R. Gubrynowicz

ZASTOSOWANIE PASMOWEJ ANALIZY WIDMOWEJ DO OKREŚLANIA CECH OSOBNICZYCH GŁOSU

26/1968

WARSZAWA

N a p r a w a c h r ę k o p i s u

D o u ż y t k u w e w n ę t r z n e g o

Zakład Badania Drgań I P P T P A N .

Nakład 200 egz. Ark. wyd. 1, Ark. druk. 1,75.
Oddano do drukarni w październiku 1968 r.
Wydrukowano w listopadzie 1968r. Nr zam. 831/68

Warszawska Drukarnia Naukowa , Warszawa ,
ul.Śniadeckich 8

R. Gubrynowicz
Pracownia Elektroakustyki
Zakładu Badania Drgań IPPT PAN

ZASTOSOWANIE PASMOWEJ ANALIZY WIDMOWEJ SYGNAŁU MOWY DO OKREŚLANIA CECH OSOBNICZYCH GŁOSU

1. Wstęp

W ostatnim dziesięcioleciu w pracach badawczych w dziedzinie akustyki cybernetycznej wiele uwagi poświęca się problemowi automatycznego rozpoznawania głosów. Od rozwiązania tego zagadnienia, które samo w sobie ma ważne aspekty praktyczne /identyfikacja głosów dla celów kryminalistyki, konstrukcja urządzeń reagujących na głos określonego dyspozytora i.t.p./, uzależniony jest również w znacznej mierze dalszy postęp prac prowadzonych nad automatycznym rozpoznawaniem elementów fonetycznych i lingwistycznych mowy, na płaszczyźnie fonemów, wyrazów lub zdań. Należy jednak podkreślić, że wszystkie opracowane do tej pory układy eksperymentalne do rozpoznawania elementów segmentalnych mowy działają skutecznie wyłącznie dla ograniczonej liczby głosów, przy czym liczba ta zależna jest od ilości przekazywanych informacji; im zawartość informacji fonetycznych lub lingwistycznych w sygnale mowy jest większa, tym mniej-

sza jest liczba głosów, dla których rozpoznawanie jest prawidłowe.

O ile subiektywne rozpoznawanie głosów nie przedstawia naogół większych trudności, zwłaszcza przy słyszeniu bezpośrednim, o tyle automatyzacja tego procesu stanowi bardzo skomplikowany problem. Przyczyną tego jest fakt, że sygnał mowy jest nośnikiem trzech rodzajów informacji, łatwo rozróżnialnych w sposób subiektywny, ale trudnych do rozdzielenia przy stosowaniu obiektywnych metod analizy technicznej. Informacje zawarte w sygnale mowy są następujące [1] :

a/ l i n g w i s t y c z n e , które określają treść przekazywanej wiadomości;

b/ s o c j o l i n g w i s t y c z n e , które pozwalają ustalić przynależność nadawcy wiadomości do określonej grupy etnograficznej, społecznej i środowiskowej;

c/ o s o b n i c z e , które umożliwiają rozpoznanie indywidualnych cech głosu nadawcy wiadomości.

Podczas procesu percepcji mowy słuchacz potrafi bez większego trudu wyeliminować z odbieranego sygnału akustycznego pierwsze dwa rodzaje informacji i skoncentrować uwagę na informacjach osobniczych, umożliwiających mu prawidłowe zidentyfikowanie głosu nadawcy wiadomości. Natomiast w procesie automatycznego rozpoznawania głosów eliminowanie zbędnych w tym przypadku informacji lingistycznych i socjolingwistycznych musi być dokonywane przy pomocy specjalnych metod [2]. Jedna z nich polega na ustaleniu /stabilizacji/

informacji lingwistycznych zawartych w analizowanym sygnale mowy, który zgodnie z założeniem odpowiada jednemu i temu samemu elementowi segmentalnemu mowy /np. określonemu wyrazowi kodowemu/. Inna metoda, w odróżnieniu od poprzedniej , opiera się na analizie parametrów statystycznych, niezależnych od treści lingwistycznej zawartej w sygnale mowy. Parametry te zostają uśrednione nie dla określonych segmentów mowy, lecz dla dowolnego tekstu o długości większej od pewnej minimalnej wartści. W tym przypadku sygnał mowy może być traktowany jako stacjonarny proces stochastyczny. Takie ujęcie zagadnienia jest bardziej ogólne i naturalne, a jednocześnie lepiej nadaje się do zastosowań praktycznych obejmujących opracowanie układów do automatycznego rozpoznawania lub selekcji głosów.

