (HMM), dla M słów $\Rightarrow$ M·HMM $\{L_1, L_2, ..., L_M\}$
- procedura dopasowania polega na obliczeniu sumarycznego prawdopodobieństwa (zdarzeń i przejść), że dany ciąg zdarzeń O został wygenerowany przez dany model L
- Prawdopodobieństwo to dane jest wzorem:

$$P_m = Pr(O|L_m)$$

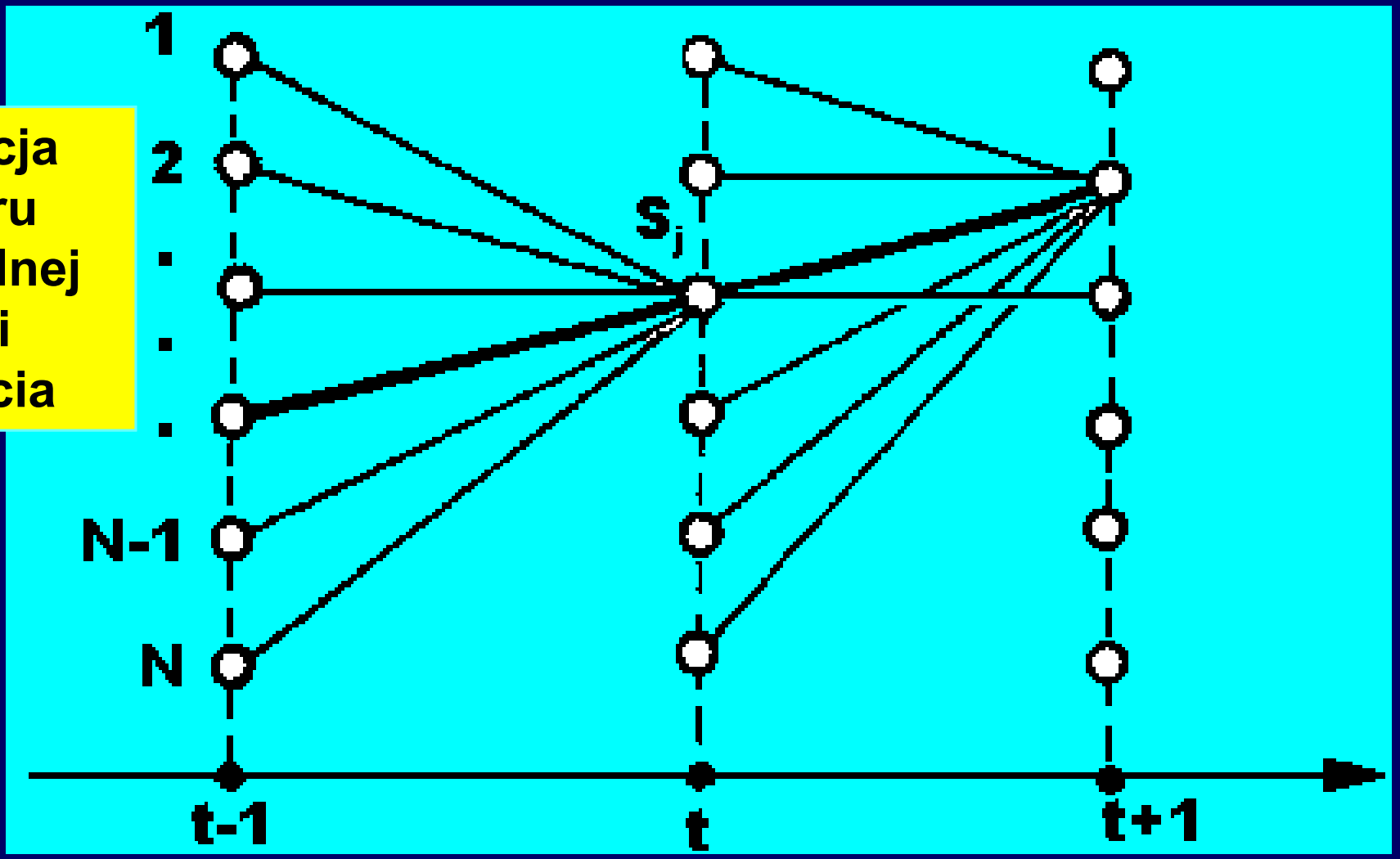
HMM



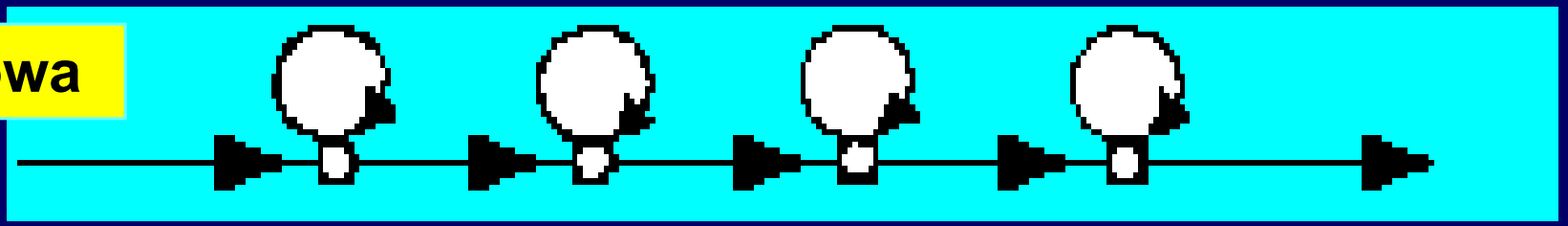
Ilustracja stanów i prawdopodobieństwa zdarzeń procesu Markowa

