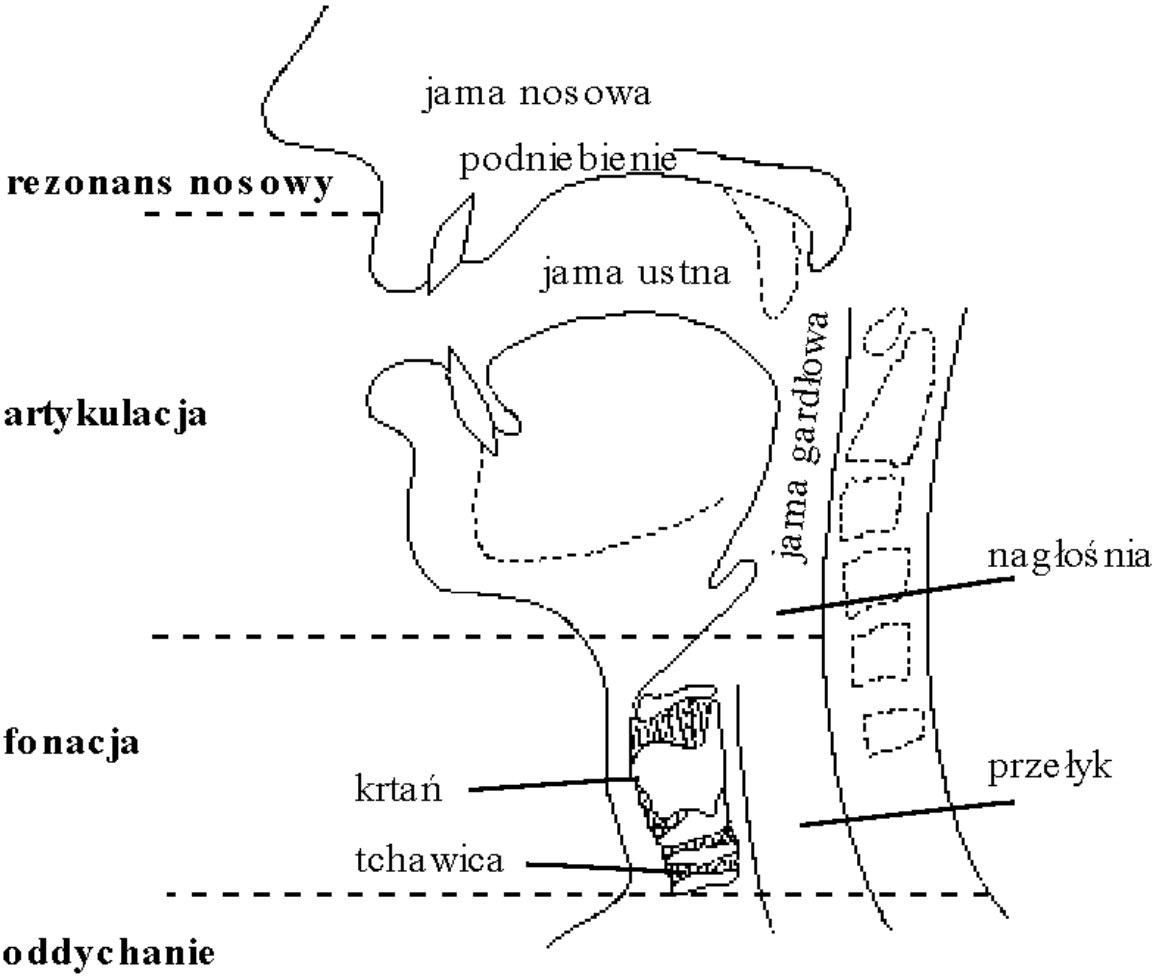
Podstawowe wiadomości na temat sygnału mowy i traktu głosowego

Artykulacja - praca organów mowy (wiązadeł głosowych, języka, jamy ustnej, i nosowej) potrzebna do wytworzenia dźwięków mowy.
Fonem - minimalny segment dźwiękowy mowy, który może odróżniać znaczenie, lub inaczej klasa dźwięków mowy danego języka o różnicach wynikających wyłącznie z charakteru indywidualnej wymowy lub kontekstu.
Alofon - wariant fonemu odróżniający się od innego alofonu cechami fonetycznymi a nie funkcją.
Diafon - przejście międzyfonemowe (inaczej difon. tranzem)
Mikrofonem - jednostka sygnału mowy o stałej długości czasowej (ok. 20-40 ms).
Formant - obszar koncentracji energii w widmie danego dźwięku mowy
lub inaczej: taki zakres widma, którego obwiednia zawiera maksimum.
Cechy dystynktywne - cechy pozwalające na rozróżnienie.
Ekstrakcja parametrów - procedura wydzielania z sygnału cech reprezentowanych przez wartości liczbowe (jest to element analizy sygnałów).
Redundancja - nadmiarowość w odniesieniu do informacji.
Logatomy - (ang. nonsense sylables) - sylaby służące do badania wyrazistości mowy w testach odsłuchowych.
HMM - (skrót od Hidden Markov Model) ukryty model Markowa używany w algorytmach do rozpoznawania mowy.
Wokodery - urządzenia służące do ograniczania objętości informacyjnej sygnału mowy metodą ekstracji parametrów i następnie po przesłaniu parametrów przez kanał telekomunikacyjny dokonujące resyntezy tego sygnału.



Narządy mowy w przekroju

Cechy mowy:
semantyczne - związane z treścią wypowiedzi
osobnicze - pozwalające rozpoznać osobę mówiącą
emocjonalne - pozwalające rozpoznać emocje osoby mówiącej; także stan zdrowia lub status społeczny
prozodyczne - odnoszące się do akcentu, głośności, intonacji, długości dźwięków i pauz

Złożoność analizy sygnału mowy:
- zakres dynamiki
- rozdzielczość częstotliwościowa i czasowa
- uwzględnienie czułości narządu słuchu
- możliwość uczenia się i dostosowywania do zmiennych warunków (np. efekt "coctail party")

Zakresy częstotliwości podstawowej tonu krtaniowego dla głosek dźwięcznych:
bas 80-320 Hz
baryton 100-400 Hz
tenor 120-480 Hz
alt 160-640 Hz
mezzosopran 200-800 Hz
sopran 240-960 Hz

Analogie elektryczno-akustyczne:

prąd <-> prędkość objętościowa U:

$$U=v \cdot A$$

v - prędkość liniowa drgań cząstek środowiska
A - pole powierzchni przekroju poprzecznego układu akustycznego

definicja ogólna:

impedancja akustyczna:

$$Z_a=p/U$$

p - ciśnienie akustyczne

W dziedzinie czasu sygnał mowy można opisać jako splot:

$$p(t)=e(t)*m(t)$$

$e(t)$ – sygnał pobudzenia
$m(t)$ – odpowiedź impulsowa układu biernych efektorów artykulacyjnych (traktu głosowego)

W dziedzinie zespolonej (transformacja Laplace'a) sygnał mowy można opisać:

$$p(s)=E(s) \cdot M(s)$$

$E(s)$ - pobudzenie
$M(s)$ – transformata Laplace'a odpowiedzi impulsowej układu biernych efektorów artykulacyjnych (traktu głosowego)
$s=\sigma+j\omega$ - częstotliwość zespolona
σ - tlumienie, ω - pulsacja

na okręgu jednostkowym (transformacja Fouriera)

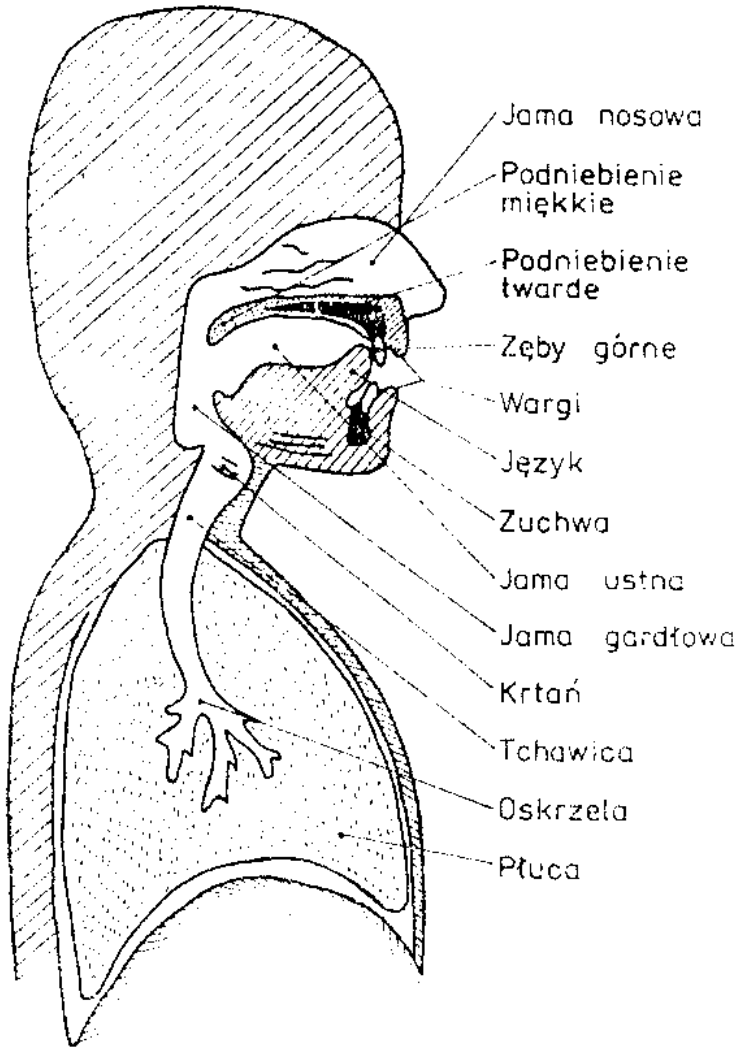
$$p(j\omega)=E(j\omega) \cdot M(j\omega)$$

lub para równań:

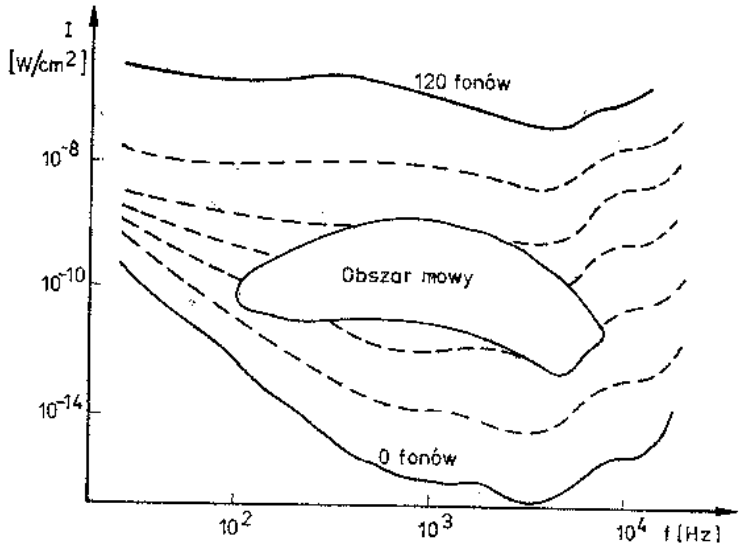
$|p(f)|=|E(f)|*|M(f)|$ - amplitudowe

$\phi[p(f)]= \phi[E(f)]+ \phi[M(f)]$ - fazowe

zalezności fazowe jednak nie mają wpływu na percepcję mowy

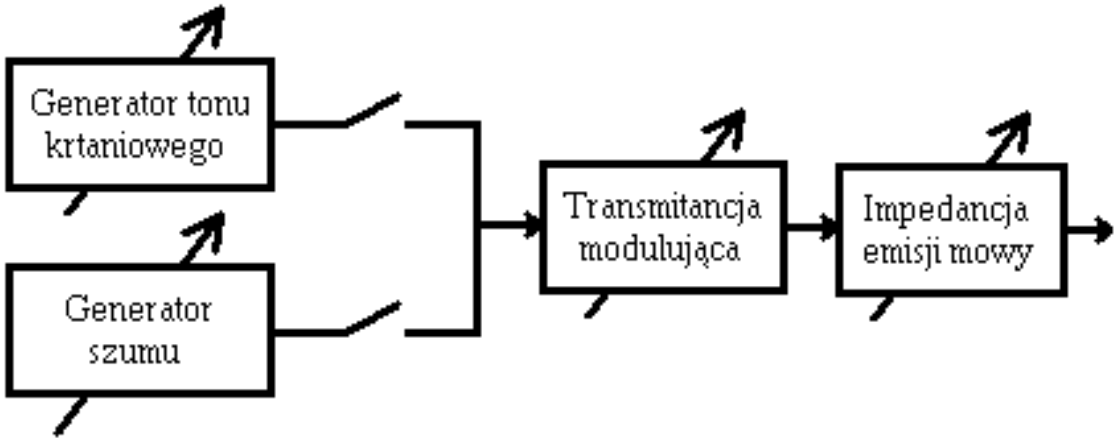


Uproszczony schemat traktu głosowego w przekroju



Wykres krzywych izofonicznych z zaznaczonym obszarem zajmowanym przez naturalny sygnał mowy

Teoria wytwarzania dźwięków mowy



Schemat zastępczy układu wytwarzania dźwięków mowy

Formanty numeruje się: F1, F2, F3 itd., a odpowiadające im częstotliwości w Hz oznacza się jako F_1, F_2, F_3

Największe znaczenie mają dwie wnęki jamy ustnej wynikające z obecności języka (dwa formanty F1 i F2),
inne wnęki - jama gardłowa, ustna i nosowa.

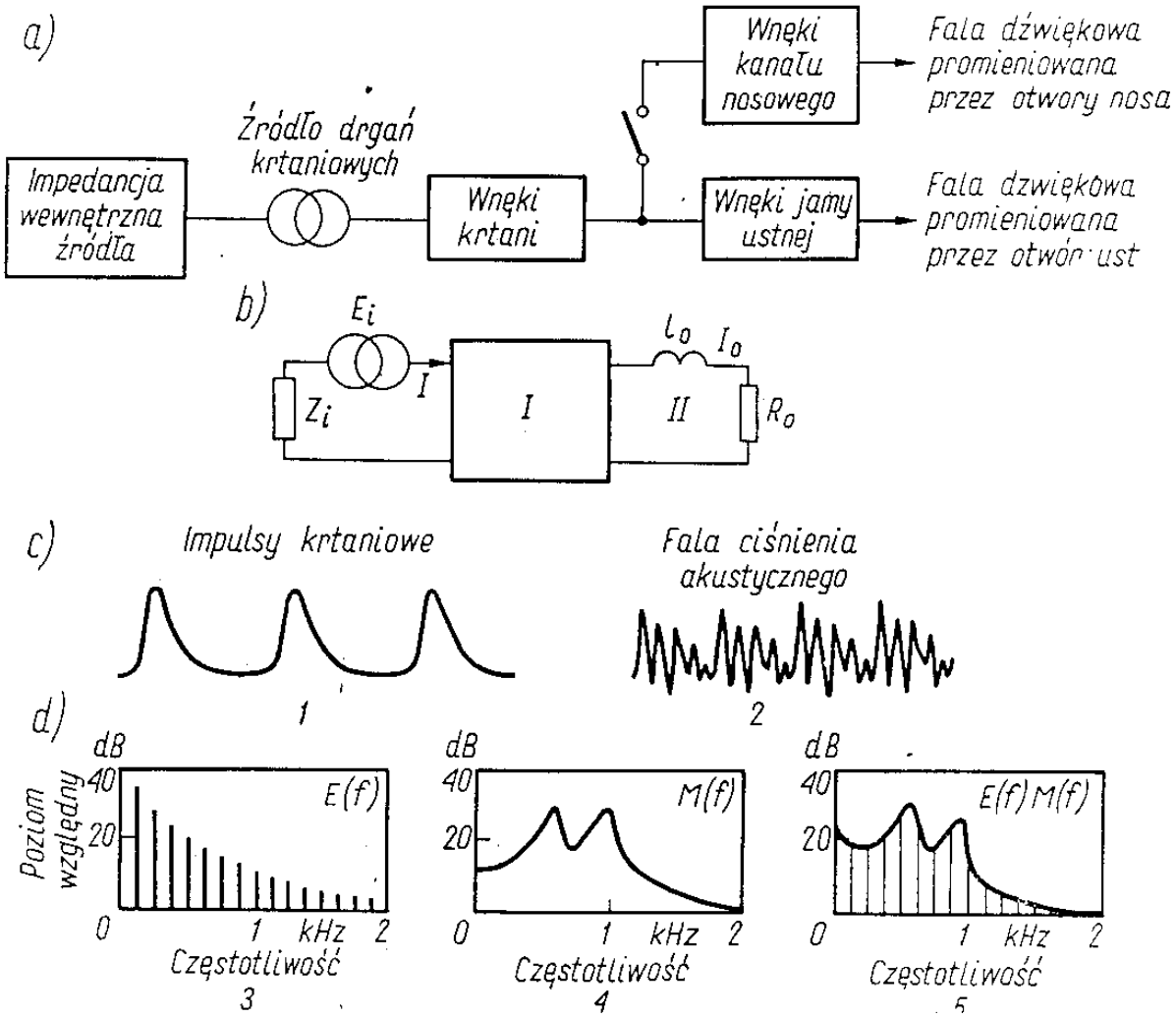
Podstawowe założenie teorii wytwarzania dźwięków mowy:
Niezależność rezonansowych właściwości i charakterystyk efektorów artykulacyjnych i źródła tonu krtaniowego

Parametry formantowe zależą zarówno od tonu krtaniowego jak i od właściwości rezonansowych organu mowy - traktu głosowego

Wyznaczenie struktury formantowej widma sygnału mowy:
uśrednianie kształtu jego obwiedni w przedziałach częstotliwości o szerokości 250-300 Hz (w zakresie dolnym widma < 1500 Hz) oraz 500-700 Hz (w górnym zakresie >2500 Hz) –
ogólnie: powinno to być realizowane przy pomocy filtracji zbliżonej do przypadku zastosowania filtrów o stałej dobroci.

struktura formantowa samogłosek w mowie ciągłej zależy także od fonemu poprzedzającego

stała czasowa słuchu: narastanie 20-30 ms, zanikanie 100-200 ms



Mechanizm wytwarzania dźwięków mowy jako proces kształtowania widma tonu krtaniowego (impulsów krtaniowych)

a) elektryczny układ zastępczy

b) czwórnikowy układ zastępczy dla głosek nienosowych

c) przebiegi czasowe

d) charakterystyki częstotliwościowe, kolejno: tonu krtaniowego, traktu głosowego, sygnału wynikowego

Modelowanie mechanizmów wytwarzania dźwięków mowy

TON KRTANIOWY (POBUDZENIE DLA GŁOSEK DŹWIĘCZNYCH)

Jest często nazywany formantem F0 – jego częstotliwość w konsekwencji to parametr $F_{0,}$ powstaje jako wynik modulacji strumienia powietrza wypływającego z płuc przez wiązadła głosowe

- wyniki modelowania prowadzą do przybliżenia wartości nachylenia obwiedni widma tunu krtaniowego jako –6...-12 dB/oktawę,
- jako przybliżenie przebiegu tonu krtaniowego często stosuje się przebieg piłokształtny, którego obwiednia widma ma nachylenie -6 dB/oktawę/

Przyjmuje się, że ton krtaniowy to sygnał o częstotliwości podstawowej wynikającej z charakteru głosu mówcy (np. tenor - 120-480 Hz) i o widmie składającym się z wszystkich składowych harmonicznych z obwiednią o nachyleniu od –6 do –12 dB/oktawę

POBUDZENIE SZUMOWE

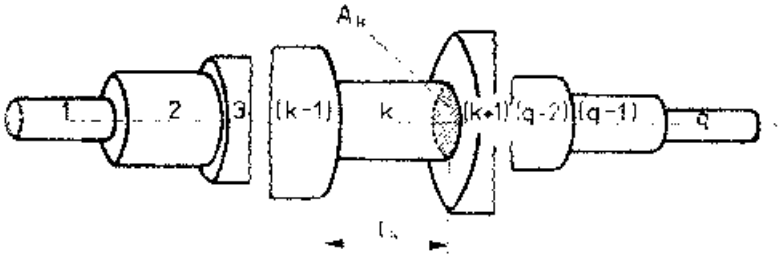
Szumy turbulencyjne - wtórny efekt działania strumienia powietrza
fala udarowa (przy nagłym otworzeniu drogi przepływu) sama staje się źródłem fal (spółgłoski zwarte)

obwiednia widma - 6 dB/oktawę

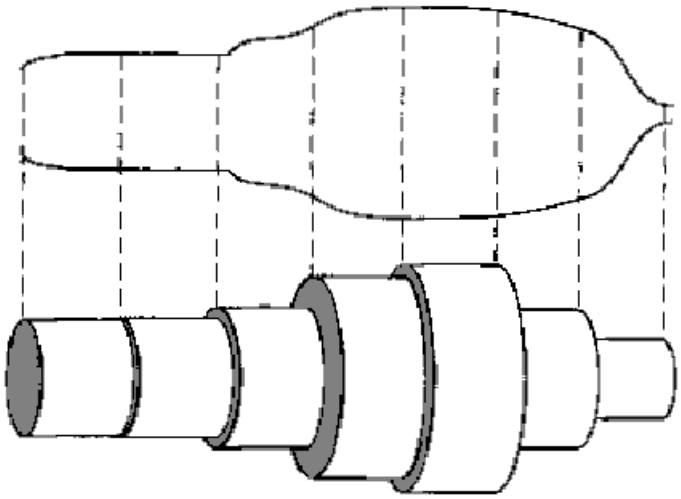
TRAKT GŁOSOWY

Jest modelowany jako układ fragmentów ściętych stożków lub układ walców. W tym pierwszym przypadku powstaje model tubowy, zachowujący ciągłość przekroju, w drugim model cylindryczny. Fakt, że ten drugi model jest łatwiejszy do analizy powoduje jego rozpowszechnienie do różnych symulacji:

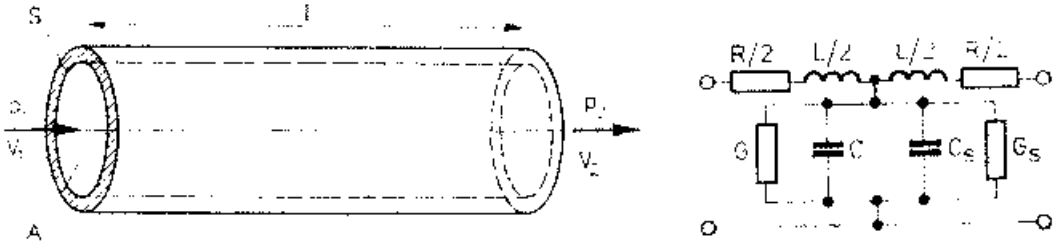
- rezonator Helmholtza (umożliwia modelowanie pojedynczego formantu)
- podwójny rezonator Helmholtza (umożliwia modelowanie dwóch formantów)
- modele złożone z kilku rur zakończonych płaską tarczą kołową (odgrodą) imitującą charakterystykę promieniowania ust jako nadajnika dźwięku
- trójparametrowy model Fanta, uwzględniający rozkład biegunów i zer na płaszczyźnie zespolonej i podstawowe trzy parametry: miejsce artykulacji (miejsce największego przewężenia kanału), stopień tego przewężenia (powierzchnia przekroju) oraz kształt otworu wylotowego ust
- model Markela-Graya



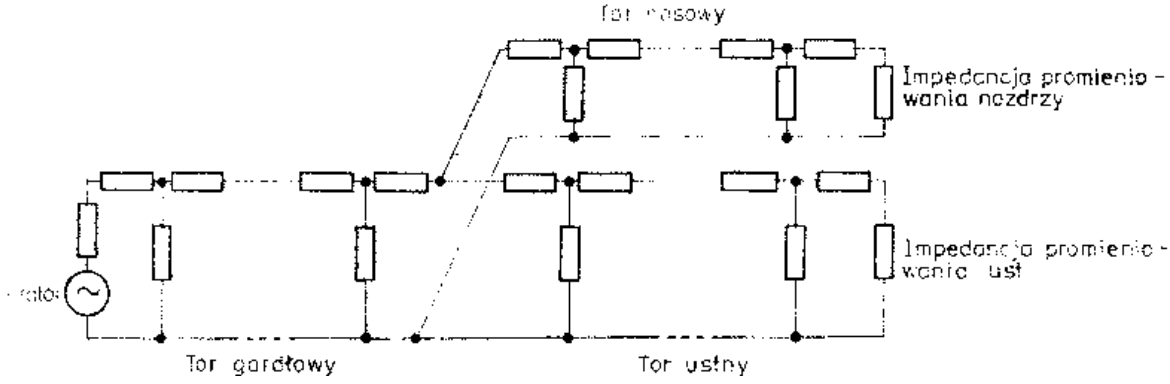
Uproszczony model traktu głosowego (w ogólnym przypadku poszczególne elementy nie są równe)



Model traktu głosowego – fizyczny i cylindryczny



Elementarny fragment modelu traktu głosowego (z lewej strony) i czwórnik elektryczny stosowany jako analogia elementarnego odcinka (z prawej)



Ogólna struktura modelu elektrycznego

Uproszczenia fizycznego modelu cylindrycznego:

1. niezgodność kształtu przekroju poprzecznego
2. brak płynności zmian przekroju
3, nieuwzględnienie elastyczności – sztywności ścianek

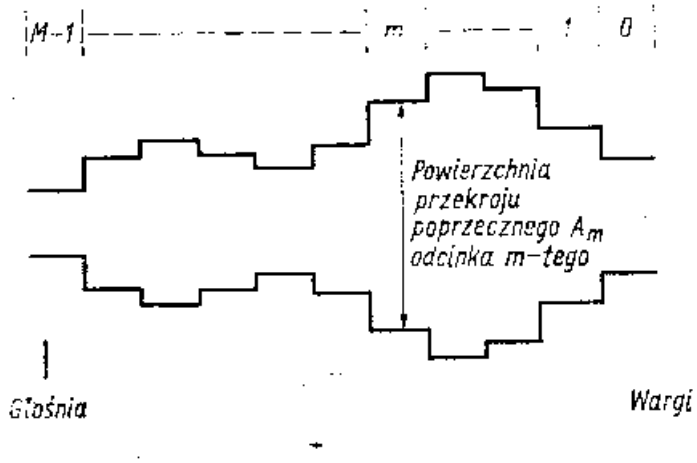
płuca, oskrzela mają niewielki wpływ na sygnał mowy (różnica 2 rzędów wielkości)

główny podział głosek polskich: dźwięczne i bezdźwięczne

częstotliwości własne wnęk są bliskie częstotliwościom formantowym

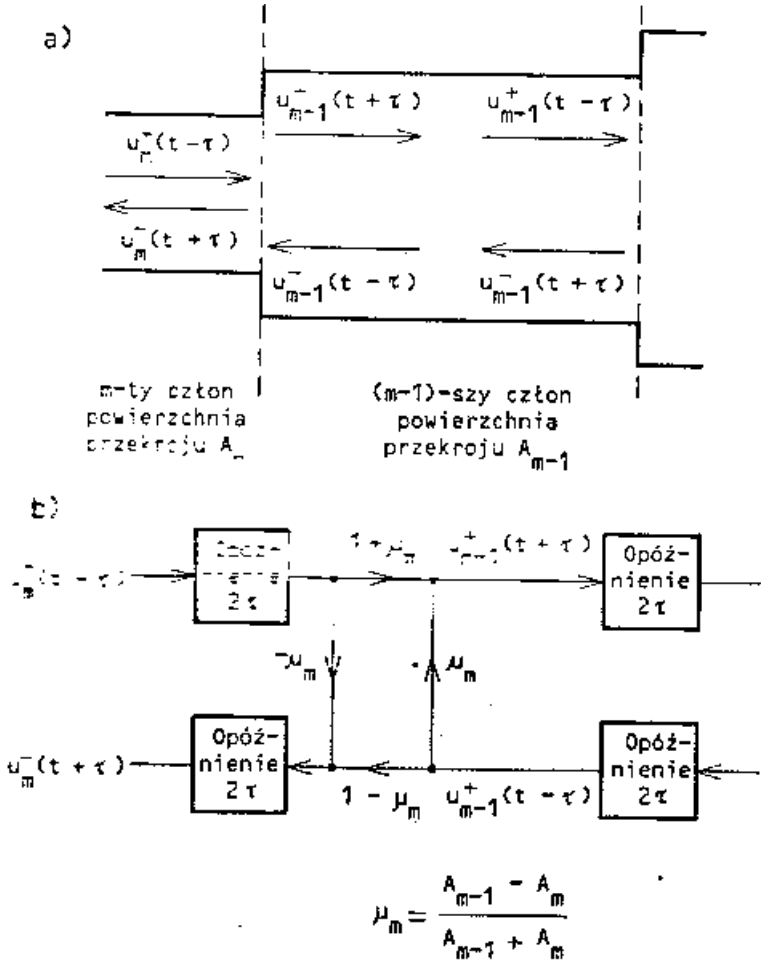
Model Markela-Graya:

- kanał głosowy jest zamodelowany jako kaskadowe połączenie cylindrycznych rur o jednakowej długości
- dźwięk rozchodzi się jako fala płaska, brak strat wewnętrznych i brak sprzężenia pomiędzy kanałem głosowym i głośnią



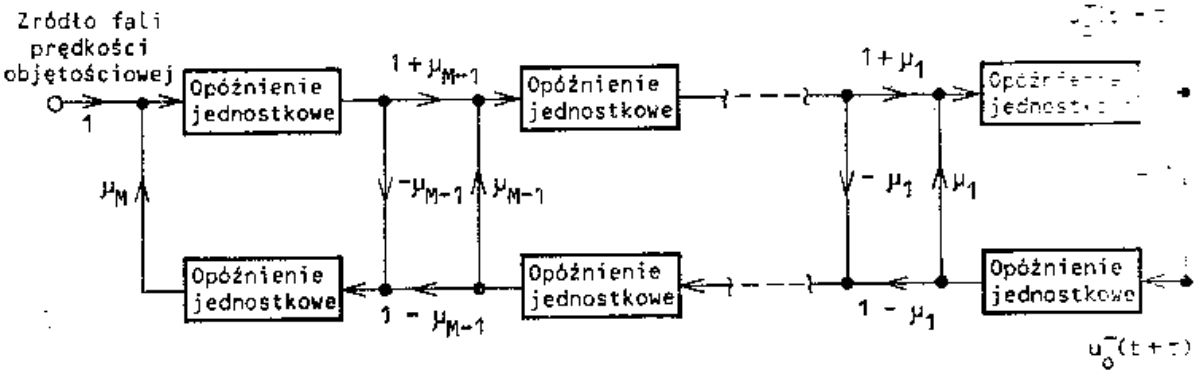
Model konfiguracyjny kanału głosowego jako zbiór kaskadowo połączonych odcinków cylindrycznych o jednakowych długościach i zmieniającym się przekroju

ciśnienie lub prędkość objętościową przedstawia się jako funkcję czasu i położenia wzdłuż osi rury
zachowana jest ciągłość na granicy dwóch członów, co prowadzi do odbicia fal w tym miejscu
związki pomiędzy tymi falami można przedstawić w postaci grafu przepływowego

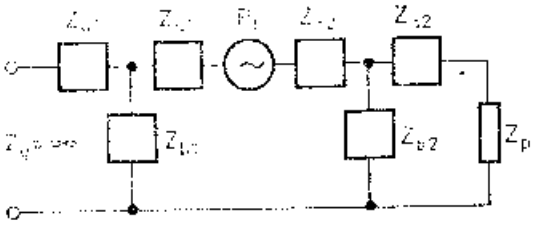


$$\mu_m = \frac{A_{m-1} - A_m}{A_{m-1} + A_m}$$

Dwa człony rury akustycznej z zaznaczeniem fal prędkości bieżącej i powrotnej (a) i graf przepływu sygnału dla prędkości objętościowej (b)



