Wykład 2/3. Analiza mowy – parametry opisu sygnału mowy, metody klasyfikacji

Język (szczególnie mowa) jest naturalnym i jednym z najbardziej efektywnych sposobów przekazywania myśli i odczuć człowieka. W ciągu dnia człowiek wypowiada do 40 000 słów.

Jako początek systemów automatycznego rozpoznawania mowy można uznać rok 1920, gdy wyprodukowano zabawkę „Rex". Był to mechaniczny pies, który po usłyszeniu swojego imienia wychodził z budy. Zabawka ta działała na zasadzie wykrywania energii w zakresie 500 Hz (odpowiada on pierwszemu formantowi głoski e). Za pierwszy „profesjonalny" system rozpoznawania mowy uznaje się system, który powstał na początku lat pięćdziesiątych XX wieku w laboratorium Bella. System umożliwiał rozpoznawanie 10 cyfr (wymagał uczenia przez konkretnego mówcę).

Badaniem struktury dźwiękowej języka zajmuje się fonetyka. Dokonuje ona selekcji dźwięków elementarnych, z których przez złożenie, powstają określone formy językowe. Dźwięki elementarne, zwane fonemami, stanowią akustyczną bazę dźwiękową. W języku polskim wyróżnia się 37 fonemów (przy uwzględnieniu dwóch samogłosek nosowych – 39), w języku angielskim 44.

Do zapisu formy dźwiękowej języka stosuje się transkrypcję z kodu literowego na kod dźwiękowy. Przykładami alfabetów fonetycznych są IPA i SAMPA.

Alfabet SAMPA został stworzony w 1995 roku przez londyńskiego profesora Johna C. Wellsa. Jego zaletą jest całkowita kompatybilność z kodem ASCII, co ułatwia pisanie na typowej klawiaturze komputera osobistego.

Transkrypcję fonetyczną mowy polskiej przedstawiono w tabeli 1.

Podstawowe terminy:

Fonem – minimalny segment dźwiękowy mowy, który może odróżniać znaczenie (inaczej: klasa dźwięków mowy danego języka o różnicach wynikających wyłącznie z charakteru indywidualnej wymowy lub kontekstu.

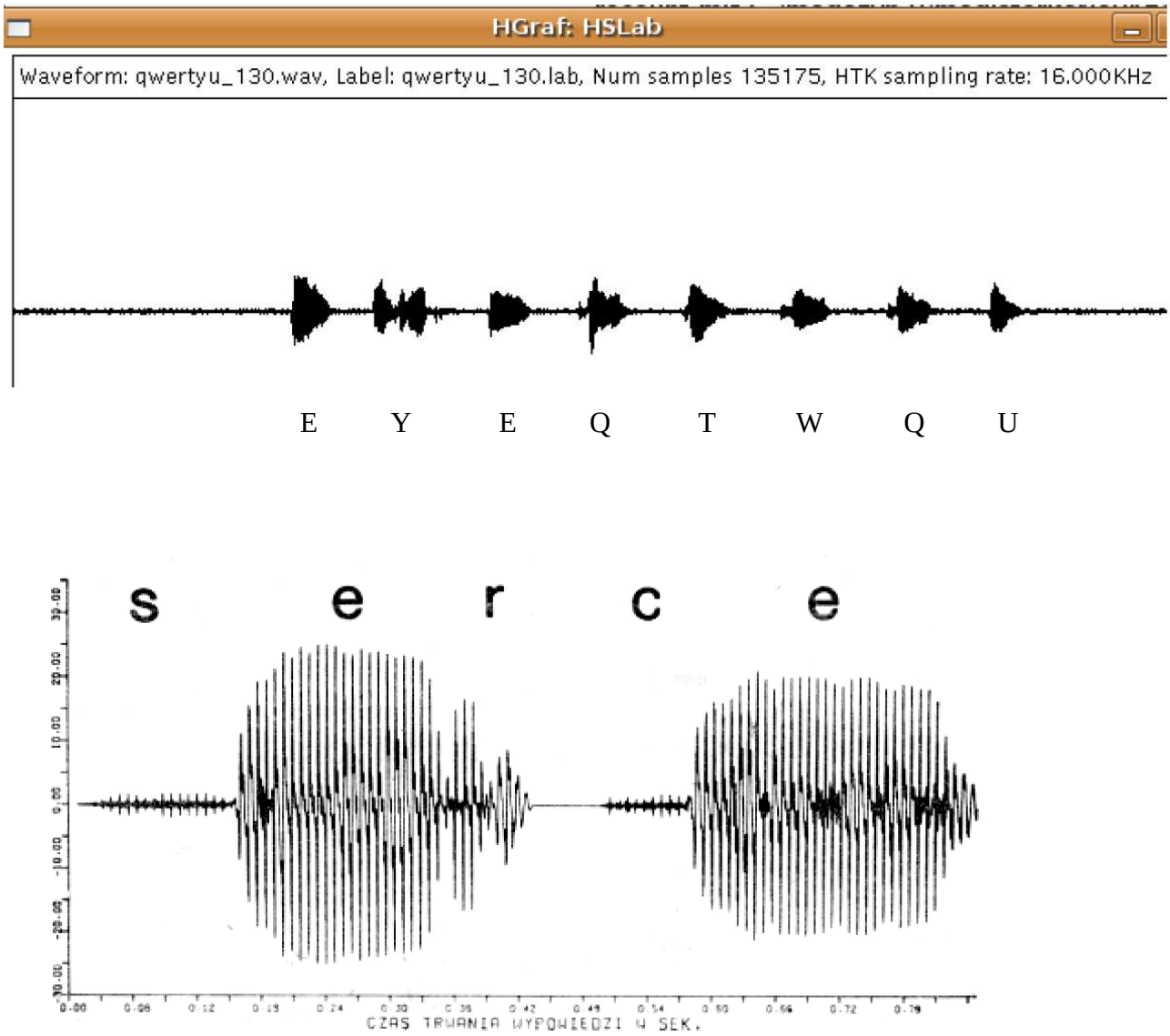
Alofon – wariant fonemu (głoski), różniący się od innego alofonu cechami fonetycznymi, a nie funkcją (np. głoska r posiada różne parametry akustyczne w wyrazach *fura*, *trawa* i *rok*).