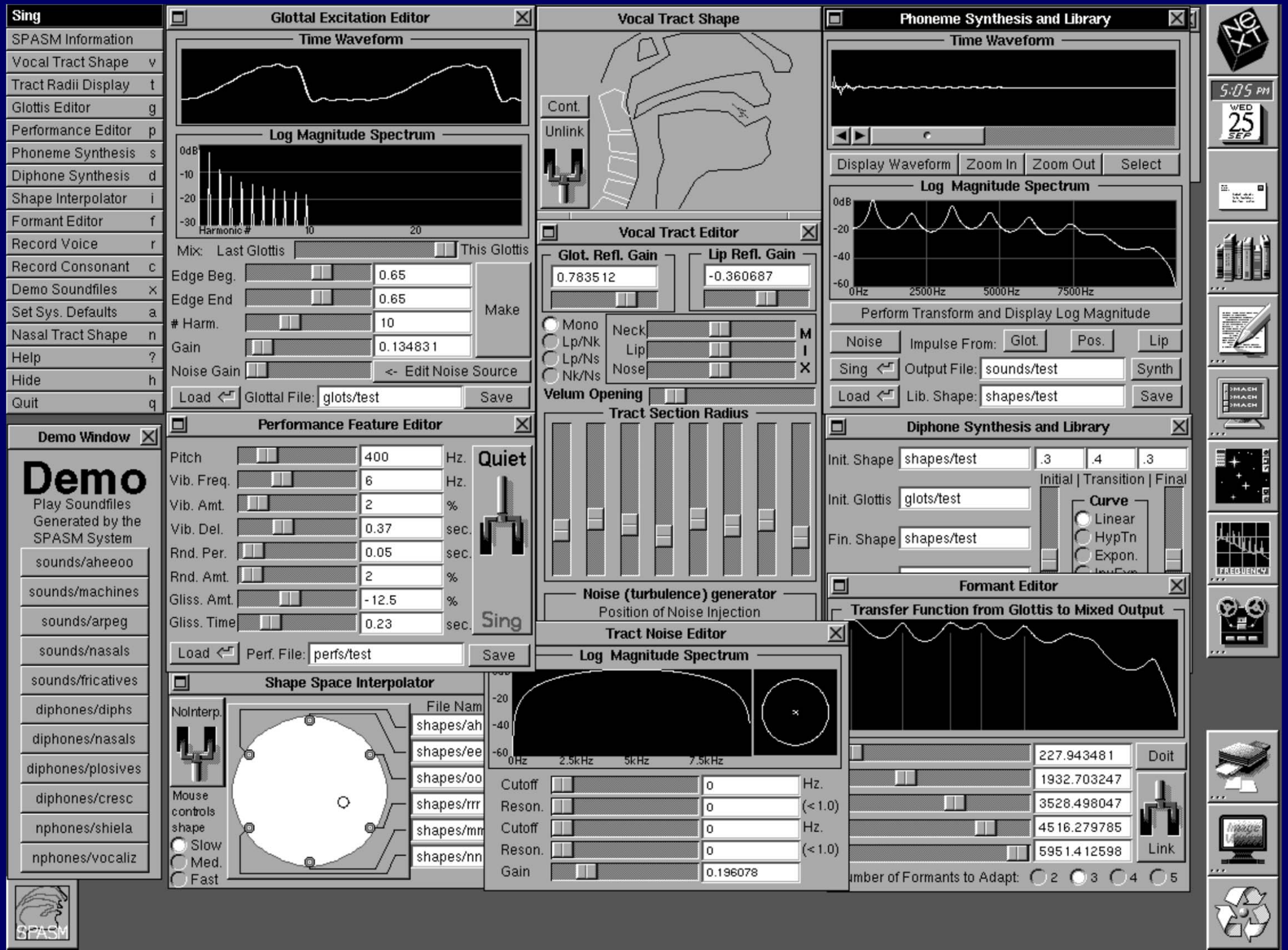
HMM

Ilustracja wyboru optymalnej drogi przejścia



Model słowa

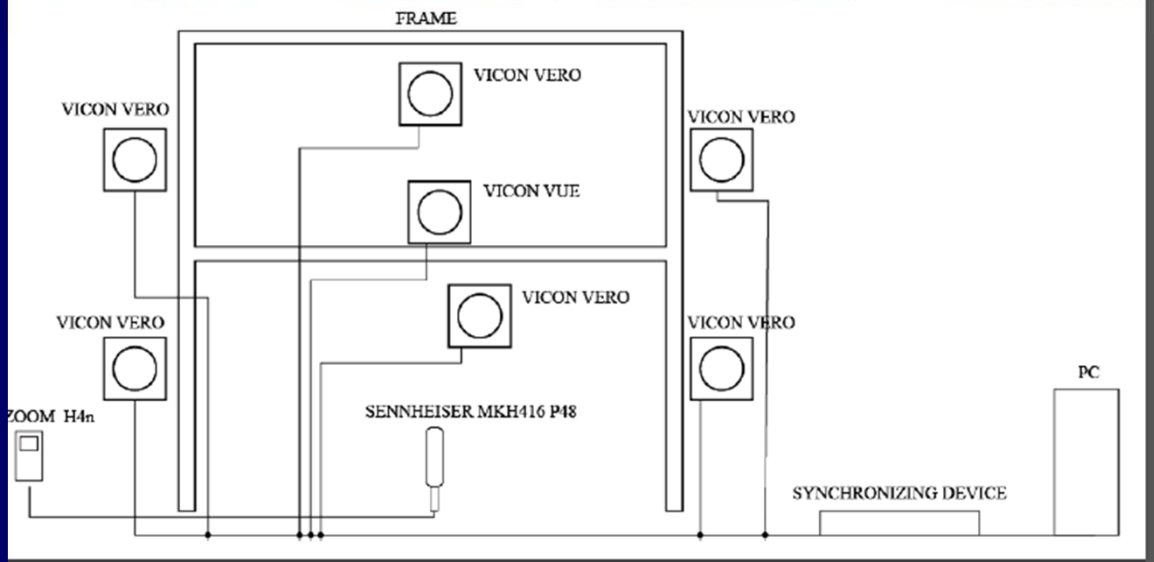


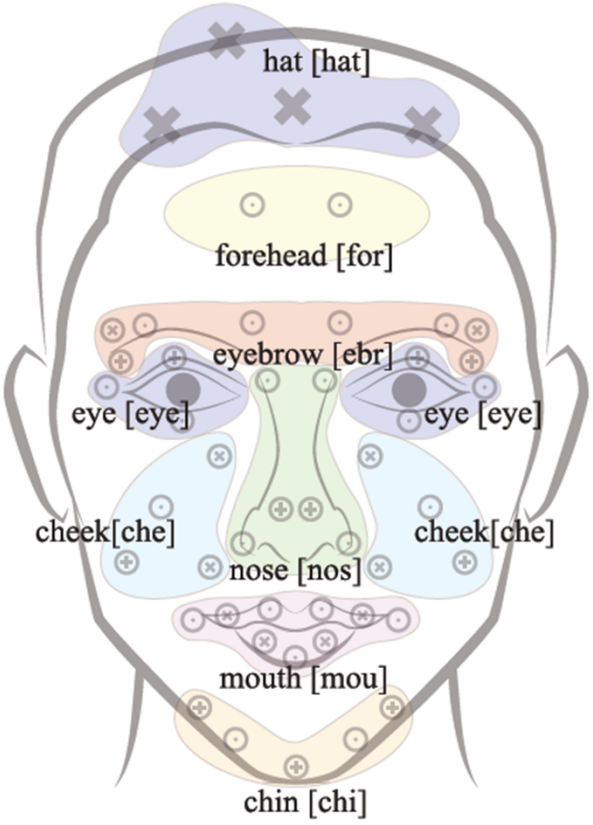

Sing
SPASM Information
Vocal Tract Shape v
Tract Radii Display t
Glottis Editor g
Performance Editor p
Phoneme Synthesis s
Diphone Synthesis d
Shape Interpolator i
Formant Editor f
Record Voice r
Record Consonant c
Demo Soundfiles x
Set Sys. Defaults a
Nasal Tract Shape n
Help ?
Hide h
Quit q
Glottal Excitation Editor
Time Waveform
Log Magnitude Spectrum
0dB
-10
-20
-30
Harmonic #
10
20
Mix: Last Glottis This Glottis
Edge Beg. 0.65
Edge End 0.65
Harm. 10
Gain 0.134831
Noise Gain
Make
<- Edit Noise Source
Load Glottal File: glots/test Save
Vocal Tract Shape
Cont.
Unlink
Vocal Tract Editor
Glot. Refl. Gain
0.783512
Lip Refl. Gain
-0.360687
Mono
Lp/Nk
Lp/Ns
Nk/Ns
Neck
Lip
Nose
M
I
X
Velum Opening
Tract Section Radius
Noise (turbulence) generator
Position of Noise Injection
Phoneme Synthesis and Library
Time Waveform
Display Waveform
Zoom In
Zoom Out
Select
Log Magnitude Spectrum
0dB
-20
-40
-60
0Hz
2500Hz
5000Hz
7500Hz
Perform Transform and Display Log Magnitude
Noise