Liniowy graf przepływu sygnału opisujący zależności pomiędzy falami prędkości bieżącej i powrotnej w całym modelu Markela-Graya



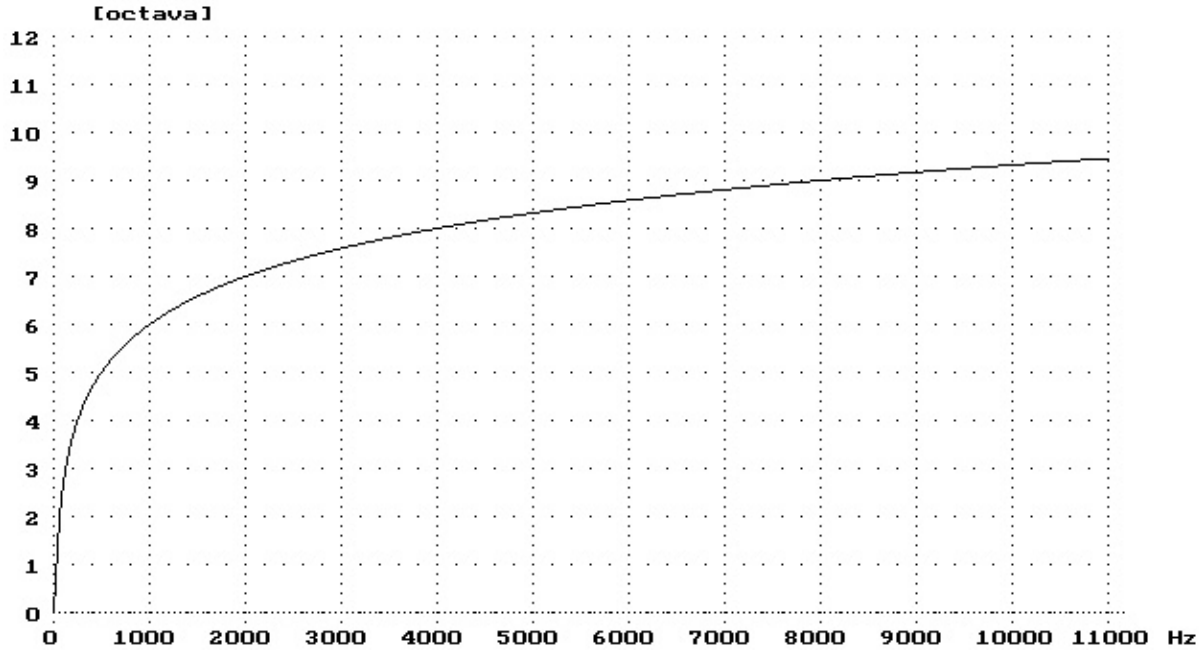
Uproszczony model procesu artykulacji głosek szumowych

Perceptualne skale częstotliwości

We wszystkich podanych poniżej wzorach na nieliniowe skale częstotliwości symbol f oznacza częstotliwość wyrażoną wkHz

Skala logarytmiczna (znana z akustyki muzycznej, odpowiada strojowi równomiernie temperowanemu):

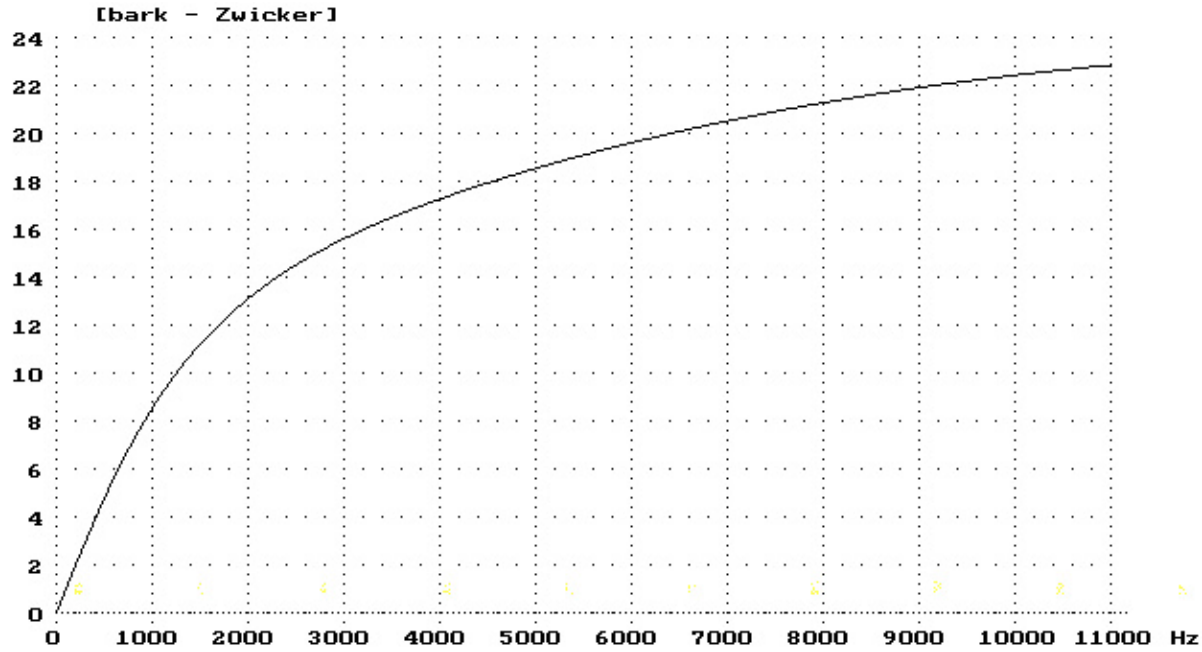
$$oktawa = \log_2(64 \cdot f)$$



Zależność pomiędzy liniową skalą częstotliwości a skalą oktawową

Skala barkowa jest związana z pojęciem pasma krytycznego, wynikającego z badań nad percepcją głośności szumu wąskopasmowego (Zwicker) lub zjawisk maskowania tonu prostego przez taki szum (Schröder). Całe pasmo słyszenia zostało podzielone na 24 pasma krytyczne. Możliwe stało się określenie zależności pomiędzy wysokością tonu w barkach a częstotliwością w hercach. Skala barkowa wg Zwickera:

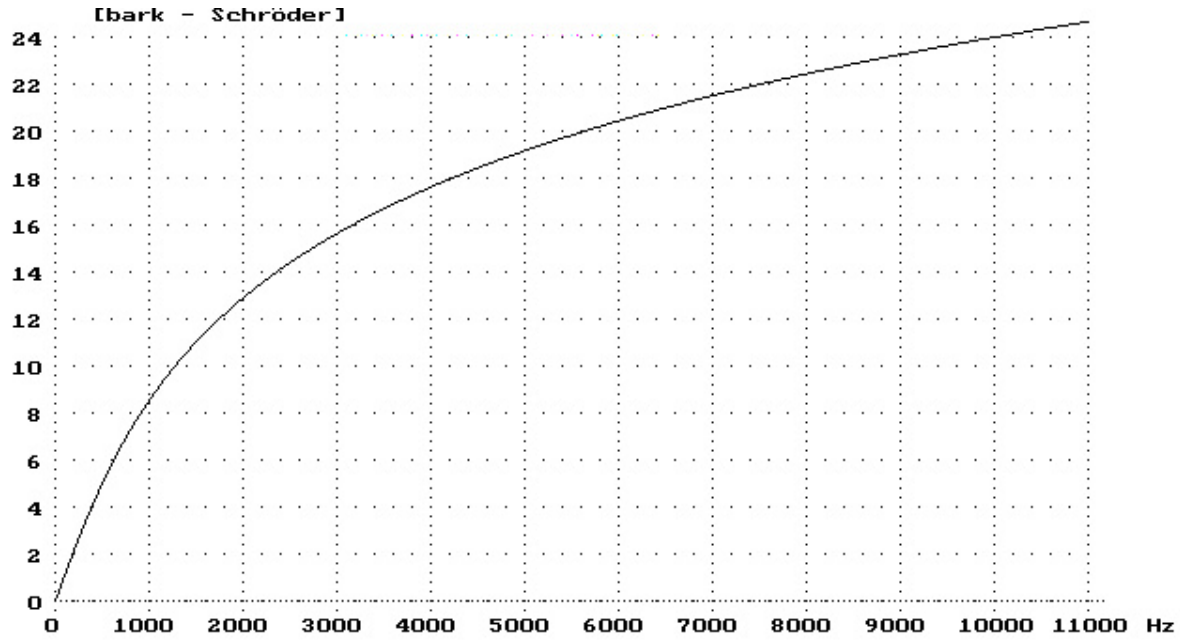
$$b = 13 \cdot \arctan(0.76 \cdot f) + 3.5 \cdot \arctan\left(\left(f/7.5\right)^2\right)$$



Zależność pomiędzy liniową skalą częstotliwości a skalą barkową Zwickera

Skala barkowa wg Schrödera:

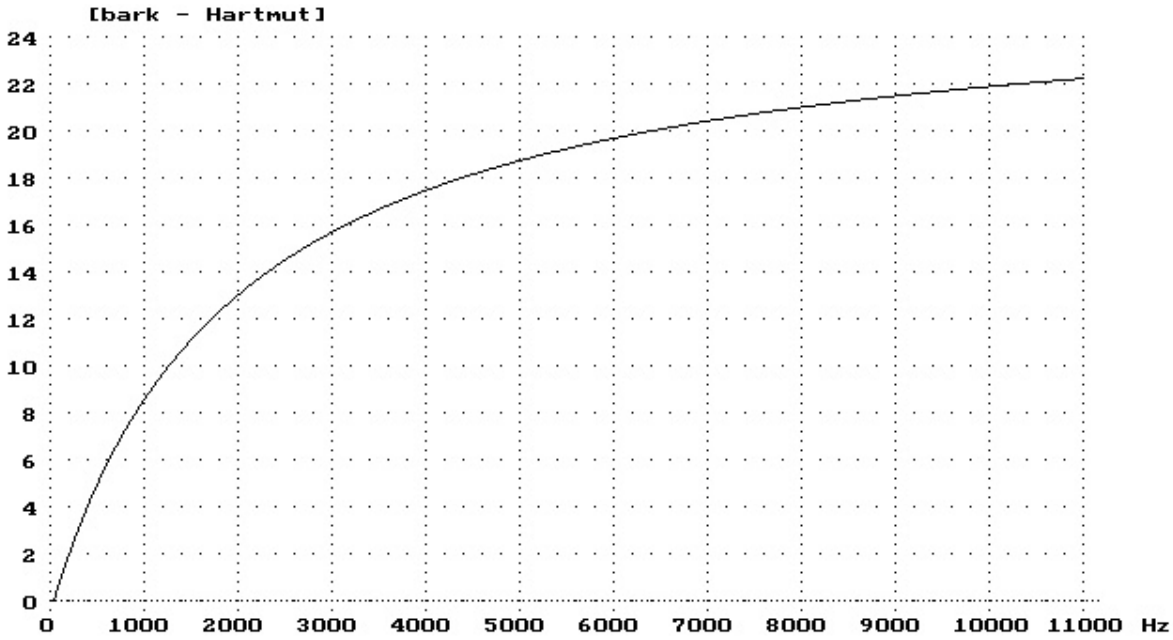
$$b = 7 \cdot \operatorname{arcsin} h\left(f / 0.65 \right)$$



