Przetwarzanie sygnału mowy

Dr. Gražina Korvel
Vilnius University Institute of Mathematics and Informatics
Akademijos str. 4
Vilnius, Lithuania
grazina.korvel@mii.vu.lt

Uniwersytet Wileński

Państwowy uniwersytet w Wilnie, założony w 1579 przez króla Polski Stefana Batorego. Jest największą uczelnią w Litwie.

Dziedziniec Macieja Kazimierza Sarbiewskiego

Wielki Dziedziniec Uniwersytetu Wileńskiego i *Kościół św*. Jana Chrzciciela i *św*. Jana Ewangelisty

Plan wykładu

- Wytwarzanie sygnału mowy
- Techniki przetwarzania sygnału mowy
  - Synteza mowy
  - Rozpoznawanie mowy
  - Rozpoznawanie mówcy

Wytwarzanie sygnału 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.

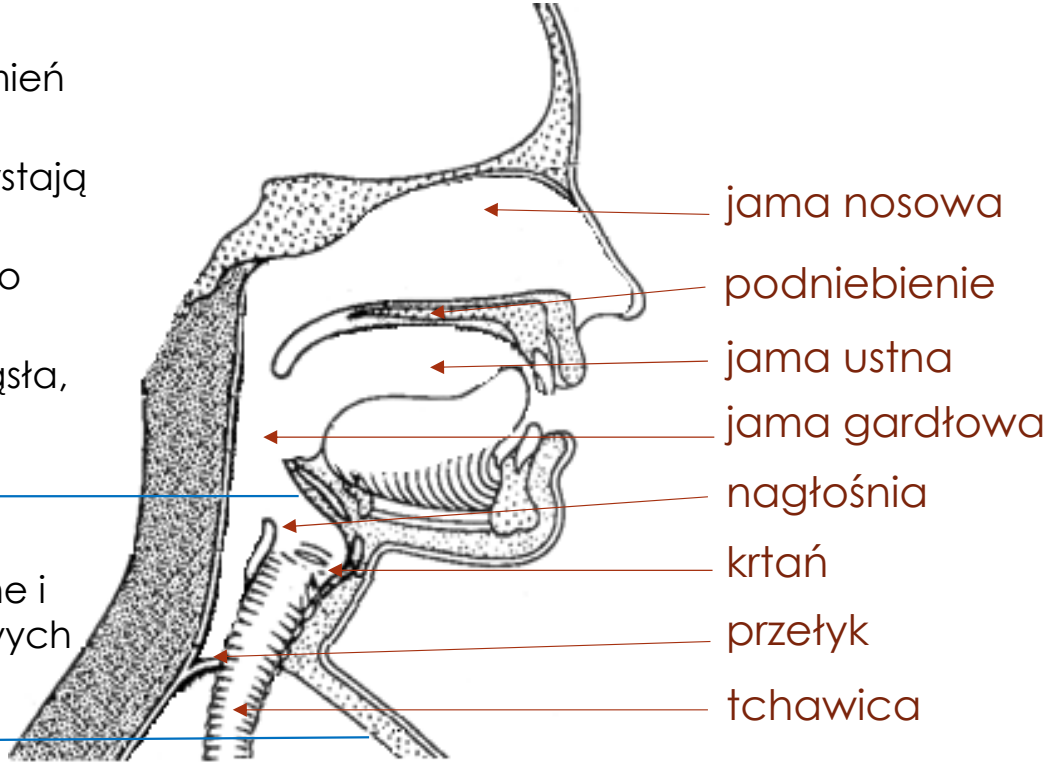
W artykulacji biorą też udział wargi, zęby, dziąsła, podniebienie twarde.