Impulse From: Glot. Pos. Lip
Sing
Output File: sounds/test
Synth
Load
Lib. Shape: shapes/test
Save
Demo Window
Demo
Play Soundfiles Generated by the SPASM System
sounds/aheeoo
sounds/machines
sounds/arpeg
sounds/nasals
sounds/fricatives
diphones/diphs
diphones/nasals
diphones/plosives
diphones/cresc
nphones/shiela
nphones/vocaliz
Performance Feature Editor
Pitch 400 Hz.
Vib. Freq. 6 Hz.
Vib. Amt. 2 %
Vib. Del. 0.37 sec.
Rnd. Per. 0.05 sec.
Rnd. Amt. 2 %
Gliss. Amt. -12.5 %
Gliss. Time 0.23 sec.
Quiet
Sing
Load
Perf. File: perfs/test
Save
Diphone Synthesis and Library
Init. Shape shapes/test .3 .4 .3
Initial | Transition | Final
Init. Glottis glots/test
Curve
Linear
HypTn
Expon.
Fin. Shape shapes/test
Shape Space Interpolator
NoInterp.
Mouse controls shape
Slow
Med.
Fast
File Nam
shapes/ah
shapes/ee
shapes/oo
shapes/rrr
shapes/mm
shapes/nn
Tract Noise Editor
Log Magnitude Spectrum
-20
-40
-60
0Hz
2.5kHz
5kHz
7.5kHz
Cutoff 0 Hz.
Reson. 0 (< 1.0)
Cutoff 0 Hz.
Reson. 0 (< 1.0)
Gain 0.196078
Formant Editor
Transfer Function from Glottis to Mixed Output
227.943481
1932.703247
3528.498047
4516.279785
5951.412598
Doit
Link
Number of Formants to Adapt: 2 3 4 5
SPASM