Praktyczna realizacja takiego układu wymaga rozwiązania szeregu problemów ubocznych, o charakterze podstawowym, które dotyczą wyboru odpowiedniego zespołu parametrów akustycznych sygnału mowy, najbardziej efektywnych z punktu widzenia zawartości informacji osobniczych. Odrębnym zagadnieniem jest zbadanie, jaki materiał fonetyczny na płaszczyźnie określonego języka jest najbardziej dogodny dla identyfikacji cech osobniczych głosu. Problemy te będą szerzej omówione w dalszej części pracy, której przedmiotem jest zastosowanie statystycznej metody analizy widmowej sygnału mowy do określania cech osobniczych głosu nadawcy wiadomości.

2. Opis metody pomiarowej

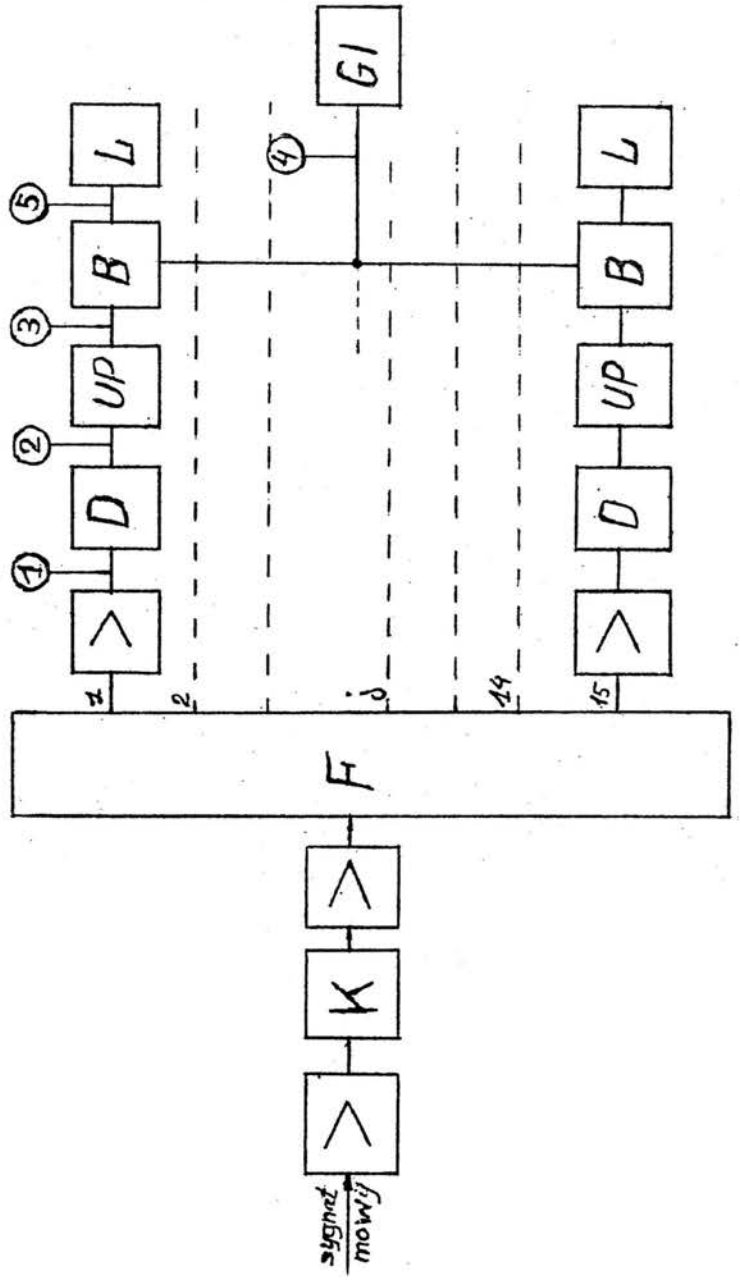
2. 1. Założenia ogólne.