APARAT FONACYJNY

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

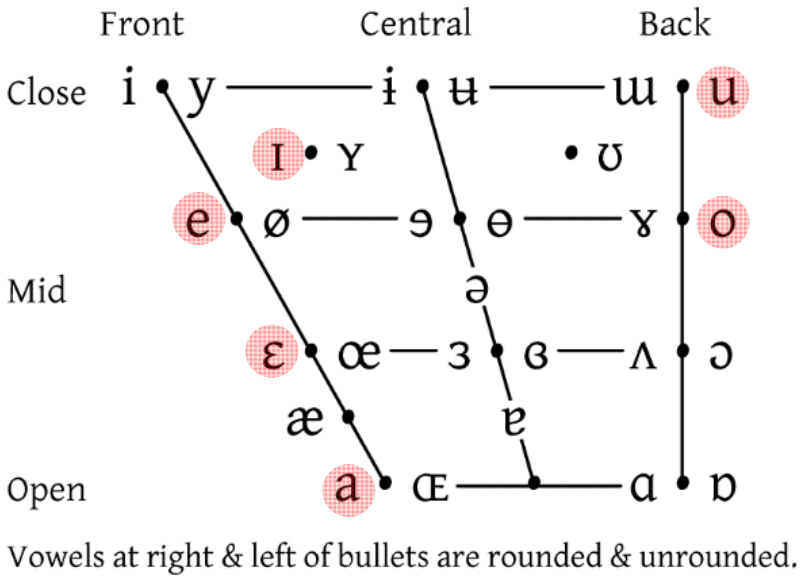
APARAT ODDECHOWY

Dostarczają energię, generującą falę dźwiękową



Międzynarodowy alfabet fonetyczny (angl. International Phonetic Alphabet)

Samogłoski



Vowels at right & left of bullets are rounded & unrounded.

Jest to standardowy zapis fonetyczny dla wszystkich języków świata.

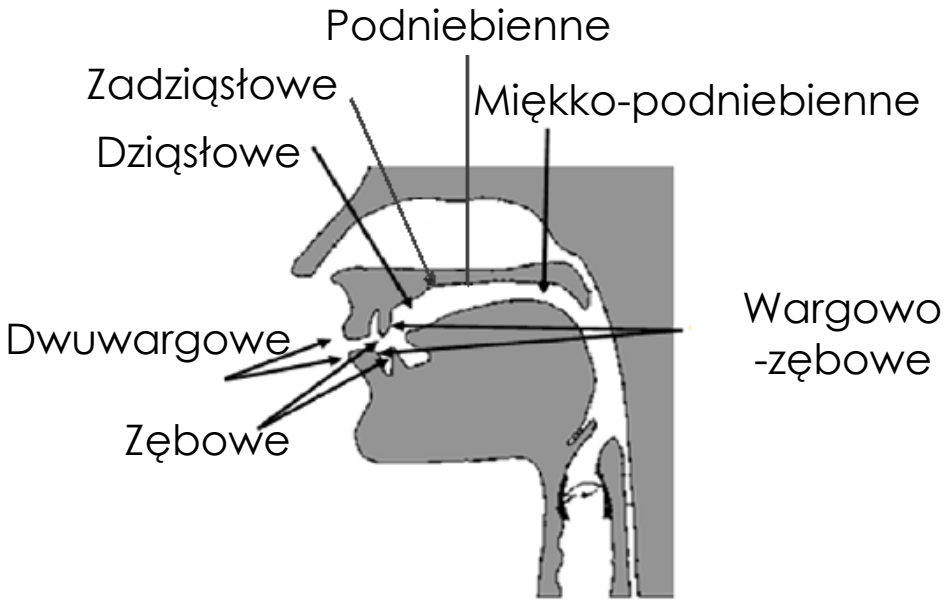
Alfabet IPA przyjęty w 1886 roku przez Międzynarodowe Towarzystwo Fonetyczne.

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)	p b		t d			ʈ ɖ	c ɟ	k g	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	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	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



Spółgłoski	IPA	Przykłady
Dwuwargowe	p, p' b, b' m, m'	**p**óvas, **p**etelìškė **b**rolis, la**b**iáu ã**m**atas, s**m**ẽgenys
Wargowo-zębowe	f, f'	**f**ãbrikas, **f**igūrá
Zębowe	t, t' d, d'	**t**ãkas, šal**t**ẽkšnis **d**árbas, liū**d**esỹs
Dziąsłowe	s, s' z, z, n, n' l, l'	**s**áulė, vaĩ**s**ius **z**ýlė, **z**irzė́ti **n**ãmas, **n**ėšti vá**l**sas, va**l**ià
Zadziąsłowe	ʃ, ʃ' ʒ, ʒ, r, r'	**š**akà, **š**iáudas **ž**vãkė, **ž**iógas **r**ãtas, k**r**iáušė
Podniebienne	j	á**i**das
Miękko-podniebienne	k, k' g, g' x, x' ɣ, ɣ'	**k**ãtinas, **k**iaũlė **g**añdras, **g**érvė **ch**òras, **ch**èmija **h**armònija, **h**iacìntas