Diafon – przejście międzyfonemowe (inaczej tranzem), które zawiera informację przejściową od środka kwazistacjonarnych odcinków par samogłoska-spółgłoska lub spółgłoska-samogłoska.

Tabela 1: Notacje fonemów języka polskiego [Ba92]

Lp.	Przykładowe użycie fonemu	IPA	SAMPA	Użyta
1	lis, PIT	i	i	i
2	ryba, typ	ɨ *	I	y
3	rzeka, test	e	e	e
4	rak, pat	a	a	a
5	bok, pot	o	o	o
6	kurs, puk	u	u	u
7	gęś	ẽ	e~	e~
8	wąs	õ	o~	o~
9	pas, pik	p	p	p
10	but, bit	b	b	b
11	but, test	t	t	t
12	kody, dym	d	d	d
13	rzeka, kit	k	k	k
14	noga, gen	g	g	g
15	kaftan, fan	f	f	f
16	wrak, wilk	v	v	w
17	pasek, syk	s	s	s
18	koza, zbir	z	z	z
19	szafa, szyk	ʃ	S	sz
20	żaba, żyto	ʒ	Z	rz
21	ktoś, świt	ɕ	s' (s\)	si
22	więź, źle	ʑ	z'	zi
23	chata, hymn	x	x	h
24	moc, cyk	t͡s	ts	c
25	rydze, dzwon	d͡z	dz	dz
26	kaczka, czyn	t͡ʃ	tS	cz
27	móżdżek, dżem	d͡ʒ	dZ	dż
28	ćwikła, ćma	t͡ɕ	ts'	ci
29	dziwak, dźwig	d͡ʑ	dz'	dzi
30	mur, mysz	m	m	m
31	nos, nasz	n	n	n
32	niski, koń	ɲ	n'	ni
33	sęk, pęk	ŋ	N	en
34	cykl, luk	l	l	l
35	ropa, ryk	r	r	r
36	miał, łyk	w	w	l~
37	kraj , jak	j	j	j

Sygnał mowy, traktowany jako przebieg czasowy ma skomplikowany przebieg, będący odzwierciedleniem złożonego charakteru jego artykulacji. Na parametry sygnału ma wpływ jego źródło (drgające wiązadła głosowe, turbulentny przepływ powietrza przez narządy mowy) i własności dynamiczne kanału głosowego, ostatecznie formującego strukturę sygnału. Przykładowe przebiegi czasowe sygnału mowy przedstawiono na rysunku 1.



Rys. 1. Przykładowe przebiegi czasowe sygnału mowy

Sygnał akustyczny mowy (traktowany jako źródło informacji) zawiera dwa typy informacji:

- informacje bezpośrednio dotyczące przekazywanej treści
- informacje dodatkowe (treści pozatekstowe).

Treści pozatekstowe związane są z cechami osobniczymi głosu, stanem emocjonalnym mówcy (np. zdenerwowanie), zjawiskami patologicznymi (np. chrypka, katar). Wpływ na postać sygnału akustycznego mogą mieć również wpływ zakłócenia pochodzące z otoczenia.

Prace dotyczące analizy mowy można podzielić na trzy obszary zainteresowań:

- automatyczne rozpoznawanie mowy (ang. – ASR, pol. ARM)
- automatyczne rozpoznawanie mówców (pol. ARG)
- synteza mowy

Automatyczne rozpoznawanie mowy to dziedzina wiedzy umożliwiająca realizację akustycznego wejścia do komputera, co pozwala na wprowadzenie informacji ludzkim głosem. W najprostszym przypadku działanie systemu polega na rozpoznaniu izolowanych słów należących do ściśle zdefiniowanego zbioru wprowadzanych przez określoną osobą.

Automatyczne rozpoznawanie mówców należy do biometrycznych metod rozpoznawania osób. Zadaniem systemu może być automatyczna identyfikacja osoby lub automatyczna weryfikacja (potwierdzenie lub odrzucenie) tożsamości osoby na podstawie wypowiedzi.