Zależność pomiędzy liniową skalą częstotliwości a skalą barkową Schrödera

Skala barkowa wg Hartmuta:

$$b = \frac{26.81}{1 + \frac{1.96}{f}} - 0.53$$



Zależność pomiędzy liniową skalą częstotliwości a skalą barkową Hartmuta

Skala barkowa wg Boersmy & Weeninka:

$$b = 7 \cdot \ln\left(f/0.65 + \sqrt{1 + f/0.65} \right)$$

Skala melowa jest skalą dotyczącą wysokości tonu, czyli wrażenia słuchowego pozwalającego na określenie położenia tonu na skali częstotliwości. Wrażenie to zależy jednak także od natężenia dźwięku i dlatego w definicji przyjęto tę wartość jako 40dB odpowiadające ciśnieniu $2 \cdot 10^{-5}$Pa

Skala melowa wg Beranka:

$$M = 1127 \cdot \ln\left(1 + {f}/{0.7}\right)$$

Zależność pomiędzy liniową skalą częstotliwości a skalą melową Beranka

Skala melowa wg Boersmy & Weeninka:

$$M = 550 \cdot \ln\left(1 + {f}/{0.55}\right)$$

Skala Königa (zakres 0 – 4000Hz):

- 10 podpasm o stałej szerokości 100 Hz dla zakresu 0 – 1000Hz
- 10 podpasm o zmiennej szerokości (logarytmicznie) dla zakresu 1000Hz - 4000Hz (zmiana szerokości o czynnik 1.193)

Metody analizy sygnału mowy

Poziomy analizy:

- akustyczny – związany z wprowadzaniem sygnału do systemu (dobór pasma, zastosowanie preemfazy, system kodowania itp.),
- parametryczny – ekstrakcja (wydzielanie) parametrów i redukcja informacji, co powinno prowadzić do równoważnego zapisu parametrycznego pod względem identyfikacyjnym,
- strukturalny – podział sygnału na segmenty, które powinny podlegać rozpoznawaniu,
- leksykalny – powinien prowadzić do syntezy rozpoznawanych elementów fonetycznych w całościowe elementy rozpoznania - najczęściej wyrazy,
- syntaktyczny – analiza gramatyczna wypowiedzi,
- semantyczny – identyfikacja treści wypowiedzi i wydobycie jej „sensu”

DZIEDZINA CZASU

Funkcja autokorelacji *r(i)* sygnału *x(i)* może być przedstawiona przy pomocy ogólnego równania:

$$r(m) = \frac{\sum_{i=q}^{q+N-1} x(i)x(i+m)}{\sum_{i=q}^{q+N-1} [x(i)]^2}$$

lub inaczej funkcja autokorelacji to:

$$R(n) = \frac{\sum_{i=1}^{k} \left(X_i - \overline{X_{k,i}}\right)\left(X_{i+n} - \overline{X_{k,i+n}}\right)}{\sqrt{\sum_{i=1}^{k} \left(X_i - \overline{X_{k,i}}\right)^2 \sum_{i=1}^{k} \left(X_{i+n} - \overline{X_{k,i+n}}\right)^2}}$$

gdzie:

$$\overline{X_{k,i}} = \frac{1}{k} \sum_{j=i}^{k+i} X_j$$

Metoda AMDF (Average Magnitude Differential Function), nazywana również metodą filtru grzebieniowego, stanowi modyfikację metody autokorelacyjnej. Metoda ta polega na badaniu różnicy pomiędzy sygnałem, a jego przesunięciem w dziedzinie czasu:

$$AMDF(m) = \sum_{i=q}^{q+N-1} \left|x(i) - x(i+m)\right|^k$$

Wykładnik *k* może przyjmować różne wartości, np. jeśli zostanie przyjęty jako 2 to wzór ten będzie przypominać podobny wzór służący do obliczenia błędu średniokwadratowego.
Obie te metody mogą służyć do badania okresowości sygnału, w przypadku sygnału mowy do określenia dźwięczności danego fragmentu i ewentualnie estymacji częstotliwości tonu krtaniowego.

Preemfaza 6 dB/oktawa jest równoważna operacji różniczkowania:

$$x_p(t) = \frac{d}{dt}[x(t)]$$

lub dla sygnału skwantowanego w dziedzinie czasu:

$$x_p(n) = x(n+1) - x(n)$$

Preemfazę stosuje się w celu stłumienia niskich częstotliwości i wyeliminowania składowej stałej (np. podczas analizy przejść przez zero lub kodowania sygnału).

DZIEDZINA CZĘSTOTLIWOŚCI

Transformata Fouriera sygnału:
gdzie:* f – *częstotliwość,

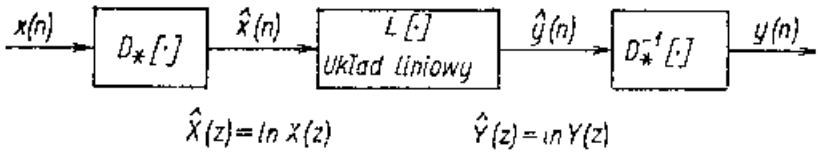
$$X(f) = \int_0^T y(t) \cdot e^{-j \cdot 2\pi \cdot f \cdot t} dt$$

t – *czas,*
y(t) – funkcja czasu (sygnał),
T – długość przedziału całkowania; interpretacja wyników zależy od charakteru sygnału i od doboru wartości przedziału całkowania (tutaj przyjęto *<0,T>*)

lub w skrócie:

$$X(f) = F[y(t)]$$

Analiza homomorficzna jest używana do tzw. rozplotu sygnału mowy (operacja odwrotna do splotu). Sygnał mowy jest splotem funkcji pobudzenia i odpowiedzi impulsowej kanału głosowego, stąd rozplot prowadzi do rozdzielenia obu tych przebiegów.



Postać kanoniczna systemu homomorficznego

Układ D∗[˙] przekształca splot sygnałów w sumę (sygnał na wyjściu tego układu to cepstrum zespolone – cepstrum to anagram słowa spectrum), która w tym wypadku dla małych n oznacza współczynniki cepstralne opisujące trakt głosowy, a dla wyższych n wpółczynniki te opisują pobudzenie.

Układ L[˙] poprzez zastosowanie odpowiedniego okna prostokątnego dokonuje wyboru jednego lub drugiego składnika.
Końcowy układ poprzez operację pozwala uzyskać odpowiednie przebiegi czasowe lub też wcześniej ich widma (np. transmitancja traktu głosowego – widmo wygładzone cepstralnie.)

Cepstrum zespolone sygnału jest zdefiniowane jako:

$$\hat{X}(T) = F\left[\ln\left(X(f)\right)\right]$$

gdzie: T – dziedzina czasu dla cepstrum,

Cepstrum mocy (transformacja Fouriera):

$$\hat{X}(T) = F\left[\ln\left|X(f)\right|\right]$$

Cepstrum mocy sygnału (transformacja kosinusowa):

$$\hat{X}_c(k) = \sum_{n=0}^{N-1}\left[\ln\left|X(n)\right|\right]\cdot\cos\left(\frac{(n-0.5)\cdot k\cdot\pi}{N}\right)$$

gdzie: $X(n)$ – dyskretne widmo mocy
n –numer prążka widma
N–numer maksymalnego prążka widma analizowanego pasma częstotliwości,
k–numer współczynnika cepstralnego

Mel-cepstrum (współczynniki mel-cepstralne) to cepstrum w skali melowej (transformacja kosinusowa):

$$M(k) = \sum_{n=1}^{N}\left[\ln\left|E(n)\right|\right]\cdot\cos\left(\frac{(n-0.5)\cdot k\cdot\pi}{N}\right)$$

Widmo wygładzone cepstralnie (transformacja kosinusowa):

$$X_c(n) = \sum_{k=0}^{K} \hat{X}_c(k) \cdot \cos\left(\frac{n \cdot k \cdot \pi}{N}\right)$$

gdzie: K – rząd wygładzania, oznacza to zastosowanie w stosunku do cepstrum okna prostokątnego o wartościach: 1 dla $k<=K$ i 0 dla $k>K$, odpowiedni dobór K zapewnia wyeliminowanie sygnału pobudzenia, czyli tony krtaniowego.

KRÓTKOOKRESOWA ANALIZA FOURIEROWSKA

$$S(\omega, n) = \sum_{k=-\infty}^{+\infty} s(k) \cdot h(n-k) \cdot e^{-j\omega k}$$

gdzie: $s(n)$ – spróbkowany sygnał mowy
$h(n)$ – funkcja okna

$$S(\omega, n) = \left[s(n) \cdot e^{-j\omega n}\right] * h(n)$$

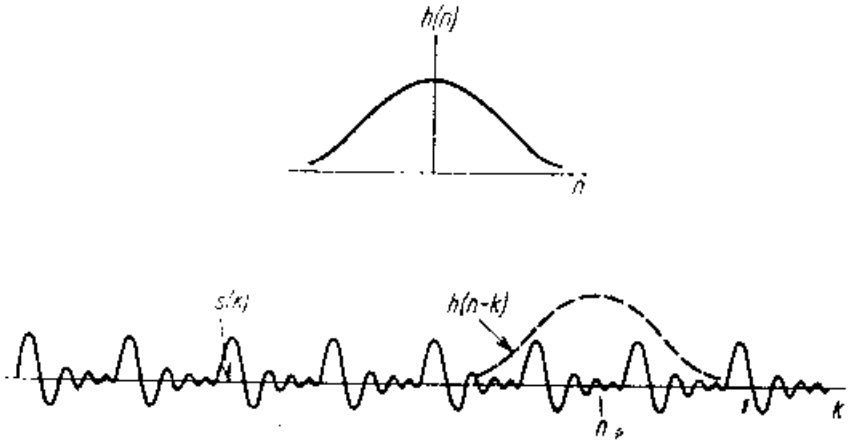
jest to realizacja analizy poprzez zestaw filtrów

$$S(\omega, n) = e^{-j\omega n} \cdot \sum_{k=-\infty}^{+\infty} s(k) \cdot h(n-k) \cdot e^{j\omega(n-k)}$$

$$S(\omega, n) = e^{-j\omega n} \cdot \left\{s(k) * \left[h(n) \cdot e^{j\omega n}\right]\right\}$$

gdzie:

$h(n) \cdot e^{-j\omega n}$ - filtr środkowoprzepustowy o częstotliwości środkowej ω



Przedstawienie krótkookresowej transformacji Fouriera

ANALIZA LPC (linear predictive code)

Ogólna postać transmitancji wymiernej opisującej kanał głosowy przedstawia się następująco:

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

gdzie:
G - wzmocnienie,
b_l – współczynniki opisujące zera transmitancji,
a_k – współczynniki opisujące bieguny transmitancji.