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 dostarcza się 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 šakà, šiáudas žvãkė, žiógas chòras, chèmija harmònija, hiacìntas

Techniki przetwarzania sygnału mowy

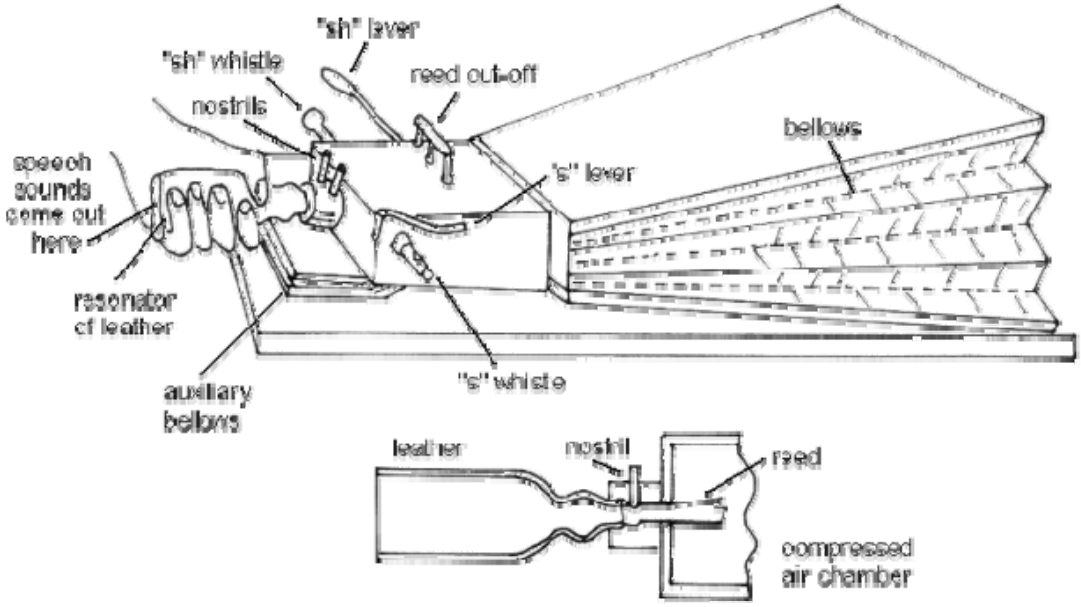
- Rozpoznawanie mowy
- Rozpoznawanie mówcy
- Synteza mowy
- Poprawa jakości sygnału
- Kodowanie mowy

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„)



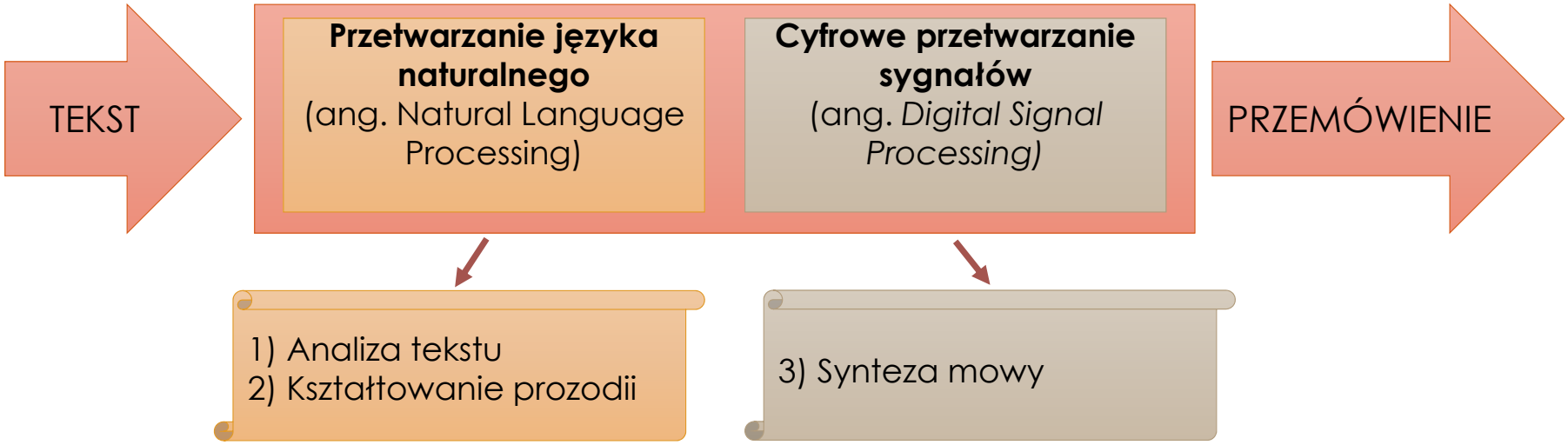
Pierwszy mechaniczny syntezator (von Kempelen, 1791)