Synteza mowy umożliwia dwustronną komunikację głosową człowiek-komputer. Umożliwia wyprowadzanie informacji z komputera w postaci mowy np. systemy automatycznej konwersji tekstu na postać akustyczną (TTS – z ang. Text to Speech).

Struktura systemu rozpoznawania mowy:

W bloku wstępnego przetwarzania następuje odbiór sygnału mowy z mikrofonu oraz jego wstępne przetworzenie: wzmocnienie, filtracja (ograniczenie pasma od dołu i od góry) oraz przetworzenie na postać cyfrową (przetwornik a/c). W obecnych systemach funkcję tą spełnia karta dźwiękowa.

W bloku ekstrakcji parametrów następuje analiza sygnału mowy, w wyniku której otrzymuje się wartości parametrów niosących informację o treści wypowiedzi i niezależnych od indywidualnych cech głosu mówcy. Zbiór tych parametrów tworzy wektor (lub macierz) cech, na podstawie którego dokonuje się klasyfikacji sygnału.

W bloku klasyfikacji następuje porównanie nadchodzących ciągów obrazów wypowiedzi ze znajdującymi się w pamięci wzorcami, które stanowią uogólniony (uśredniony) opis klas dźwięków. Klasą mogą być fonemy, wyrazy lub nawet całe zdania. Obrazy wzorcowe są tworzone w procesie uczenia.

Cechy charakterystyczne sygnału mowy

W sygnale mowy wyodrębnia się pewne cechy charakterystyczne, których pomiar i dalsza analiza umożliwia określenie istotnych zbiorów parametrów wykorzystywanych w dalszych etapach przetwarzania i rozpoznawania mowy. Stosowane są następujące kryteria wyboru parametrów:

- skuteczność (zdolność rozróżniania dźwięków)
- łatwość pomiaru (złożoność procedur pomiarowych)
- stabilność, powtarzalność (mały zakres zmienności mierzonych parametrów)
- odporność na zakłócenia (otoczenie, tor transmisyjny np. łącze telefoniczne)

Parametry wykorzystywane do rozpoznawania mowy można podzielić na następujące grupy:

- parametry w dziedzinie czasu
- parametry w dziedzinie częstotliwości
- parametry liniowego kodowania predykcyjnego LPC

Parametry w dziedzinie czasu

Jest to grupa parametrów, które można określić bezpośrednio na podstawie struktury czasowej sygnału mowy. Parametry te można podzielić na dwie grupy:

- parametry będące funkcjami czasu
- parametry związane z pomiarem punktów, w których sygnał mowy zmienia znak (przejście sygnału mowy przez zero).

Do pierwszej grupy parametrów zaliczamy:

- natężenie sygnału w funkcji czasu

$$I(t) = a \int_{t-T/2}^{t+T/2} u^2(\tau)\, d\tau ,$$

gdzie: a – współczynnik normalizacji,

T – interwał uśredniania (szerokość okna pomiarowego)

- obwiednia amplitudy sygnału

$$O(t) = \sqrt{|u^2(t) + H^2[u(t)]}\ ,$$

gdzie: H- transformata Hilberta sygnału mowy

Dyskretny przebieg obwiedni O(t) obliczany jako uśredniona wartość w pewnych przedziałach czasu można przedstawić w postaci wektora:

$$O(t) = \{o_1, o_2, ..., o_N\}$$

Do drugiej grupy parametrów zaliczamy:

- gęstość przejść przez zero $\rho_0(t)$

$$\rho_0(t) = \{\rho_0(1), \rho_0(2), ..., \rho_0(N)\}$$

W celu określenia gęstości przejść przez zero należy zdefiniować funkcję przejścia przez zero P(u, t). Przebieg gęstości reprezentuje znormalizowany wektor wyznaczony dla N przedziałów czasowych.

- rozkład interwałów pomiędzy kolejnymi przejściami przez zero

Dla danego sygnału mowy określamy pewną funkcję R(t), która wyraża rozkład ciągły interwałów czasowych między kolejnymi przejściami przez zero w pewnym okresie czasu T :

$$(t_1, t_2, ..., t_j, ..., t_N)\ ,\ \text{gdzie}\ \sum_{j=1}^{N} t_j = T$$

Z przeprowadzonych doświadczeń wynika, że wartości graniczne interwałów czasowych wynoszą: $t_{min} = t_d = 0.2$ ms i $t_{max} = t_g = 6.2$ ms. Przedział $[t_d, t_g]$ dzielimy na P przedziałów czasowych, zwanych kanałami czasowymi.