Metoda oparta jest na założeniu, że w strukturze anatomicznej organu mowy danego osobnika, określającej zarówno charakterystykę źródła tonu krtaniowego, jak i funkcję transmitancji kanału głosowego i nosowego, zawarte są cechy indywidualne głosu, umożliwiające jego identyfikację [3] [4]. Podczas procesu artykulacji efektory artykulacyjne organu mowy są w ustawicznym ruchu, wskutek czego parametry akustyczne aparatu głosowego muszą być traktowane jako funkcje czasu. W istocie zmiany konfiguracji geometrycznej organu mowy określają treść fonetyczną zawartą w wypowiedzi.

Uśredniając jednak te parametry w odpowiednio długim okresie czasu można otrzymać ich wartości średnie, które zależą jedynie od indywidualnej struktury anatomicznej organu mowy nadawcy wiadomości oraz od struktury fonetycznej języka użytego przez niego do przekazywania informacji. Opierając się na tym założeniu można przyjąć, że widmo amplitudowe sygnału akustycznego, wyemitowane podczas określonej wypowiedzi i uśrednione za czas jej trwania, odzwierciedla jedynie fizjologiczne i anatomiczne cechy aparatu głosowego, a więc tym samym, że w widmie tym zawarte są głównie informacje osobnicze charakteryzujące nadawcę wiadomości.

2. 2. Metodyka badań eksperymentalnych.

W celu doświadczalnego sprawdzenia tej hipotezy zaprojektowano i opracowano układ pomiarowy, umożliwiający prze-

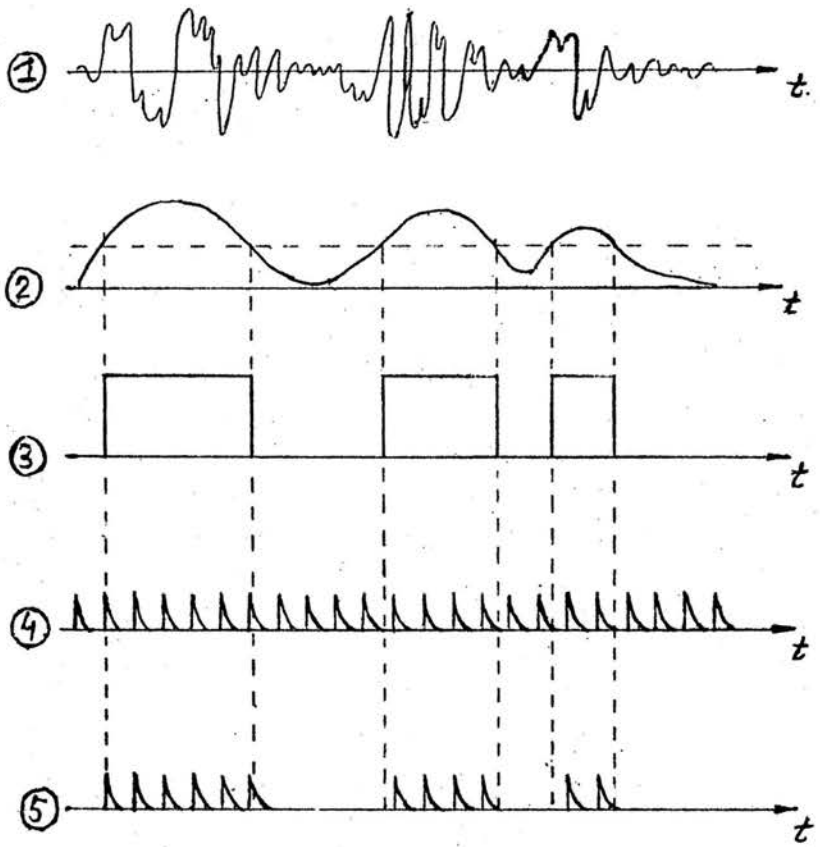


Rys. 1. Schemat blokowy układu pomiarowego.

prowadzenie pasmowej analizy widmowej sygnału mowy z bezpośrednim przedstawieniem wyników w postaci liczbowej. Schemat tego układu podano na rys. 1.