Synteza mowy (ang. Text-To-Speech)

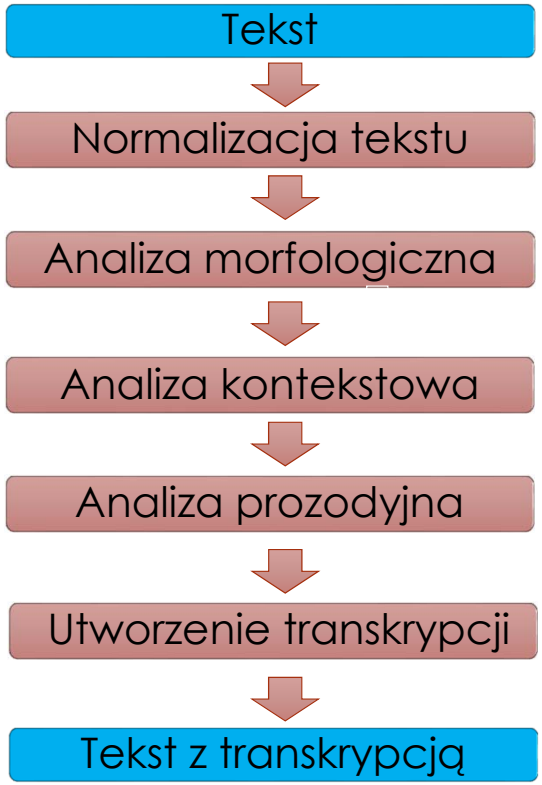
Zmiana tekstu na sygnał akustyczny

Podstawowe cele:

- Zrozumiałość wypowiedzi
- Naturalny dźwięk



Przetwarzanie języka naturalnego
Analiza tekstu



Cel analizy:

Przekształcenie tekstu na zapis fonetyczny

Normalizacja tekstu

Zamiana znaków nieliterowych i skrótów na ciąg fonemów.

Proces normalizacji obejmuje:

- zmianę liter na małe lub wielkie
- rozwinięcie skrótów, akronimów
- usunięcie znaków interpunkcyjnych i diaktrycznych

Przykłady:

- 10 $-> dziesięć dolarów
- rys. 6. -> rysunek szósty

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 prozodyjna

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")

PROZODIA

AKCENT
Wyróżnienie pewnych sylab w wyrazach

INTONACJA
Wyróżnienie pewnych wyrazów i zdań

- Podwyższenie lub obniżenie tonu
- Zwiększenie lub zmniejszenie intensywności amplitudy
- Wydłużenie lub skrócenie czasu trwania głoski/wyrazu