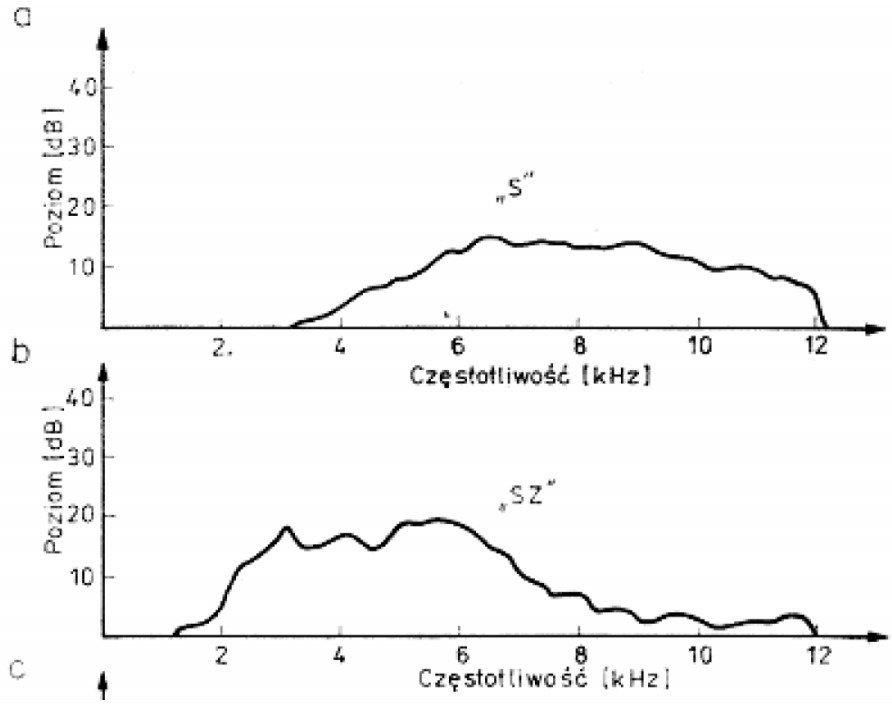
Rozkład interwałów czasowych pomiędzy przejściami przez zero otrzymujemy w wyniku zliczenia liczby interwałów t_j należących do poszczególnych kanałów czasowych. W wyniku otrzymujemy następujący wektor:

$$x(p) = \{x_1, x_2, ..., x_i, ..., x_P\}$$

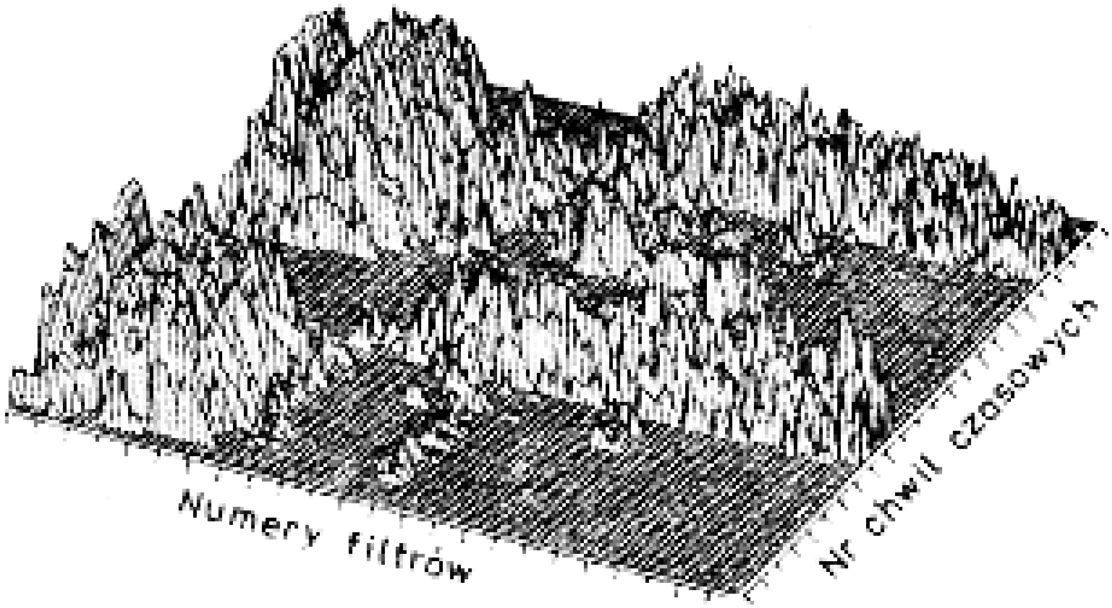
Parametry w dziedzinie częstotliwości

Podstawową formą prezentacji wizualnej sygnału mowy w dziedzinie częstotliwości jest trójwymiarowy obraz, który przedstawia amplitudę sygnału w funkcji czasu i częstotliwości.