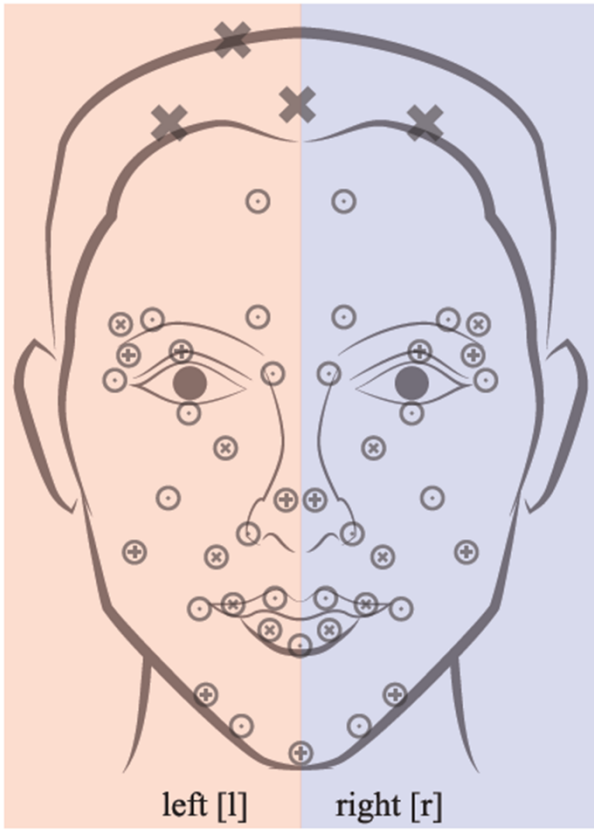
Audiowizualne rozpoznawanie mowy

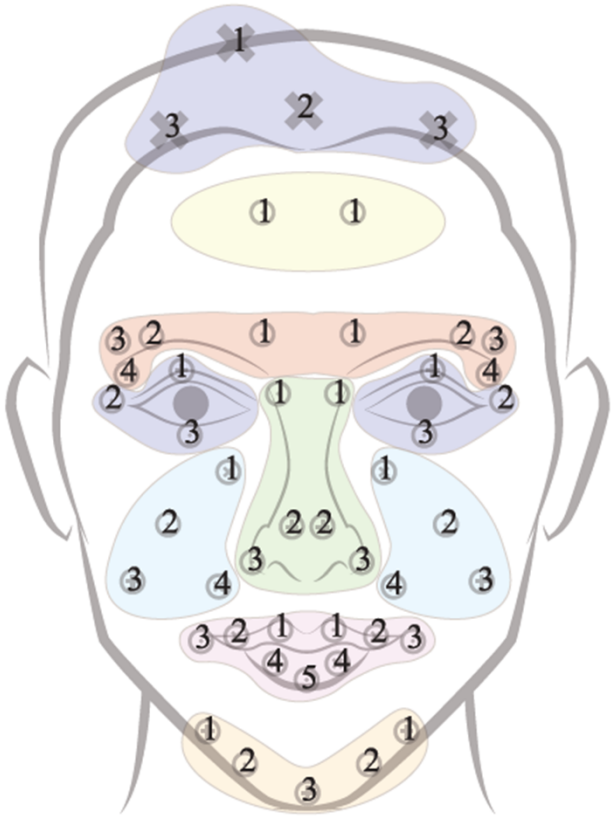
Autor:
Piotr Bratoszewski

WYKORZYSTANE URZĄDZENIA



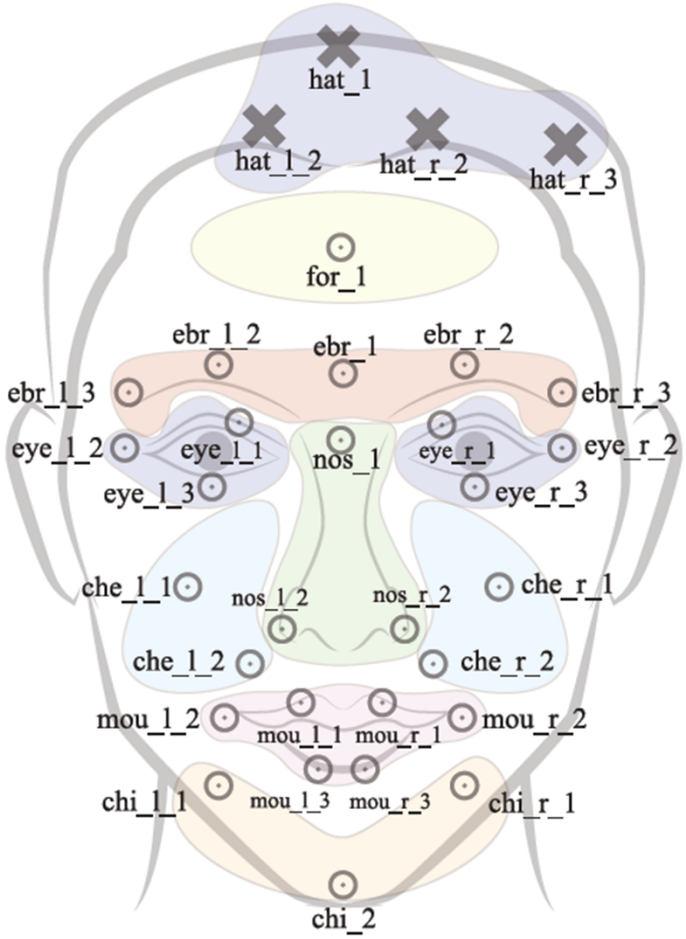


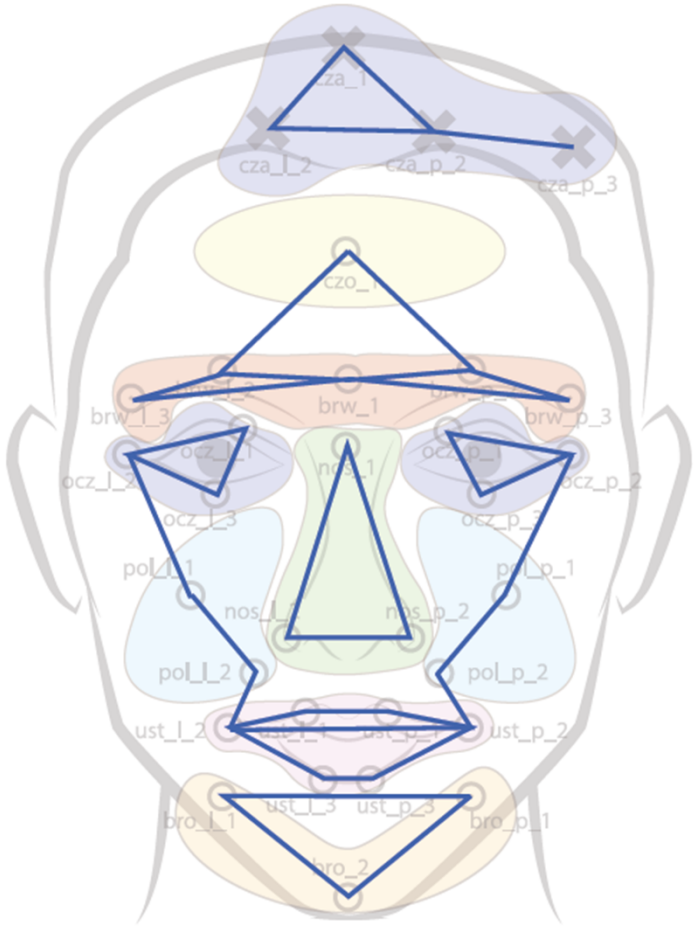




SPOSÓB NAZEWNICTWA MARKERÓW

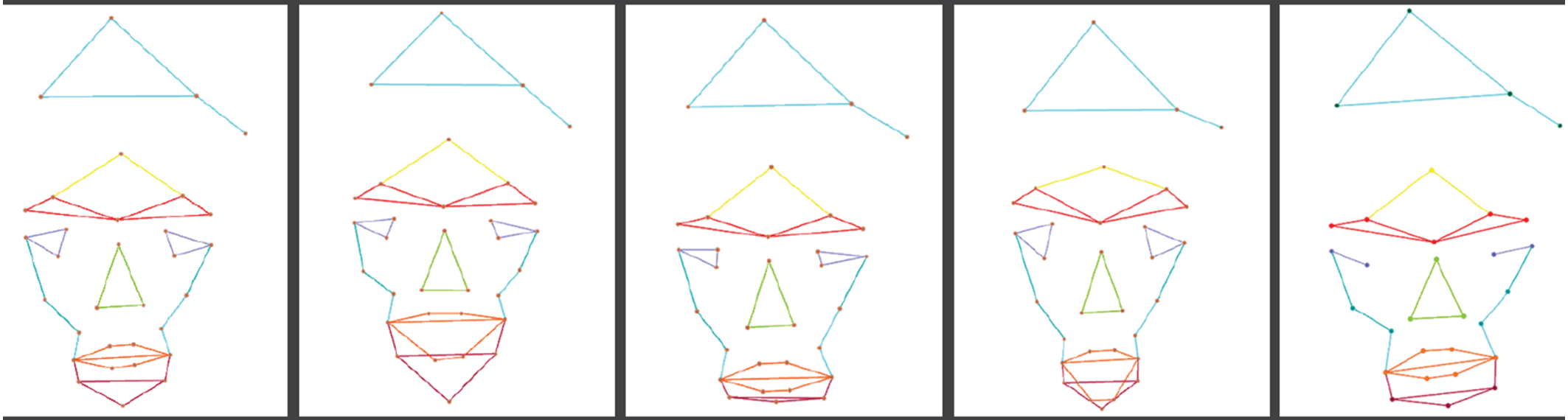
WIZUALIZACJA PUNKTÓW W PROGRAMIE BLADE





FINALNE UŁOŻENIE I NAZEWNICTWO

OSOBY BIORĄCE UDZIAŁ W NAGRANIACH



ZESTAW PRZYKŁADOWYCH EKSPRESJI OSOBY NR 3

MODALITYcorpus

Home License About Explore corpus

What is **Modality Corpus?**

The MODALITY corpus consists of over **30 hours** of multimodal recordings. The database contains high-resolution, high-framerate stereoscopic video streams and audio signals obtained from a microphone array and a laptop microphone. The corpus can be employed to develop an AVSR system, as every utterance was labelled. Recordings in noisy conditions can be used to test the robustness of speech recognition systems.