Rodzaje syntezy mowy

- **Metoda formantowa**

  Odwzorowanie widma sygnału mowy

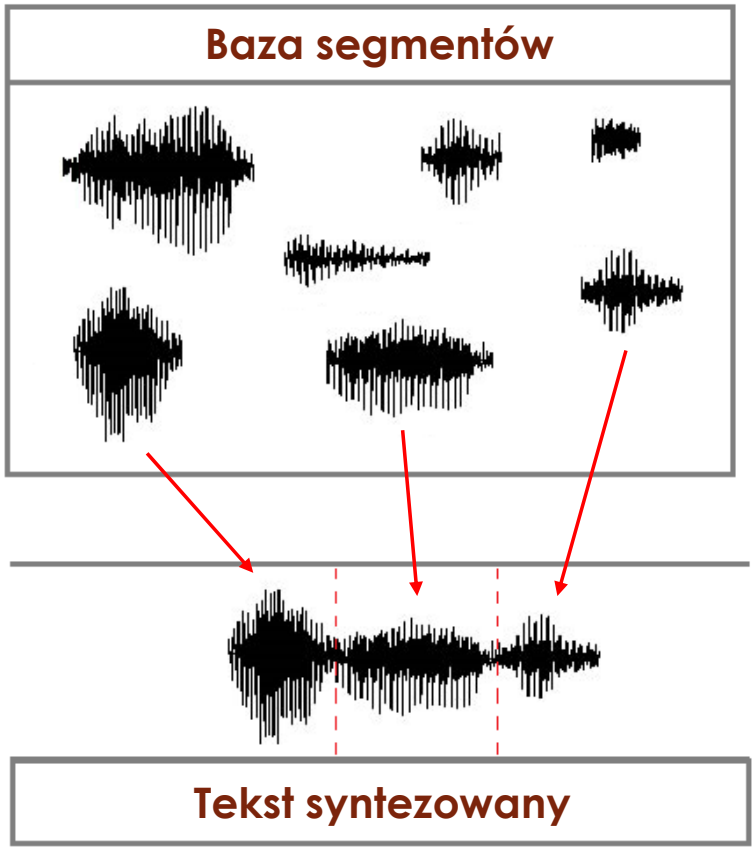
- **Metoda artykulacyjna**

  Fizyczne odwzorowanie mechanizmów wytwarzania mowy

- **Metoda konkatenacyjna**

  Wykorzystanie nagranych próbek sygnału mowy

Konkatenacyjna synteza mody



Łączenie wypowiedzi z mniejszych jednostek nagranych przez lektora

Wykorzystywane jednostki:

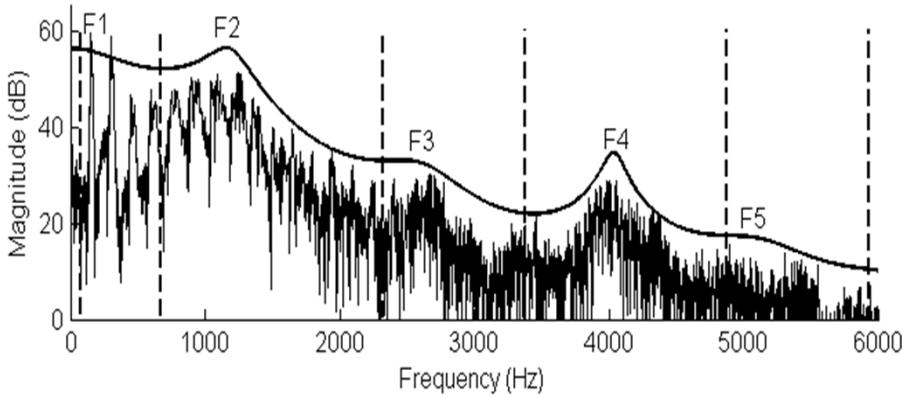
- fonem (głoska)
- difony
- trifony
- sylaby
- całe wyrazy

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.

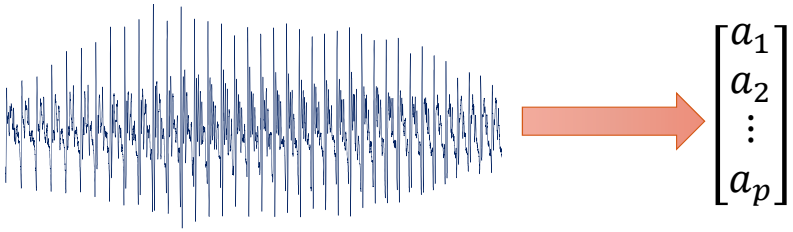
Artykulacyjna synteza mowy

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)