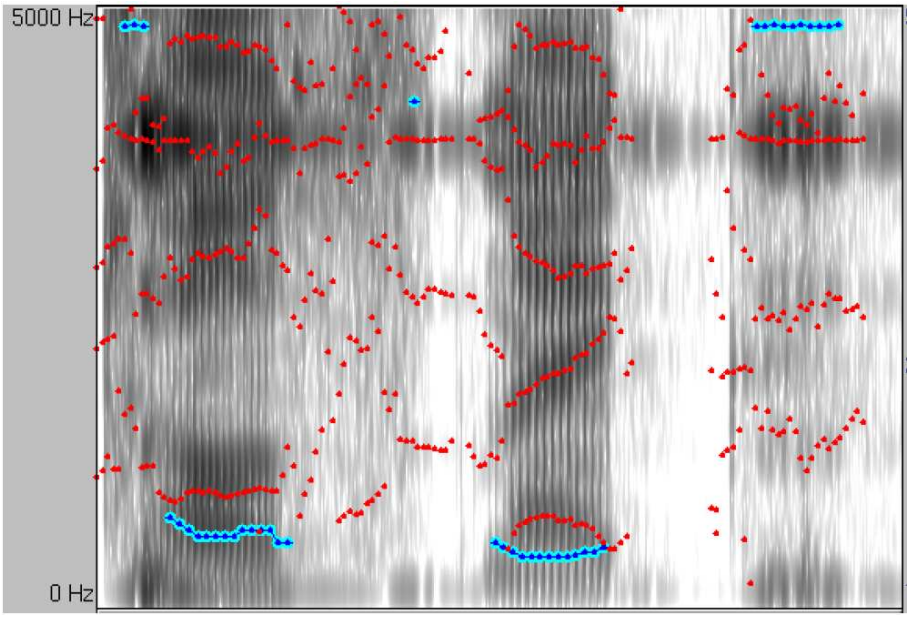
Poniższy rysunek przedstawia widmo pojedynczych głosek „s” i „sz”.



Kolejny rysunek przedstawia zmiany widma sygnału w kolejnych chwilach czasowych:



Obraz trójwymiarowy można przedstawić pa płaszczyźnie dwuwymiarowej (stopień szarości obrazu informuje o natężeniu sygnału – ciemniejszy obraz oznacza „silniejszy” sygnał):



Przedmiotem pomiaru jest głównie widmo amplitudowe lub energetyczne sygnału. W praktyce widmo amplitudowe (lub energetyczne) wyznaczane jest za pomocą analizatorów widma, które korzystają z filtrów analogowych, filtrów cyfrowych lub analizy cyfrowej (np. FFT – szybka transformata Fouriera).

Dyskretne przekształcenie Fouriera (DFT) prowadzi do zamiany sygnału wejściowego składającego się z N próbek (N jest najczęściej potęgą liczby 2):

x[0], x[1], x[2], ... , x[N-1]

na dwa sygnały wyjściowe, z których każdy zawiera N/2+1 wartości określających amplitudy odpowiednio składowej kosinusoidalnej i sinusoidalnej sygnału:

Re X[0], Re X[1], ... , Re [N/2] - amplitudy fali kosinusoidalnej

Im X[0], Im X[1], ... , Im [N/2] - amplitudy fali sinusoidalnej

Zależność między wektorem wejściowym oraz dwoma wektorami wyjściowymi można opisać wzorem (tzw. równanie syntezy):

$$x[i] = \sum_{k=0}^{N/2} \mathrm{Re}\,\overline{X}[k]\cos(2\pi ki / N) + \sum_{k=0}^{N/2} \mathrm{Im}\,\overline{X}[k]\sin(2\pi ki / N), \text{ dla } i = 0, 1, 2, \ldots, N-1$$

gdzie: Re $\overline{X}$ [k] = Re X[k] / (N/2), Im $\overline{X}$ [k] = Im X[k] / (N/2),

Re $\overline{X}$ [0] = Re X[0] / N, Re $\overline{X}$ [N/2] = Re X[N/2] / N (przypadki szczególne).

Stosowane są trzy metody obliczania transformaty DFT:
- metodą równań równoważnych,
- metodą korelacyjną,
- za pomocą szybkiego przekształcenia Fouriera (FFT).

Obraz sygnału mowy w dziedzinie częstotliwości jest obrazem złożonym i jego bezpośrednie wykorzystanie w procesie rozpoznawania jest właściwie niemożliwe. W związku z tym poszukuje się parametrów, które dobrze odzwierciedlają jego własności. Do podstawowych parametrów w dziedzinie częstotliwości zalicza się:

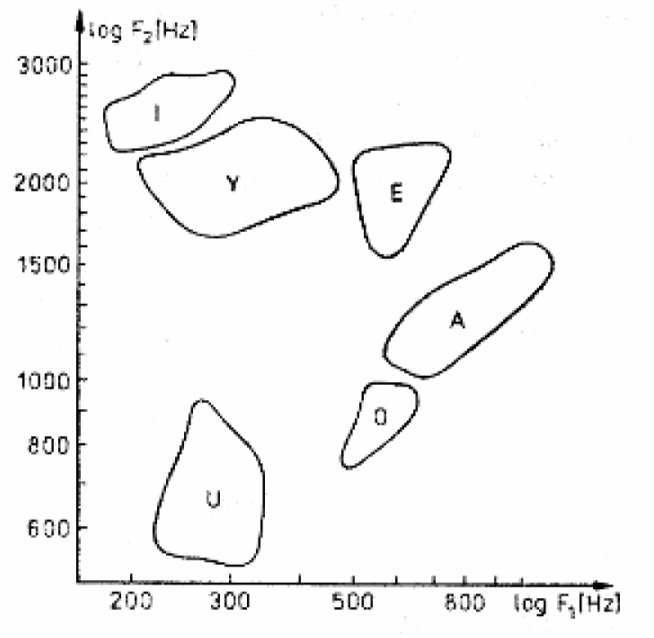
- częstotliwości formantowe (maksimum obwiedni)
- względne amplitudy poszczególnych formantów
- częstotliwości antyformantowe (minimum obwiedni)
- częstotliwość podstawowa tonu krtaniowego F_0
- wartości średnie widma liczone np. w pasmach tercjowych (1/23 oktawy)