Zasada działania tego układu jest następująca. Sygnał mowy po wzmocnieniu i kompresji, normalizującej jego amplitudę, podawany jest na zestaw $\underline{n = 15}$ filtrów tercjowych, pokrywających zakres częstotliwości od 160 Hz do 4 000 Hz i oznaczonych kolejnymi numerami od 1 do 15. Sygnały wyjściowe poszczególnych filtrów zestawu po wzmocnieniu są prostowane dwupołówkowo w układach detekcji D, a otrzymane napięcia stałe, o ile przekraczają wartość progową, wyzwalają układy progowe UP, które z kolei otwierają bramki B, przepuszczające impulsy z generatora wzorcowego GI do układów zliczających L. Zliczanie impulsów w każdym paśmie częstotliwości odbywa się jedynie w tych okresach czasu, w których w odpowiadających danemu pasmu przekroczony jest poziom wyzwalania układu progowego. Dla dokładniejszego zobrazowania działania analizatora, na rys.2 przedstawiono przebiegi napięć w poszczególnych punktach układu z rys. 1, oznaczonych cyframi arabskimi.

W ten sposób po analizie w $\underline{n = 15}$ kolejnych pasmach tercjowych otrzymuje się dla każdej wypowiedzi zestaw $\underline{n}$ liczb x_j /j = 1, 2, ..., n/ odpowiadających zliczonym w każdym kanale impulsom w czasie nadawania określonego tekstu. Każdą wypowiedź można więc opisać za pomocą wektora o $\underline{n}$ współrzędnych w postaci:



Rys. 2. Przebiegi napięć w poszczególnych punktach układu analizatora z rys. 1, oznaczonych cyframi arabskimi.

$$W_i^{(k)} = \left(x_{1i}^{(k)}, x_{2i}^{(k)}, \ldots, x_{ji}^{(k)}, \ldots, x_{ni}^{(k)} \right) \quad /1/$$

dla k-tego głosu oraz i-tej wypowiedzi.

Przedstawienie wyników analizy widmowej w postaci liczbowej ma wiele zalet, a przede wszystkim jest bardzo dogodne przy dalszej obróbce danych, zwłaszcza przy stosowaniu matematycznych maszyn cyfrowych.

Należy z kolei rozważyć, jakie czynniki wpływają na wielkość $x_{ji}^{(k)}$, czyli na liczbę impulsów zliczonych w poszczególnych kanałach analizatora. Wartości współrzędnych niewątpliwie zależą od długości tekstu, przy czym zależność ta może nie mieć charakteru liniowego, ponieważ na ilość impulsów zliczonych w ciągu danego okresu czasu wpływa statystyczna struktura fonetyczna zastosowanego tekstu. Jeśli przy zmianie długości tekstu struktura ta nie ulega zmianie, to znaczy, jeżeli częstość występowania poszczególnych głosek pozostaje nadal taka sama, wówczas można przyjąć, że zależność liczby impulsów $x_{ji}^{(k)}$ od długości tekstu jest w przybliżeniu liniowa. Jeśli ponadto stosuje się teksty fonetycznie zrównoważone, w których głoski występują z taką samą częstością jak w mowie potocznej, wówczas wybór tekstu może być dowolny. Oczywiście rola pozostałych czynników wpływających na wartość $x_{ji}^{(k)}$, a mianowicie: sposobu artykulacji /zwłaszcza szybkości mówienia/ oraz cech indywidualnych aparatu głosowego, określonych jego strukturą anatomiczną, nie zostaje w ten sposób wyeliminowana i mogą one być wykorzystane w procesie identyfikacji. Stosując dostatecznie długi

tekst wybrany z mowy potocznej, można przyjąć , że jest on fonetycznie zrównoważony. Wynika stąd, że jeżeli czas trwania wypowiedzi przekracza pewną wartość graniczną t_{min}, to widmo sygnału mowy uśrednione za czas $t \geqslant t_{min}$ jest dla danego osobnika stałe i zawiera w sobie istotne informacje osobnicze.