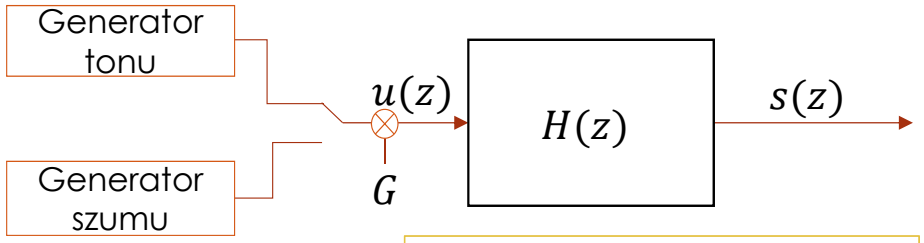
Najczęściej używa się kodowania predykcyjnego (*Linear Predictive Coding*).

1. Obliczanie *charakterystyki traktu głosowego*



Analiza LPC

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

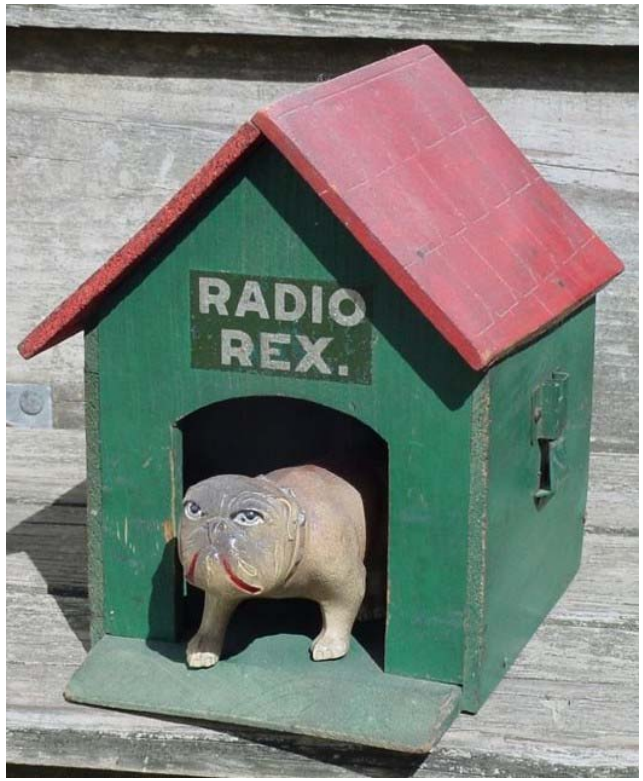


$$H(z) = G\frac{1}{1-\sum_{k=1}^{p} a_k \cdot z^{-k}}$$

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



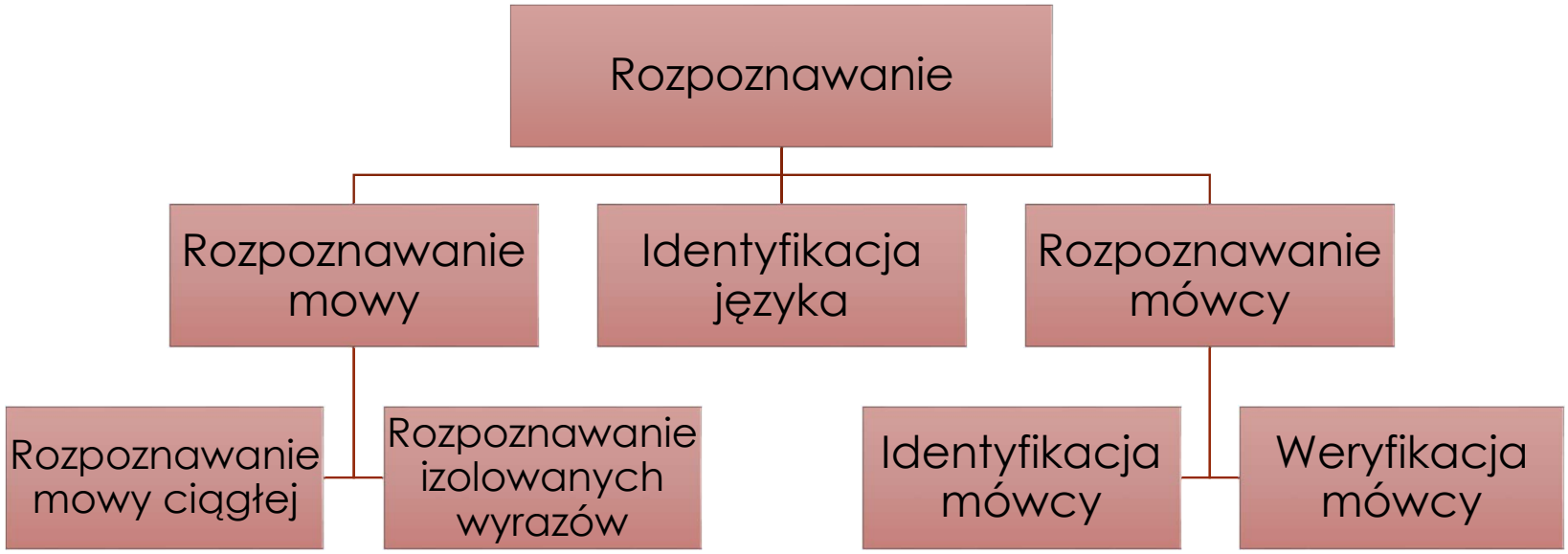
Zabawka Radio Rex powstała w 1920 roku