Formant – wartość częstotliwości, dla której w widmie częstotliwości danego dźwięku występuje lokalne maksimum.

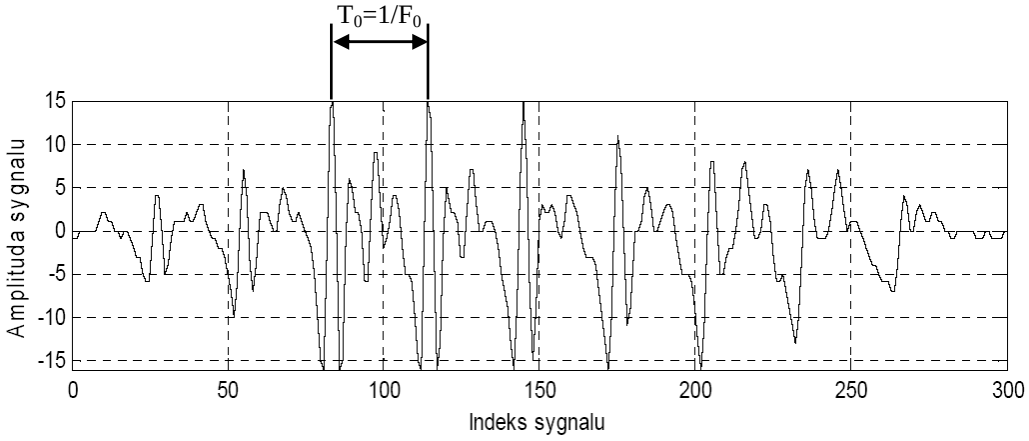
Częstotliwości formantowe określa się dla kolejnych przedziałów czasowych analizowanego widma (przyjmuje się przedziały 10-30 ms). Liczba określanych maksimów może być różna (przyjmuje się do 4). Pierwsze cztery maksima występują: w głosach męskich do częstotliwości 4 kHz, w głosach żeńskich do 5 kHz. Przykładowe średnie zakresy częstotliwości formantowych dla samogłosek przedstawia następująca tabela:

Fonem	F_1[Hz]	F_2[Hz]	F_3[Hz]	F_4[Hz}
a	700-900	1400-1750	2150-2450	-
e	450-700	1650-1850	2250-2700	-
o	500-650	950-1100	2250-2750	-
u	250-300	700- 850	2300-2550	-
i	230-320	2100-2450	2750-3200	-

Rozkład dwóch pierwszych częstotliwości formantowych dla samogłosek „polskich” przedstawia rysunek:



Częstotliwość podstawowa tonu krtaniowego F_0 (zwana również wysokością tonu) określa częstotliwość drgań wiązadeł głosowych. Odgrywa ona ważną rolę w procesie intonacji (oddaje np. charakter wypowiedzi – pytanie lub twierdzenie). Jest jednym z parametrów określających cechy osobnicze np. płeć osoby mówiącej.



<u>Analiza sygnału za pomocą banku filtrów</u>

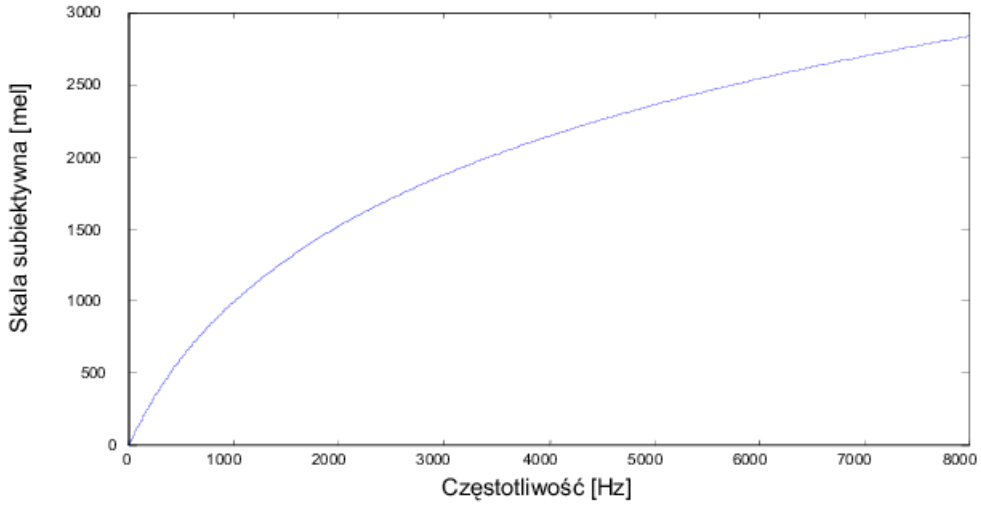
Ucho ludzkie w procesie percepcji sygnałów mowy dokonuje nieliniowej w dziedzinie częstotliwości analizy widma sygnału. Zastosowanie nieliniowego przetwarzania częstotliwości w systemach automatycznego rozpoznawania mowy prowadzi do zwiększenia ich skuteczności.