READ MORE >

License

MODALITYcorpus

Home License About Explore corpus

ALOFON
ælәfәun

Explore Alofon Corpus

Please select the type data you want to play and download

Words | Allophones

spi nac
spe ck
sp a
sp spar king
sp ot
sp ort

Select speaker

3

MODALITYcorpus

Home License About Explore corpus

ALOFON
ælәfәun

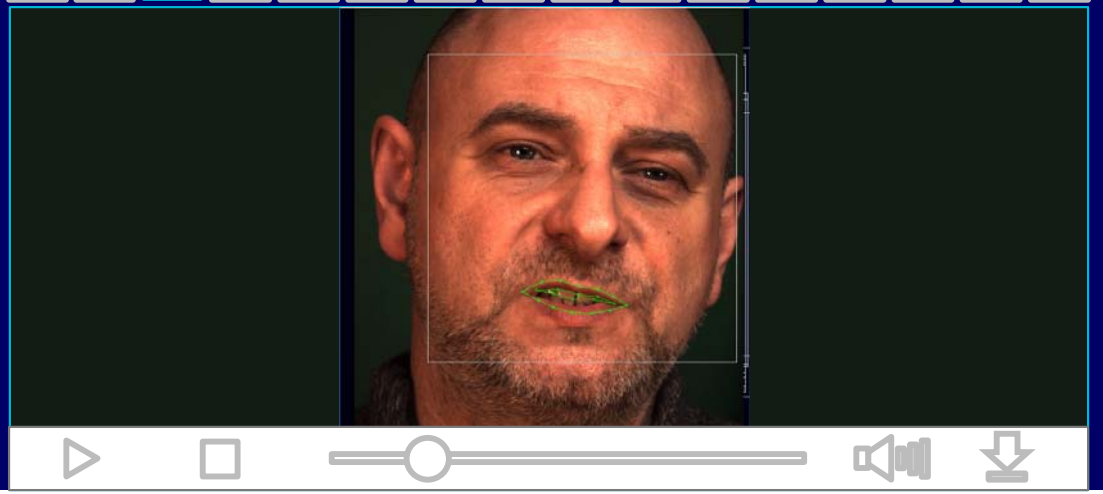
Explore Alofon Corpus

Please select the type data you want to play and download

Words | Allophones

Select speaker

3



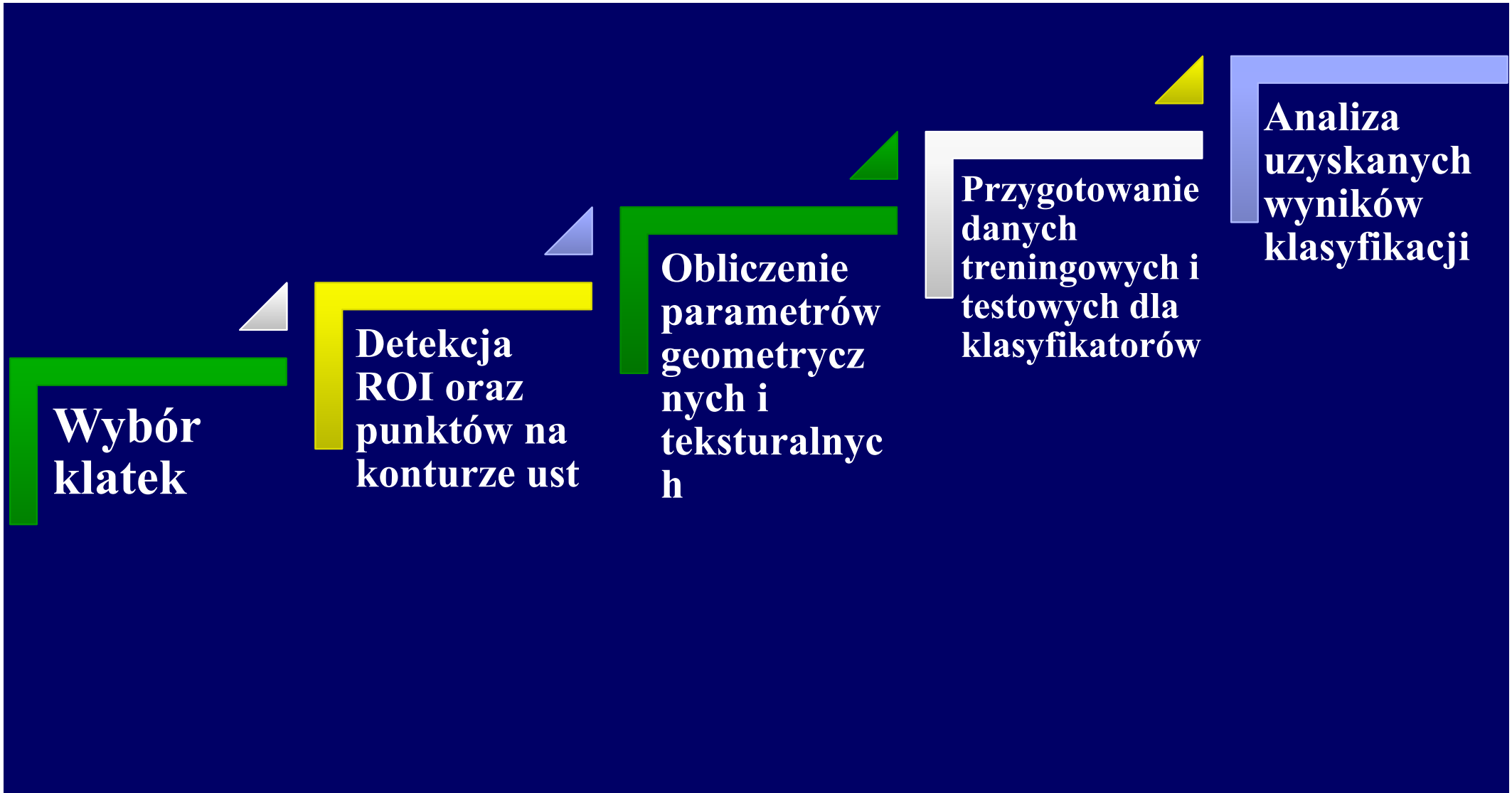
Grupy wizemów

W1	W2	W3	W4	W5	W6
WARGOWE	DZIĄSŁOWE	MIĘKKO-PODNIEBIENIOWE	ZĘBOWO-WARGOWE	PODNIEBIENIOWO-DZIĄSŁOWE	ZĘBOWE
					

Grupy wizemów c.d.

W7	W8	W9	W10	W11
ROZSZERZONE	OTWARTE-ROZSZERZONE	OTWARTE-NEUTRALNE	OTWARTE-ZAOKRĄGLONE	WYSUNIĘTO-ZAOKRĄGLONE
				

Proces postępowania z danymi

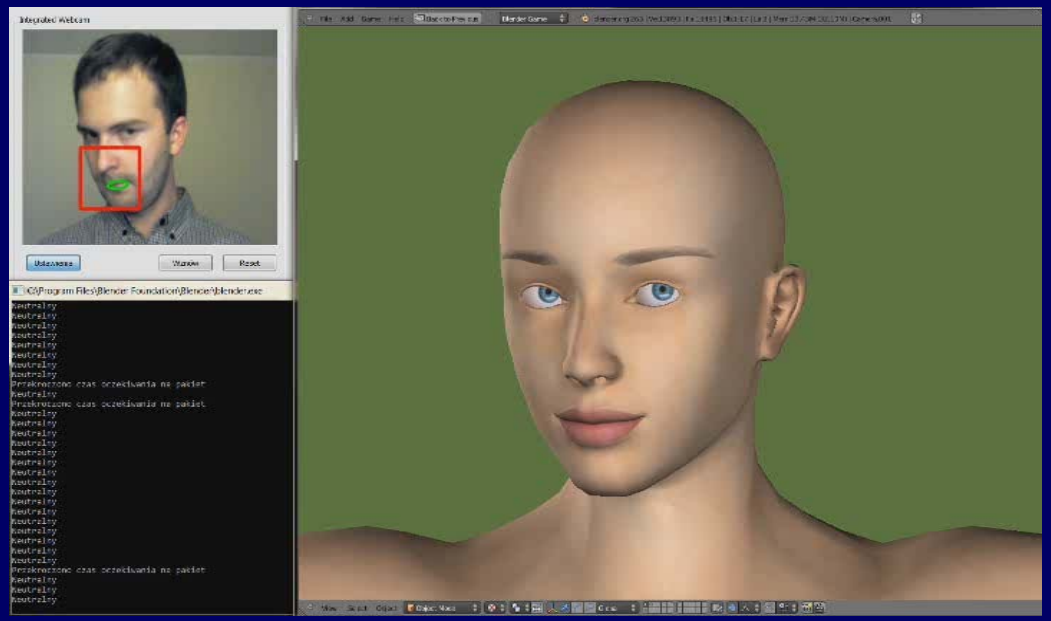


Parametry

Geometryczne	Teksturalne
• Odległościowe – 39 parametrów • Kątowe – 20 parametrów • Powierzchniowe – 8 parametrów	• Histogram obrazu ust w skali szarości – 32 parametry • Histogram obrazu ust HSV – 32 parametry • Histogram obrazu ust w skali szarości po equalizacji – 32 parametry • Histogram obrazu ust w skali szarości po filtrowaniu Clahe – 32 parametry • Wartości DCT dla obszaru ROI – 32 parametry