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



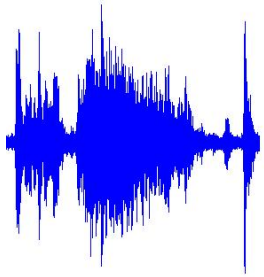
System może być zależny i niezależny od mówcy

Wielkość słownika

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

Obecnie system jest 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

Tekst

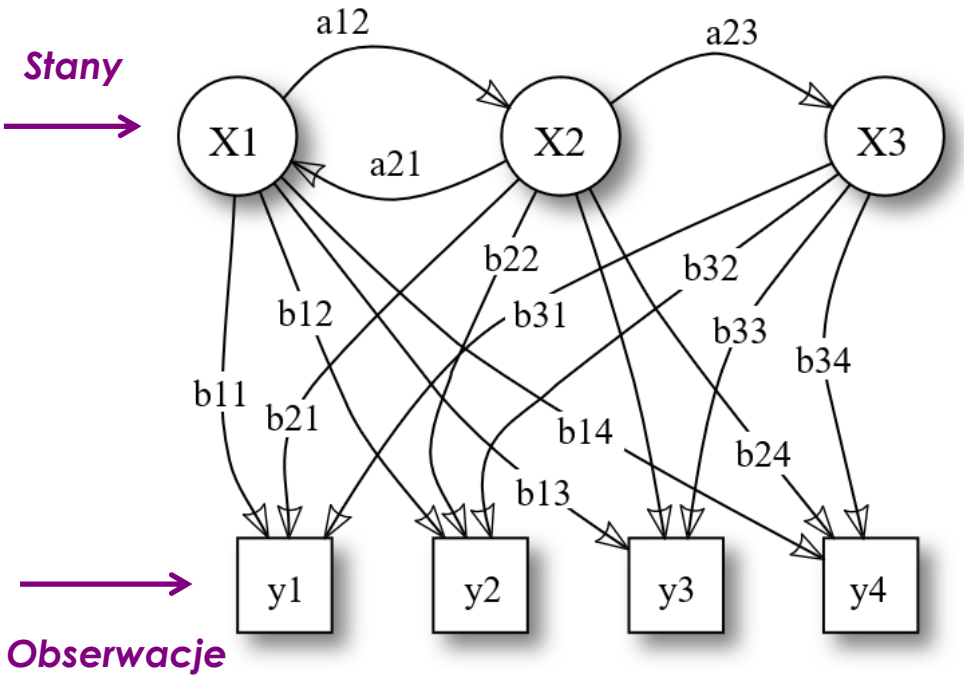
$P(w_i)$

Modele językowe (n-gramy)

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 $P(Y|w_i)$ sprowadza się do obliczenia sumarycznego prawdopodobieństwa (zdarzeń i przejść).