Otrzymane w wyniku transformacji FFT liniowe widmo sygnału przekształcane jest na mniejszą liczbę przedziałów (kanałów) za pomocą banku filtrów. Pierwszym etapem tego przekształcenia jest zmiana skali częstotliwości. W tym celu stosowana jest skala melowa (ang. mel-scale). Jednostką częstotliwości w tej skali jest mel. Zależność między mel i Hz ma charakter nieliniowy i określona jest wzorami:

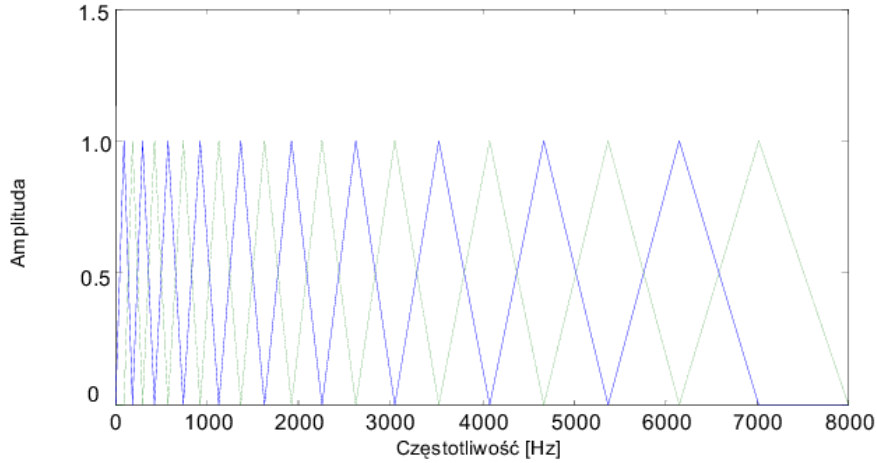
$$f_{mel}(f_{Hz}) = 2595 \log_{10} (1 + f_{Hz}/700)$$

$$f_{Hz}(f_{mel}) = 700 (10^{fmel/2595} -1)$$

Związek między liniową skalą częstotliwości (w Hz) a „fizjologiczną” skalą melową przedstawia rysunek:



Na bazie skali mel tworzony jest bank filtrów, który dokonuje nieliniowej analizy częstotliwościowej (podobnie jak ucho ludzkie). Filtry mają trójkątne charakterystyki amplitudowe i są równomiernie rozłożone w częstotliwościowej skali mel. Poniższy rysunek przedstawia bank 20 filtrów o szerokości pasma 300 mel przesuniętych względem siebie o 150 mel:



Najczęściej liczba kanałów w banku filtrów zawiera się w zakresie od 12 do 20.

W celu określenia parametrów opisujących sygnał mowy za pomocą banku filtrów należy:

- wyznaczyć widmo amplitudowe za pomocą przekształcenia Fouriera (np. FFT),
- oddzielnie dla każdego kanału w banku obliczyć sumę współczynników widma amplitudowego ważonych odpowiadającymi im wartościami charakterystyk amplitudowych filtru.

Dokonując następnie transformaty cepstranej otrzymanych parametrów (za pomocą przekształcenia kosinusowego) otrzymuje się tzw. współczynniki cepstralne (MFCC, z jęz. ang. Mel-Frequency Cepstral Coefficients).

Liniowe kodowanie predykcyjne LPC

Metoda ta polega na określeniu pewnych współczynników a_p (p = 1, 2, ..., P; P - rząd predykcji), zwanych współczynnikami predykcji, które umożliwiają wyznaczenie kolejnej wartości wybranego parametru sygnału mowy (np. amplitudy u(n)) na podstawie znajomości p wcześniejszych wartości. Jeżeli sygnał mowy przedstawimy jako ciąg kolejnych próbek:

$$u(n) = u(0), u(1), \ldots u(N-1)$$

wówczas kolejne wartości można otrzymać jako kombinację liniową P poprzednich próbek:

$$u'(n) = -\sum_{p=1}^{P} a_p u(n-p), \quad \text{dla } P < n < N-1$$

Efektywną metodę wyznaczania współczynników predykcji opracowali Levinson i Robinson. Na jej podstawie rekurencyjną procedurę wyznaczania współczynników przedstawił Durbin. Otrzymane współczynniki są następnie wykorzystywane w procesie klasyfikacji sygnału mowy.