Wirtualny terapeuta – zastosowania w logopedii

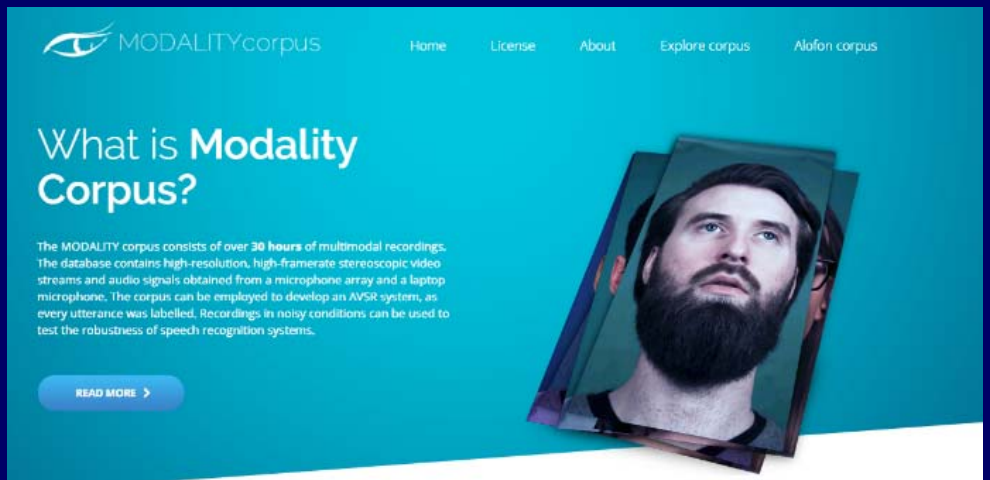
- Osadzona sztuczna sieć neuronowa (perceptron wielowarstwowy) wytrenowana na obrazach

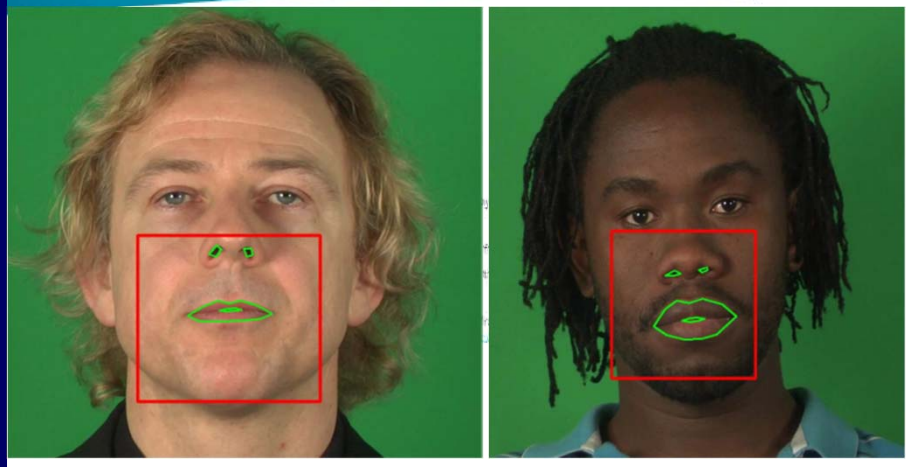


Agata.wmv

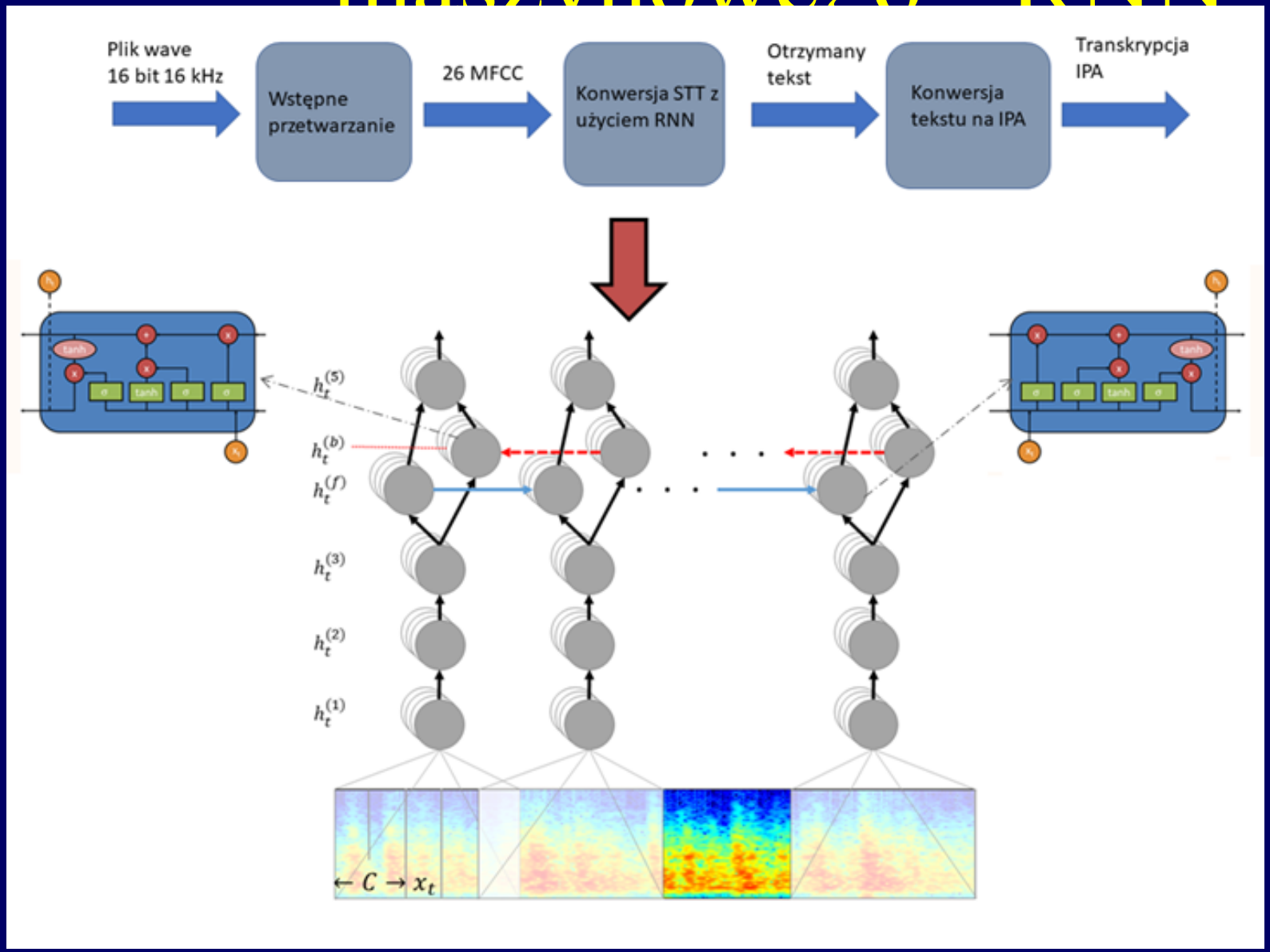
Audiowizualna analiza mowy

- **Korpus MODALITY** (31 godzin nagrań audio-video mówców) www.modality-corpus.org
- **Korpus ALOFON**
- (1,5h godziny nagrań audio, vide o i Facial Motion Capture)