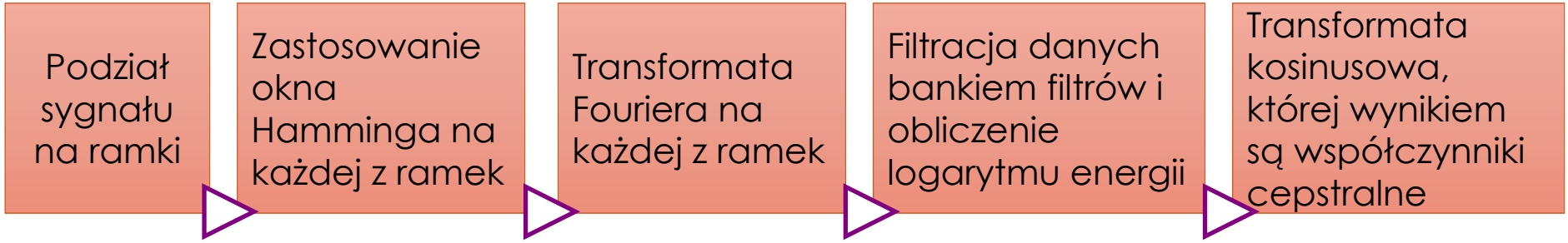
W ukrytym modelu Markowa stan nie jest widoczny, jednak wyjście zależne od niego jest znane.

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

Ekstrakcja parametrów - metody cepstralne

ang. Mel Frequency Cepstral Coefficient (MFCC)

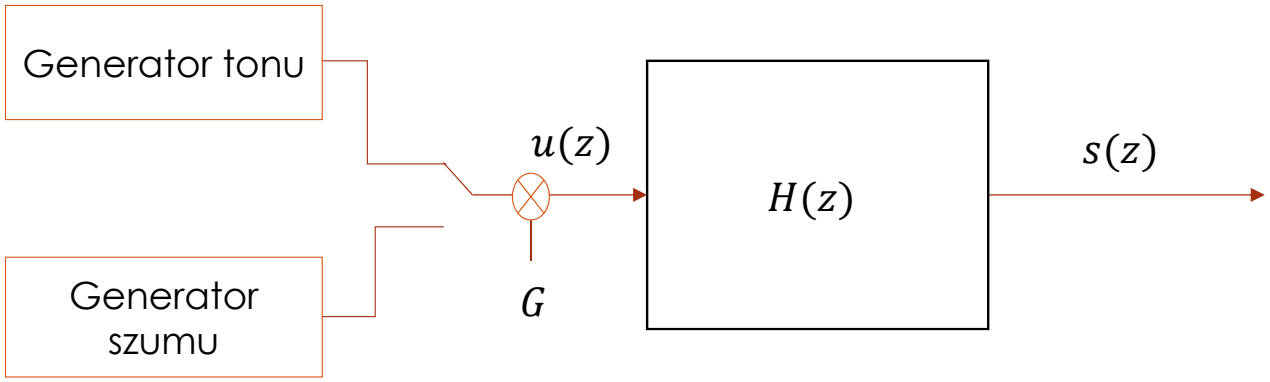
- Cepstrum - to transformata Fouriera logarytmu widma $\hat{X}(T) = F[\ln(X(f))]$.
- Skala cepstrum odpowiada dziedzinie czasu
- Współczynniki cepstralne niosą informacje o trakcie głosowym i o tonie krtaniowym
- Skala melowa, określająca subiektywny odbiór wysokości dźwięku przez ludzkie ucho względem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



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

Rozpoznawanie mówcy

Weryfikacja mówcy
Potwierdzenie deklarowanej przez mówcę tożsamości

- mówca współpracuje
- treść wypowiedzi znana
- sprawdzenie jednego wzorca

Identyfikacja mówcy
Wyznaczenie, który z mówców się wypowiada

- mówca może nie współpracować
- treść wypowiedzi nieznana
- obowiązkowe jest sprawdzenie wielu wzorców

Możliwe błędy podczas weryfikacji:

odrzucenie uprawnionego mówcy
zaakceptowanie nieuprawnionego mówcy

Możliwe błędy podczas identyfikacji:

błędna identyfikacja mówcy

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

Podsumowanie

Tematyka komputerowego przetwarzania sygnału mowy obejmuje niżej wymienione dziedziny:

- Cyfrowe przetwarzanie sygnału
- Przetwarzanie języka naturalnego
- Podstawy akustyki
- Informatykę i matematykę

Dziękuję za uwagę