Wybrane metody klasyfikacji

Proces automatycznego rozpoznania mowy polega na przyporządkowaniu analizowanego sygnału do odpowiedniej klasy. W zależności od przeznaczenia systemu rozpoznawanie może odbywać się na różnych poziomach: alofonów, fonemów, sylab, wyrazów lub nawet całych fraz (zdań). W większości tych systemów właściwa praca systemu (czyli rozpoznawanie) poprzedzona jest fazą uczenia. W wyniku uczenia powinniśmy otrzymać w przestrzeni parametrów rozłączne obszary odpowiadające poszczególnym klasom rozpoznawanych sygnałów.

Wprowadźmy następujące oznaczenia:

D – zbiór wszystkich obiektów podlegających rozpoznaniu,

D_i – podzbiór zbioru D reprezentujący klasę obiektów,

i – indeksy klas (i=1, 2, ..., N; N - liczba klas)

d_i – element zbioru D ($d_i \in D$)

X – przestrzeń cech

$\bar{x} = (x_1, x_2, ..., x_n) \in X$ – wektor parametrów opisujących sygnał mowy

Wówczas:

$$D = \bigcup_{i=1..N} D_i$$

$$\forall_{i,j \in 1..N} (D_i \cap D_j = \emptyset)$$

Proces rozpoznawania nieznanego elementu d_j polega na:

$f_1 : d_j \rightarrow (x_{j1}, x_{j2}, ..., x_{jn})$

$f_2 : (x_{j1}, x_{j2}, ..., x_{jn}) \rightarrow i$; i = 1.. N (numer klasy)

Do podstawowych metod rozpoznawania mowy zalicza się:

- algorytmy statystyczne parametryczne (np. algorytm Bayesa)
- algorytmy statystyczne nieparametryczne (np. algorytm NN – najbliższy sąsiad)
- algorytmy oparte o funkcję podobieństwa
- programowanie dynamiczne
- ukryte modele Markowa HMM
- sieci neuronowe

<u>Algorytm NN (najbliższy sąsiad)</u>

W algorytmie tym zostaje zapamiętany cały ciąg uczący U. Proces rozpoznawania polega na obliczeniu funkcji podobieństwa (np. odległości w przestrzeni parametrów X) między nieznanym obrazem $\mathbf{x}_j$ a wszystkimi obrazami ciągu uczącego $\mathbf{x}_k \in U$ (k =1, 2, ... , I; I – liczba elementów ciągu uczącego) i znalezieniu takiego obrazu $\mathbf{x}_k$ dla którego funkcja ta przyjmuje wartość największą (a odległość wartość najmniejszą). Obraz $\mathbf{x}_j$ zostaje zakwalifikowany do tej samej klasy co obraz $\mathbf{x}_k$. Jeżeli $d(\mathbf{x}_j, \mathbf{x}_k)$ oznacza odległość między wektorami x_j i x_k w przestrzeni X wówczas:

$$FP(\mathbf{x}_j, \mathbf{x}_k) = \frac{1}{d(x_j, x_k) + \varepsilon}$$

Szukamy x_k zgodnie z regułą:

$$x_k : \max_{x_k \in U} FP(\mathbf{x}_j, \mathbf{x}_k)$$

Do obliczenia odległości $d(\mathbf{x}_j, \mathbf{x}_k)$ stosowane są różne metryki:

$$d(\mathbf{x}_j, \mathbf{x}_k) = \sqrt{\sum_{i=1}^{n} (x_{ji} - x_{ki})^2} \quad \text{(metryka Euklidesa)}$$

$$d(\mathbf{x}_j, \mathbf{x}_k) = \sum_{i=1}^{n} |x_{ji} - x_{ki}| \quad \text{(metryka Hamminga)}$$

Zalety: duża prostota. **Wady:** konieczność pamiętania całego ciągu uczącego, dużo obliczeń, duża wrażliwość na błędy w ciągu uczącym.

Algorytm αNN

Jest to modyfikacja algorytmu NN, która ma zmniejszyć wrażliwość na błędy w ciągu uczącym. Podobnie jak w algorytmie NN obliczamy odległości obrazu x_j od wszystkich obrazów ciągu uczącego x_k. Następnie otrzymane odległości porządkujemy w kolejności rosnącej i wybieramy α pierwszych elementów otrzymanego ciągu (najczęściej α jest małą liczbą całkowitą). Wśród α pierwszych obrazów α_1 należy do pierwszej klasy, α_2 do drugiej, α_N do N-tej klasy. Obraz x_j zostanie przyporządkowany do tej klasy obiektów, dla której α_i jest największe.

Algorytm NM (najbliższa średnia)

Algorytm ten eliminuje konieczność pamiętania wszystkich obrazów ciągu uczącego. Wprowadza się pojęcie wzorca klasy określanego jako wartość średnia lub wartość modalna (najbardziej typowa dla danej klasy). Pamiętane są wówczas tylko parametry opisujące wzorce poszczególnych klas. Działanie algorytmu polega wówczas na obliczeniu funkcji podobieństwa (odległości) między rozpoznawanym obrazem a wzorcami poszczególnych klas, co znacznie zmniejsza liczbę obliczeń.