Odpowiedź impulsowa oraz charakterystyka częstotliwościowa odpowiadające tej transmitancji są nieliniowymi funkcjami współczynników licznika i mianownika, zatem obliczenie tych parametrów polega na rozwiązaniu układu równań nieliniowych.

Podejście to jest ogólne w tym sensie, że zakłada jednoczesną obecność zer i biegunów w rozpatrywanej transmitancji. Dla często przyjmuje się opis transmitancji jako zawierającej wyłącznie zera (stopień mianownika p=0) lub wyłącznie bieguny (stopień licznika q=0). W każdym z tych przypadków rozwiązanie opiera się na układzie równań liniowych. Ten drugi przypadek (wyłącznie bieguny) jest o tyle uzasadniony, że prowadzi do aproksymacji charakterystyki kanału głosowego w postaci ukazującej częstotliwości rezonansowe, czyli ujawniającej naturę formantową sygnału mowy.

Równanie to w przypadku pominięcia zer upraszcza się do postaci:

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

Odpowiedź impulsowa dla powyższej transmitancji jest opisana przez równanie różnicowe:

$$v(n) = G \cdot \delta(n) + \sum_{k=1}^{p} a_k \cdot v(n-k)$$

Dla $n>0$ równanie upraszcza się do postaci:

$$v(n) = \sum_{k=1}^{p} a_k \cdot v(n-k)$$

Prawa strona powyższego równania to kombinacja liniowa p poprzednich wartości odpowiedzi impulsowej, stąd pochodzi nazwa predykcja liniowa. Ze względu na to, że model jest jedynie przybliżeniem rzeczywistej sytuacji, można jedynie zminimalizować błąd ***e(n)*** pomiędzy wartościami obserwowanymi ***v(n)*** a otrzymanymi z modelu $\hat{v}(n)$:

$$e(n) = v(n) - \hat{v}(n) = v(n) - \sum_{k=1}^{p} a_k \cdot v(n-k)$$

Za kryterium służącym do obliczenia współczynników predykcji a_k przyjmuje się minimum błędu średniokwadratowego:

$$E = \sum_{n=1}^{N-1} e^2(n) = \sum_{n=1}^{N-1} \left[v(n) - \sum_{k=1}^{p} a_k \cdot v(n-k) \right]^2$$

W powyższym wzorze górna granica sumowania N-1 oznacza liczbę dostępnych próbek ciągu $v(n)$. Obliczenie współczynników predykcji sprowadza się więc do rozwiązania układu p równań:

$$\frac{\partial E}{\partial a_i} = 0$$

gdzie ***i*=1, 2 ...*p***.

Do rozwiązania powyższego układu równań stosowane są zazwyczaj dwie metody: autokowariancji lub częściej zalecana metoda autokorelacji. Każda z tych metod ma wady i zalety: pierwsza z nich jest dokładniejsza, ale może prowadzić do niestabilnych rozwiązań. Druga natomiast zapewnia stabilność, czyli lokalizację rozwiązań wewnątrz jednostkowego okręgu na płaszczyźnie zespolonej. Ponadto współczynniki autokorelacji są elementami macierzy Toeplitza, co umożliwia zastosowanie szybkiego algorytmu iteracyjnego odwracania macierzy (algorytmy Levinsona, Robinsona i Durbina). Dodatkowo przy zastosowaniu algorytmu Durbina uzyskuje się tablicę współczynników odbicia, co stanowi nawiązanie do cylindrycznego modelu traktu głosowego zaproponowanego przez Markela-Graya.

Metoda Durbina:

$$k_i = \frac{\sum_{j=1}^{i-1} \alpha_j^{(i-j)} R(i-j) - R(i)}{E_{i-1}}$$

$$a_i^{(i)} = -k_i$$

$$a_j^{(i)} = a_j^{(i-1)} + k_i \cdot a_{i-j}^{(i-1)}$$

$$E_i = \left(1 - k_i^2\right) \cdot E_{i-1}$$

gdzie:
j=1…i-1

przy czym:
$a_j^{(i)}$ dla j=1,2…, i – współczynniki predykcji układu i-tego rzędu,

Zbiór równań rozwiązuje się rekurencyjnie dla i=1,2…, p, zaczynając od E_0=R(0)

Rozwiązanie końcowe:
$a_j = a_j^{(p)}$
j=1,2…, p
k_j – współczynniki odbicia

Standardy μ-law i A-law

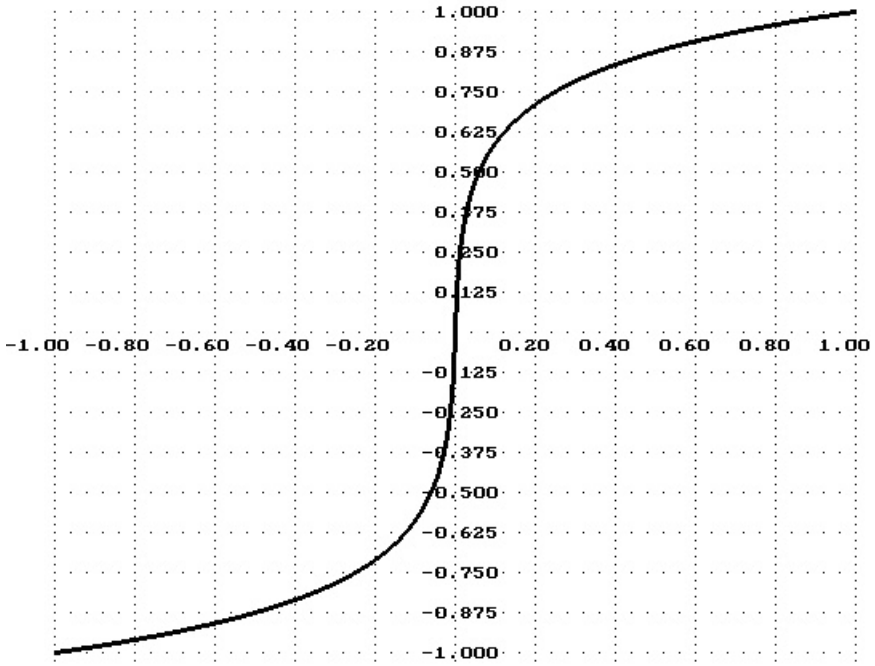
Podstawą dla nieliniowej kwantyzacji jest
prawo Webera-Fechnera:
Minimalny dostrzegalny przyrost dowolnego bodźca Δp jest proporcjonalny do wartości tego bodźca, względem którego dokonuje się tego porównania:

$$\Delta p = k \cdot p$$

Występują jednak ograniczenia zakresu stosowalności prawa Webera-Fechnera - dotyczą one skrajnych zakresów skali: dolnej - w pobliżu progu czułości i górnej, gdzie występuje zjawisko nasycenia.
Z prawa Webera-Fechnera wynika celowość stosowania skali logarytmicznej w celu dokonania kompresji amplitudy sygnalu przed jego transmisją lub przetwarzaniem. Funkcję realizującą takie przekształcenie nazywa się funkcją kompresji. Oczywiście dla odtworzenia pierwotnego sygnału należy zastosować funkcję do niej odwrotną.
W praktyce stosowane skale są zmodyfikowane w sposób pozwalający na uniknięcie obliczania logarytmu z zera.

Nieliniowa kwantyzacja μ-law (amerykańska):

$$F(x) = \operatorname{sgn}(x) \cdot \frac{\ln(1+\mu \cdot |x|)}{\ln(1+\mu)} \qquad dla \quad -1 \le x \le 1$$



Wykres zależności pomiędzy skalą liniową a skalą μ-law

Nieliniowa kwantyzacja A-law (europejska – Niemiecki Urząd Poczt):

$$F(x)=\operatorname{sgn}(x)\cdot\frac{A\cdot|x|}{1+\ln(A)} \qquad dla \quad \frac{1}{A}\le x\le 1 \quad oraz \quad -1\le x\le -\frac{1}{A}$$

$$F(x)=\operatorname{sgn}(x)\cdot\frac{1+\ln(A\cdot|x|)}{1+\ln(A)} \qquad dla \quad -\frac{1}{A}\le x\le \frac{1}{A}$$

Wartości funkcji kompresji dla wybranych punktów skali nieliniowych:

μ-law (μ = 247):

x	0.5	0.25	0.125	0.0625	0.03125	0.015625
F(x)	0.87501	0.75074	0.62789	0.50777	0.39276	0.28674

A-law (A = 87.7):

x	0.5	0.25	0.125	0.0625	0.03125	0.015625
F(x)	0.87337	0.74675	0.62012	0.49349	0.36686	0.24024

skala logarytmiczna:

x	0.5	0.25	0.125	0.0625	0.03125	0.015625
F(x)	0.875	0.750	0.625	0.500	0.375	0.250

Zastosowanie powyższych standardów pozwala na zwiększenie dynamiki sygnału o około 24dB, tzn. sygnał zakodowany na 8 bitach odpowiada sygnałowi o kwantyzacji liniowej 12 bitów.

Standardy te są punktem odniesienia dla obliczeń stopnia kompresji sygnału mowy w przypadku wokoderów (czyli: częstotliwość próbkowania = 8kHz, liczba bitów na próbkę = 8, co oznacza szybkość transmisji 64 kilobity/sek.). Przykładowo dla kompresji 1:10 szybkość transmisji wynosi 6,4 kb/sek.

Parametryzacja sygnału mowy

DZIEDZINA CZASU:

Możliwe są dwa podejścia:
1. Oparte na tzw. makrostrukturze sygnału – obliczenia są wykonywane w odcinkach czasowych po wstępnej segmentacji, uzyskane parametry to amplituda i szybkość zmian.
2. Oparte na tzw. mikrostrukturze sygnału, czyli przebiegu czasowym, analizującym przejścia sygnału mowy przez zero. Prowadzi to uzyskania dwóch rodzajów parametrów: gęstość przejść przez zero i rozkład interwałów czasowych. Analiza przejść przez zero powstała w oparciu o spostrzeżenie, że sygnał mowy zachowuje zrozumiałość w przypadku dokonania przekształcenia na falę prostokątną (mimo dużych zniekształceń i utraty jakości). Zostaje wówczas zachowana jedynie informacja o momentach czasowych, w których sygnał przechodzi przez zero. Odpowiada to kodowaniu jednobitowemu.
Zaletą parametryzacji czasowej jest prostota i szybkość algorytmu.
W praktyce okazało się, że parametry czasowe nie są najlepsze pod względem skuteczności rozpoznawania mowy, pomimo stosowania dodatkowych zabiegów na sygnale: preemfaza 6dB/oktawę (różniczkowanie), preemfaza 12dB/oktawę (dwukrotne różniczkowanie), deemfaza (całkowanie) i inne.
Lepsze okazały się parametry częstotliwościowe.