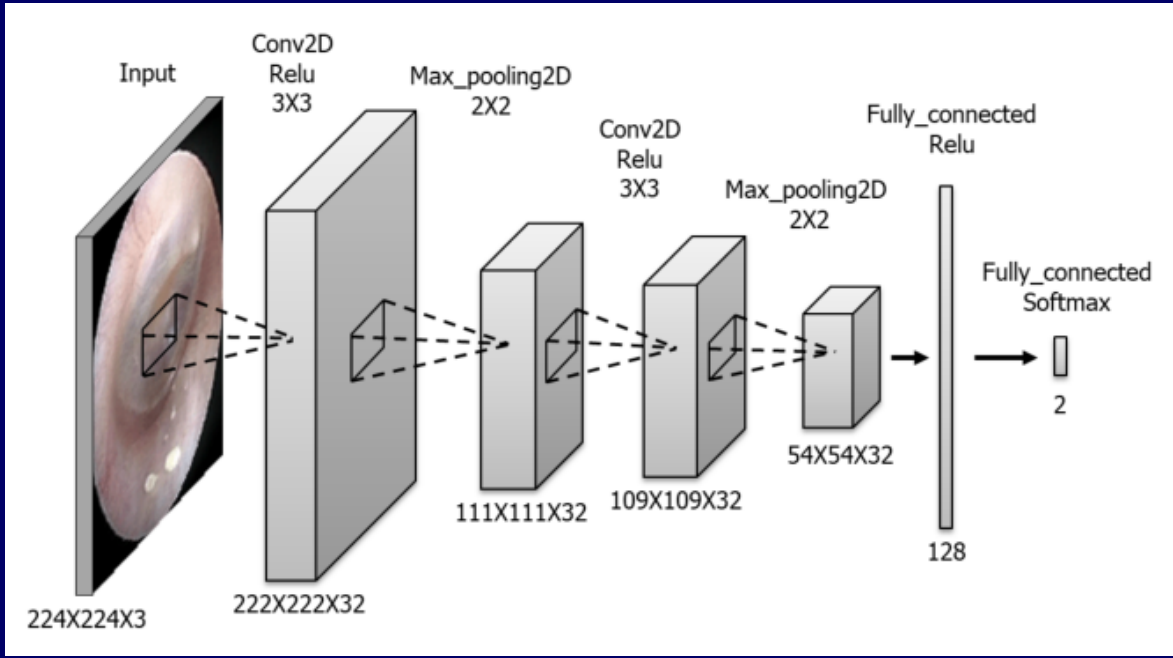


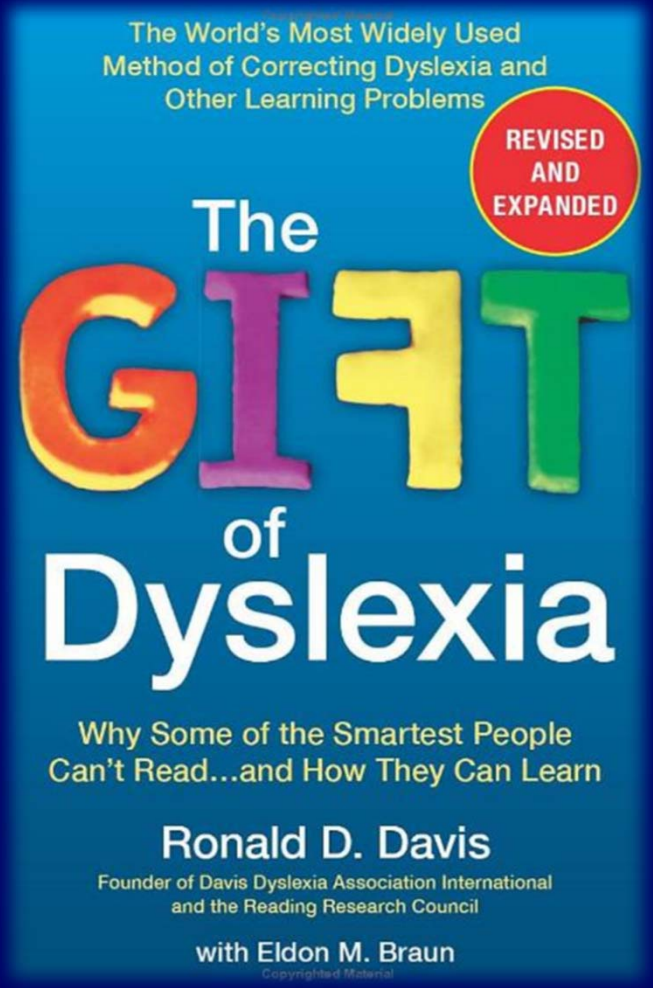
Wykorzystanie uczenia maszynowego - RNN



„Dar dysleksji” – mózg i głębokie uczenie

Davis D. Ronald Eldon M. Braun





Sieci neuronowe DenseNet

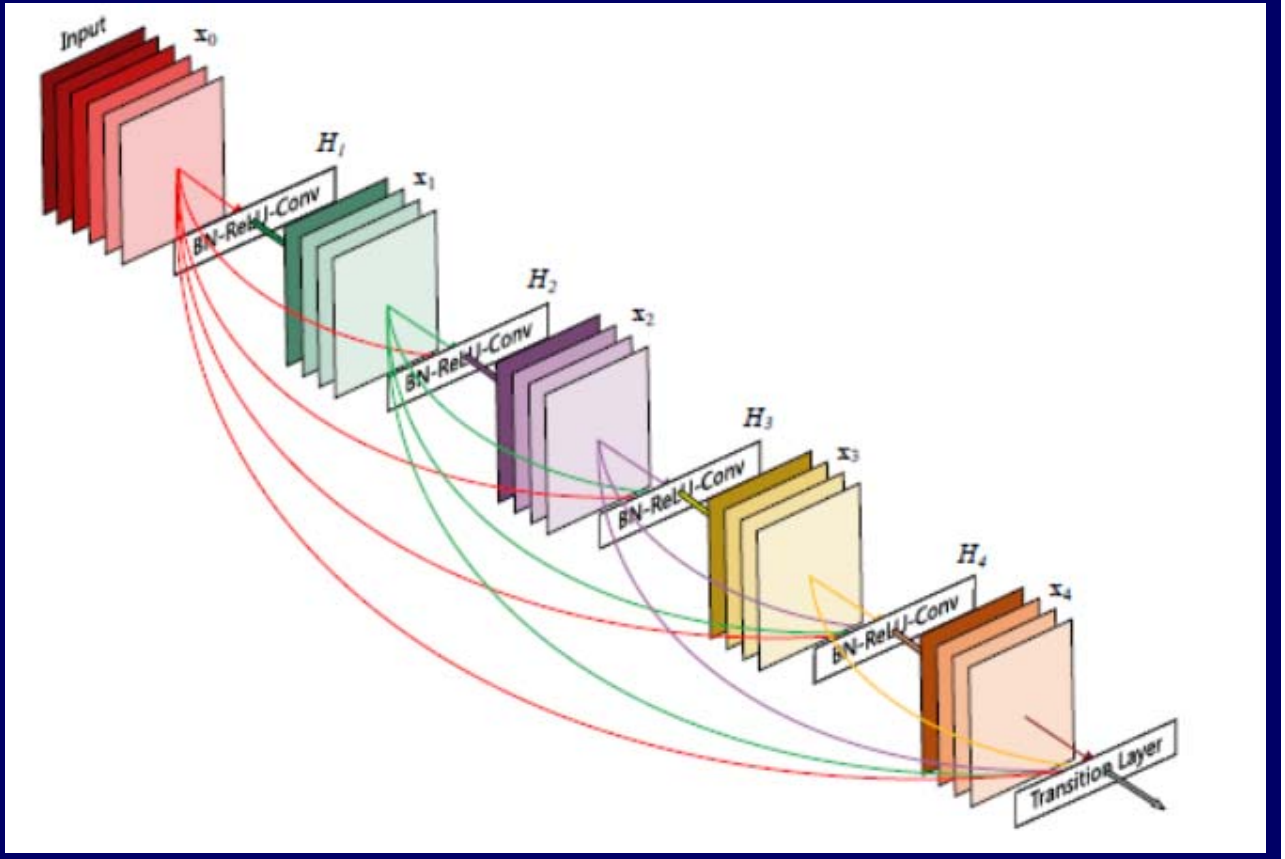
INZNAK

Prace badawcze (publikacje Cornell University)wykazały,
że sieci splotowe mogą być znacznie bardziej efektywne w treningu, jeśli zawierają krótsze połączenia między warstwami w pobliżu wejścia i tymi zbliżonymi do wyjścia. Dense Convolutional Network (DenseNet) łączy każdą warstwę do każdej innej warstwy z wyprzedzeniem.

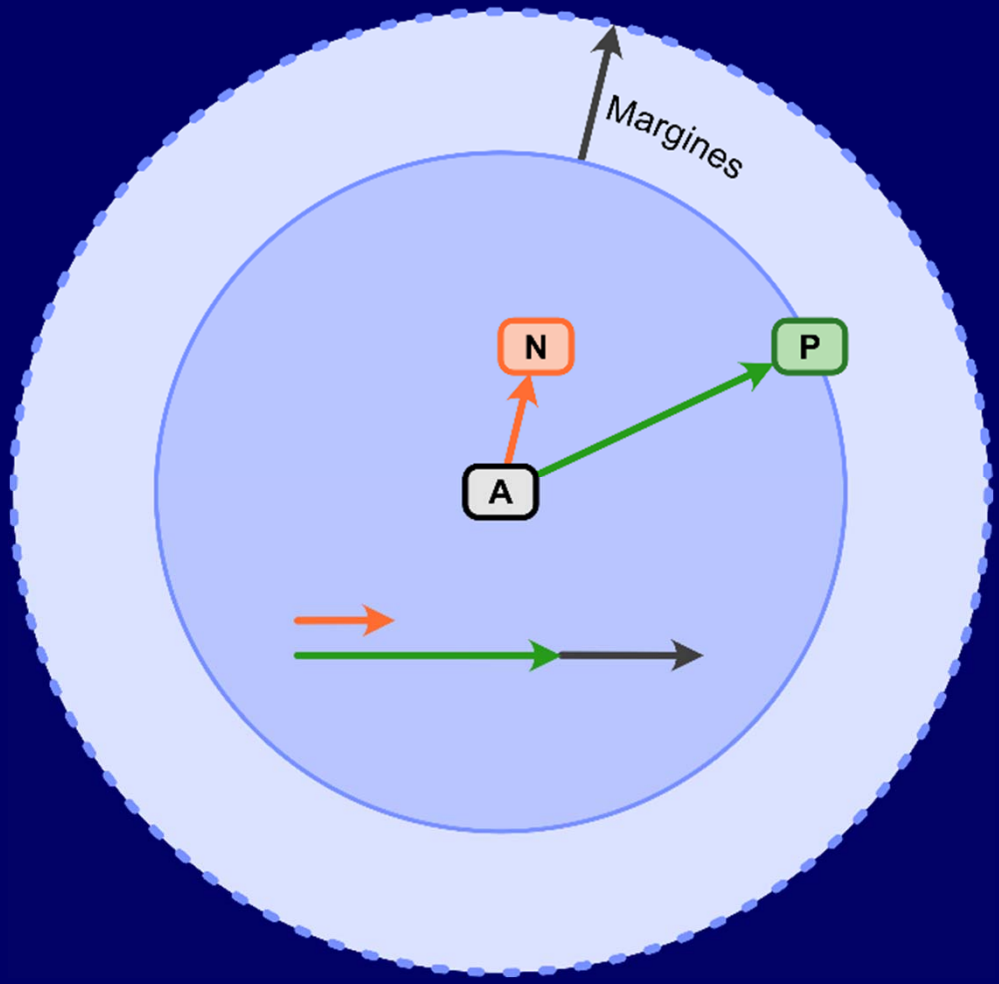
9	5	7	6	3	1	2	8	4
4	1	2	3	8	7	2	6	5	9
6	2	4	5	4	5	9	7	3	1
5	3	9	1	7	8	6	4	1	2
8	4	2	9	1	5	3	6	7
7	6	1	2	4	3	5	9	8
2	8	5	3	9	7	1	4	6
1	7	4	5	6	8	9	2	3
3	9	6	1	2	4	8	7	5

analogia do podstawowych operacji na obrazie

Kod i wstępnie wytrenowane modele są dostępne na stronie https://github.com/liuzhuang13/DenseNet.



Sieć neuronowa trenowana metodą Triplet Loss



Przykład trójki niespełniającej warunku odległości dla algorytmu triplet loss. Pomarańczowy wektor (AN) powinien być dłuższy od zielonego wektora (AP) o wartość marginesu.

Algorytm uczenia maszynowego, w którym bazowe (kotwiczące) dane wejściowe są porównywane z pozytywnymi (prawdziwymi) i negatywnymi (fałszywymi) danymi wejściowymi.

Zastosowania rozpoznawania mowy

- Programy i urządzenia przeznaczone dla osób niepełnosprawnych
- Sterowanie urządzeniami za pomocą głosu, np. telefonu komórkowego, komputera, inteligentnego domu, urządzeń samochodowych
- Nawigacja stroną internetową
- Gry edukacyjne
- Rozpoznawanie osób
- Pisanie tekstu
- Aplikacje multimedialne
- Zabawki dla dzieci
- Robotyka

Dziękuję za uwagę