DZIEDZINA CZĘSTOTLIWOŚCI:

Moment widmowy m-tego rzędu:

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

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 unormowany centralny m-tego rzędu:

$$M_{uc}(m) = \sum_{k=0}^{\infty} \frac{|G(k)| \cdot [f_k - M_u(1)]^m}{M(0)}$$

Szczególne przypadki momentów widmowych:

Moment rzędu zerowego, mający zastosowanie normalizujące, oznacza moc sygnału:

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

Moment unormowany pierwszego rzędu jest używany we wzorach do obliczeń momentów centralnych wyższych rzędów – ma interpretację środka ciężkości widma:

$$M_u(1)=\sum_{k=0}^{\infty}\frac{|G(k)|\cdot f_k}{M(0)}$$

Moment unormowany centralny drugiego rzędu – ma interpretację kwadratu szerokości widma:

$$M_{uc}(2)=\sum_{k=0}^{\infty}\frac{|G(k)|\cdot[f_k-M_u(1)]^2}{M(0)}$$

Moment unormowany centralny trzeciego rzędu to niesymetria widma, inaczej skośność (ang. skewness):

$$M_{uc}(3)=\sum_{k=0}^{\infty}\frac{|G(k)|\cdot[f_k-M_u(1)]^3}{M(0)}$$

Parametr będący miarą płaskości widma (ang. flatness):

$$kurtosis=\frac{M_{uc}(4)}{[M_{uc}(2)]^2}$$

inaczej:

$$kurtosis=\frac{1}{N}\sum_{j=1}^{N}\frac{(x_j-\bar{x})^4}{\sigma_x^4}$$

gdzie:

x_j – j-ta obserwacja spośród N dostępnych obserwacji

x – średnia arytmetyczna dla wszystkich N obserwacji

σ_x – odchylenie standardowe liczone na podstawie obserwacji jako estymator nieobciążony:

$$\sigma_x = \sqrt{\frac{1}{N-1} \cdot \sum_{j=1}^{N} \left(x_j - \bar{x}\right)^2}$$

Inny parametr służący jako miara płaskości widma (ang. spectral flatness measure):

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

gdzie: $P\left(e^{j\frac{2\pi k}{N}} \right)$ to widmowa gęstość mocy obliczona za pomocą N-punktowej transformacji Fouriera.

Momenty widmowe mogą być także liczone dla fragmentów widma, zakresy sumowania w powyższych wzorach muszą wówczas zostać zmienione z <0, ∞> na <f_d, f_g>, gdzie: f_d i f_g to punkty widma odpowiadające częstotliwości dolnej i górnej. Przykładowo pierwszy moment znormalizowany (środek ciężkości widma) liczony w zakresie pomiędzy dwoma kolejnymi minimami obwiedni widma może być interpretowany jako częstotliwość formantu znajdującego się w tym paśmie częstotliwości.

W oparciu o obliczone widmo (lub jego fragment) można dokonać analizy cepstralnej, która prowadzi do uzyskania współczynników cepstralnych, z których niskie to parametry obwiedni widma, natomiast wyższe mogą nieść informację o tonie krtaniowym o ile w wykresie cepstrum występuje wyraźne maksimum (to tylko dla fonemów dźwięcznych). W tym przypadku parametry cepstralne to wektor składający się z niskich współczynników opisujących obwiednię widma, natomiast wyższe współczynniki mogą służyć jedynie do

ekstracji tonu krtaniowego (tzn. określenia czy istnieje oraz estymacji jego częstotliwości).

Stosując wygładzanie cepstralne można uzyskać parametry fomantowe jako współrzędne lokalnych maksimów widma wygładzonego cepstralnie.

Logarytm widma wygładzonego cepstralnie (transformacja kosinusowa):

$$Y(n) = \sum_{k=0}^{K} C_k \cdot \cos\left(\frac{n \cdot k \cdot \pi}{N}\right)$$

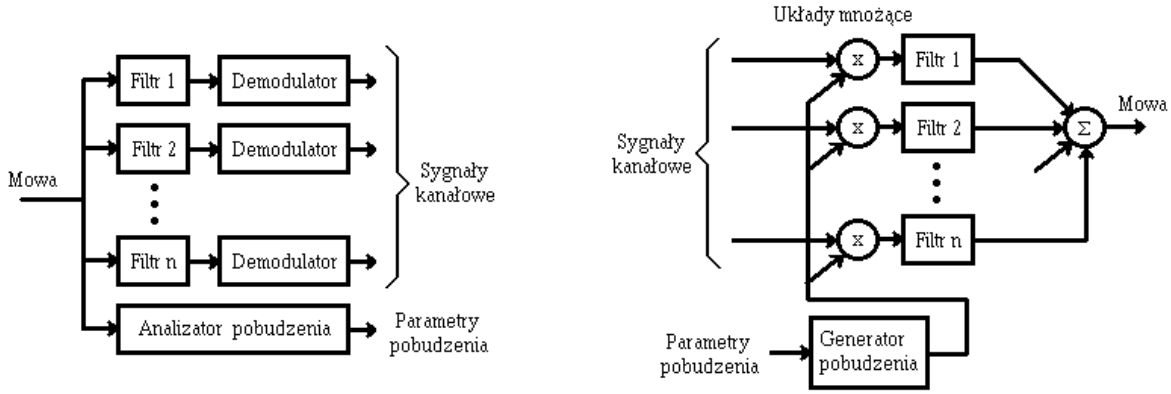
Spośród innych metod prowadzących do parametrów formantowych to klasyczna analiza przy pomocy filtrów o stałej dobroci oraz w dziedzinie cyfrowej analiza LPC.

Przykładowe parametry formantowe:

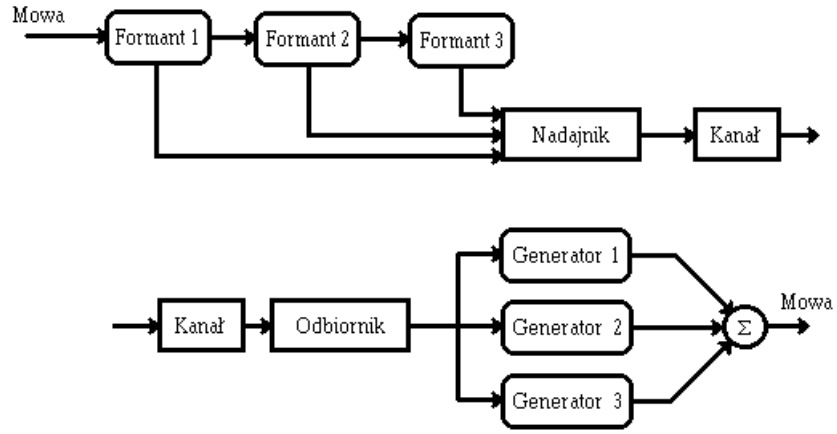
Fonem	częstotliwości [Hz]				poziomy względne [dB}			
i	210	2750	3500	4200	0	-15	-15	-27
e	380	2640	3000	3600	0	-12	-16	-20
a	780	1150	2700	3500	0	-7	-25	-25
y	240	1550	2400	3300	0	-12	-20	-30
o	400	730	2300	3200	0	-3	-30	-35
u	270	615	2200	3150	0	-13	-40	-50
w	600	1700	2900	4100	-9	0	-2	-10
sz	-	2300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

Kompresja sygnału mowy

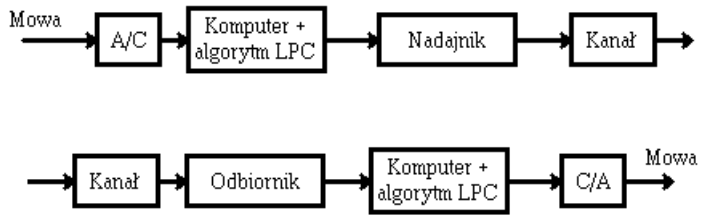
Wokodery - urządzenia służące do ograniczania objętości informacyjnej sygnału mowy metodą ekstracji parametrów i następnie po przesłaniu parametrów przez kanał telekomunikacyjny dokonujące resyntezy tego sygnału.



Struktura wokodera kanałowego (pasmowego)



Struktura wokodera formantowego



Struktura wokodera opartego na zasadzie predykcji liniowej

Podstawy automatycznego rozpoznawania mowy

Podstawy segmentacji sygnału mowy:

1. alfabet bazowy - dla mowy polskiej 37 fonemów
2. segmenty fonetyczne
   - odcinki o jednorodnej strukturze fonetycznej decydującej o przynależności do określonego fonemu
3. segmentacja stała
   - odcinki o stałej długości - kwazistacjonarne
   - "implicit segmentation" - mikrofonemy
4. segmentacja zmienna
   - segmenty zdefiniowane przez transkrypcję fonetyczną
   - "explicit segmentation" - dłuższe niż poprzednio
5. rodzaje segmentów dla sygnału mowy:
stacjonarne, transjentowe, krótkie, pauza.
6. granice segmentów:
   dźwięcznych - płynne przejścia formantów
   dźwięczny i bezdźwięczny - połączenie struktur formantowych i szumowych
   fonem i cisza - niepełna realizacja struktury widmowej

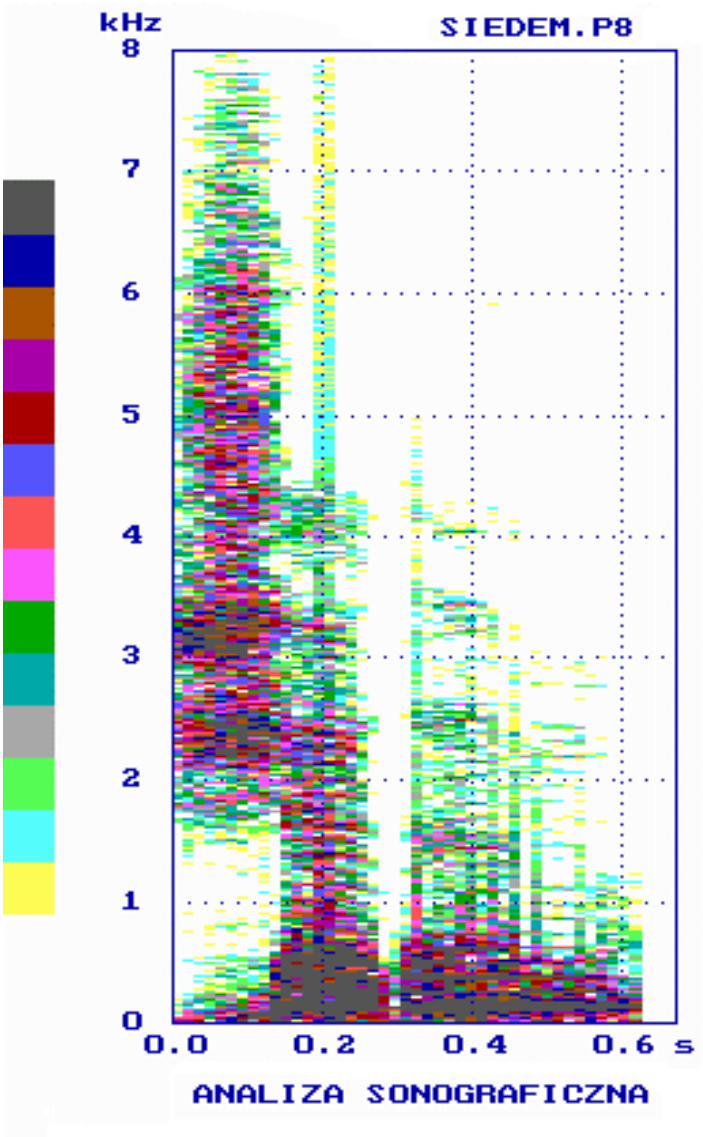
Wymagania:

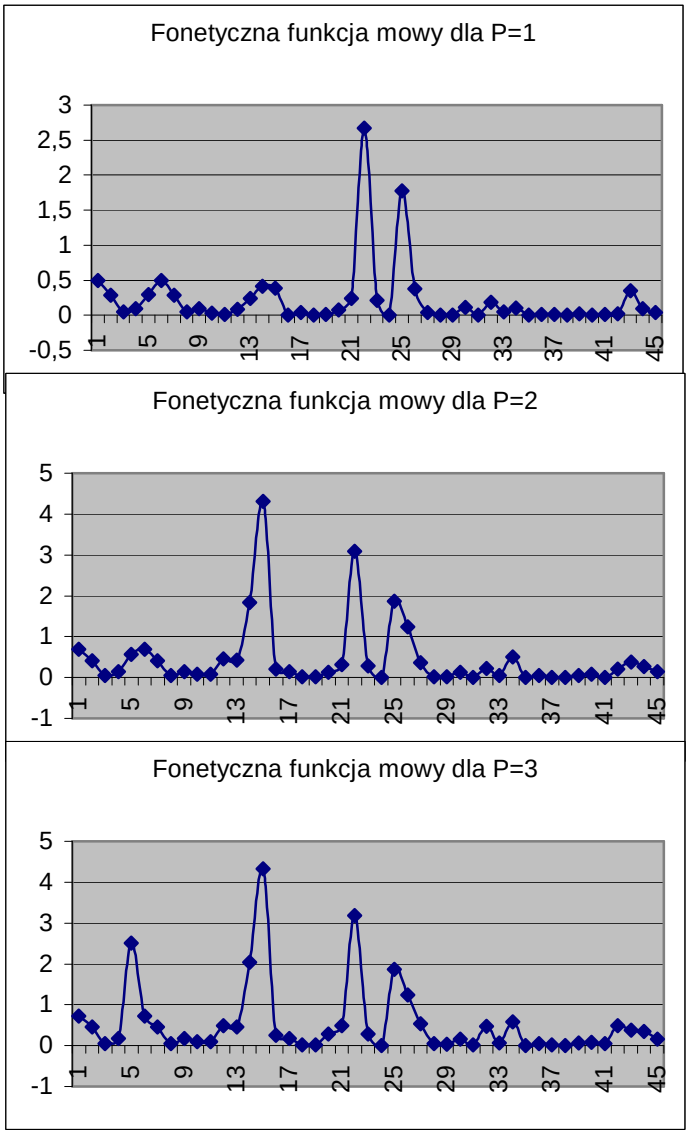
- algorytm segmentacji powinien generować funkcję czasu, na podstawie której można oznaczyć granice segmentów
- wybór metod parametryzacji
- kryteria podziału i wybór desygnatów znaczeniowych

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.





Porównanie wyników analizy sonograficznej z wynikami segmentacji dla różnych długości P wektora parametrów

Funkcje bloku segmentacji:
- parametryzacja (dla mikrofonemów)
- obliczenie fonetycznej funkcji mowy
- detekcja granic segmentów (maksima ffm)

Problemy:
- nie każde lokalne maksimum jest granicą segmentu (fitry wygładzające, algorytmy eksperckie),
- dobór wagi dla poszczególnych parametrów,
- dobór parametrów

METRYKI STOSOWANE W PRZESTRZENI PARAMETRÓW:

Euklidesa:

$$D(x,y)=\sqrt{\sum_{p=1}^{P}\left(x_p-y_p\right)^2}$$

gdzie:
x_p , y_p – wartość p-tego parametru dla porównywanych obiektów,
P – liczba parametrów,

Minkowskiego:

$$D(x,y)=\sqrt[r]{\sum_{p=1}^{P}\left|x_p-y_p\right|^r}$$

Hamminga (uliczna):

$$D(x,y)=\sum_{p=1}^{P}\left|x_p-y_p\right|$$

Euklidesa znormalizowana:

$$D(x,y)=\sqrt{\sum_{p=1}^{P}\frac{1}{S_p^{\,2}}\cdot\left(x_p-y_p\right)^2}$$

Camberra:

$$D(x,y)=\sum_{p=1}^{P}\frac{\left|x_p-y_p\right|}{\left|x_p+y_p\right|}$$

Czebyszewa:

$$D(x,y)=\max_{p}\left|x_p-y_p\right|$$

Mahalanobisa:

$$D(x,y)=\left(\bar{x}-\bar{y}\right)^T\cdot C^{-1}\cdot\left(\bar{x}-\bar{y}\right)$$

Funkcje bliskości:

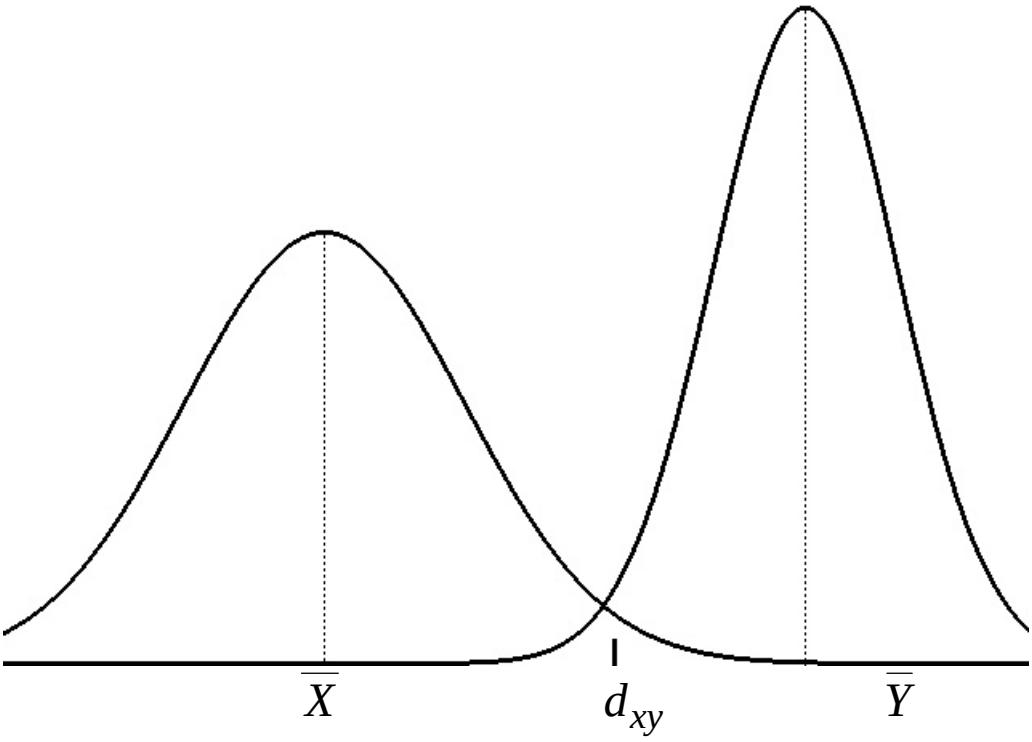
Kosinus kierunkowy:

$$B(x,y)=\frac{\bar{x}^T\bar{y}}{\|\bar{x}\|\cdot\|\bar{y}\|}$$

Tanimoto:

$$B(x,y)=\frac{\bar{x}^T\bar{y}}{\bar{x}^T\bar{x}+\bar{y}^T\bar{y}-\bar{x}^T\bar{y}}$$

Przykład jednowymiarowego optymalnego systemu dyskryminacji



Przy wyrównanym prawdopodobieństwie apriorycznym wartość dyskryminacyjna d_{xy} powinna spełniać zależność:

$$P(x>d_{xy})=P(y<d_{xy})$$

czyli:

$$\frac{1}{\sigma_1 \cdot \sqrt{2\pi}} \int_{d_{xy}}^{+\infty} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^{\,2}}\right) dx = \frac{1}{\sigma_2 \cdot \sqrt{2\pi}} \int_{-\infty}^{d_{xy}} \exp\left(-\frac{(x-\mu_2)^2}{2\sigma_2^{\,2}}\right) dx$$

zatem wartość dyskryminacyjna:

$$d_{xy} = \frac{\overline{X} \cdot S_2 + \overline{Y} \cdot S_1}{S_1 + S_2},$$

Normalizacja energetyczna (parametry czasowe – przebieg czasowy obwiedni energii, funkcja korelacji, gęstość przejść przez zero, interwały czasowe przejść przez zero, trajektorie czasowe innych parametrów)
i czasowa sygnału mowy (dynamiczne dopasowanie czasowe - time warping)

Segmentacja elementów fonetycznych i leksykalnych.
alofony, fonemy, diafony, sylaby, słowa

Metody parametryzacji mowy.
(prawdopodobieństwo średniego błędu rozpoznawania)

Separowalność parametrów.
- kryteria i metody oceny skuteczności parametrów:
1. macierze kowariancji (rozproszeń)
2. iloraz średniej odległości między klasami i średniego promienia odległości wewnątrz klas

redukcja przestrzeni parametrów
cel:
1. skrócenie etapu treningu
2. zwiększenie szybkości obliczeń klasyfikatora
3. obniżka kosztów

metody (transformacje liniowe):
1. rozwinięcie Karhunena-Loeve'go
2. rozwinięcie w szeregi funkcji ortogonalnych
3. analiza dyskryminacyjna Fishera

Pozostałe informacje nt. rozpoznawania mowy są zawarte:

http://sound.eti.pg.gda.pl/student/pdio/mowa.ppt

Materiały pomocnicze do zajęć ->
Przetwarzanie dźwięku i obrazu ->
Algorytmy komputerowego rozpoznawania mowy