W celu uniezależnienia wartości współrzędnych $x'^{(k)}_{ji}$ od długości tekstu, a także częściowo od szybkości mówienia, wprowadzono transformację $\underline{T}$ współrzędnych wektorów $W'^{(k)}_i$, którą przy założeniu liniowej zależności $x'^{(k)}_{ji}$ od czasu trwania wypowiedzi można zapisać w następujący sposób:

$$\underline{T} : W'^{(k)}_i \longrightarrow W^{(k)}_i = \left(x^{(k)}_{1i}, x^{(k)}_{2i}, \ldots x^{(k)}_{ji}, \ldots x^{(k)}_{ni} \right), \qquad /2/$$

gdzie

$$x^{(k)}_{ji} = \frac{x'^{(k)}_{ji}}{\sum_{j=1}^{n} x'^{(k)}_{ji}}$$

Dla przykładu w Tablicy 1 podano składowe wektorów $W'^{(k)}_i$ oraz $W^{(k)}_i$ o $\underline{n=5}$ współrzędnych, otrzymane dla i=4 wypowiedzi tego samego tekstu [1] czytanego przez jedną osobę z różną szybkością: dość szybko /wersja szybka/ i dość wolno /wersja wolna/. Ocena szybkości czytania dokonywana była w sposób subiektywny.

Z otrzymanych danych wynika, że o ile przed zastosowaniem

---

[1] Tekst stanowił wycinek komunikatu o treści ogólnej, wybranego z popularnego dziennika.

Tablica 1. Wpływ szybkości mówienia na wartości współrzędnych wektorów $W_i^{(k)}$ oraz $W_i^{(k)}$ dla n=5. Głos WM /k=1/.

a/ Wektor $W_i^{(k)}$

Wersja tekstu	$x_{1i}^{(1)}$	$x_{4i}^{(1)}$	$x_{7i}^{(1)}$	$x_{8i}^{(1)}$	$x_{15i}^{(1)}$
szybka i = 1	7 749	15 678	13 168	4 620	8 423
wolna i = 2	14 658	23 364	18 145	9 193	14 127
szybka i = 3	11 592	16 213	12 292	3 312	8 716
wolna i = 4	15 192	22 603	16 211	7 530	12 634

b/ Wektor $W_i^{(k)}$

Wersja tekstu	$x_{1i}^{(1)}$	$x_{4i}^{(1)}$	$x_{7i}^{(1)}$	$x_{8i}^{(1)}$	$x_{15i}^{(1)}$
szybka i = 1	0,155	0,316	0,268	0,093	0,170
wolna i = 2	0,184	0,298	0,238	0,116	0,178
szybka i = 3	0,222	0,312	0,236	0,064	0,168
wolna i = 4	0,205	0,305	0,219	0,100	0,170

transformacji $\underline{T}$ układu współrzędnych zmiany wartości tego samego parametru dla różnych szybkości mówienia osiągają nawet wartości rzędu 200%, to w nowym układzie współrzędnych zmiany te w najniekorzystniejszym przypadku nie przekraczają 80%.

Jeśli chodzi o wpływ pozostałych czynników artykulacyjnych /np. częstotliwości tonu krtaniowego, akcentu, dynamiki, itp/, to mimo że w krótkich odcinkach czasu mogą one ulegać dość znacznym, przypadkowym zmianom, można przyjąć, że dla dostatecznie długiej wypowiedzi sposób artykulacji jest niezmienny, a nawet typowy dla danego osobnika /akcent, cechy nawykowe, charakterystyczne defekty wymowy, itp./.

3. Problem tekstu wypowiedzi

Jak już wspomniano na wstępie, długość tekstu wypowiedzi ma istotny wpływ na sposób rozpoznawania głosów. Jest to związane z wyeliminowaniem z sygnału mowy zbędnych informacji lingwistycznych, utrudniających obiektywne rozpoznawanie głosów. Również od długości tekstu zależy metoda ekstrakcji parametrów charakteryzujących głos nadawcy wiadomości. Im krótszy jest nadany tekst, tym dokładniejsze muszą być stosowane metody analizy. Dla krótkich tekstów stosowane są na ogół metody typu sonograficznego /por. np. [5] [4] [3] [8] /. Natomiast gdy tekst oparty na mowie potocznej jest dostatecznie długi, to dla celów identyfikacji treść jego może być dowolna i nie musi być taka sama jak treść wypowiedzi, dla których utworzono wektory wzorcowe reprezentu-

jące poszczególne głosy. Ponadto można stosować prostsze metody analizy, otrzymując przy tym wystarczającą do identyfikacji ilość informacji, co wpływa w istotny sposób na szybkość procesu rozpoznawania.

Jednakże ilość informacji osobniczych, zawartych w nadanym sygnale mowy, zależy nie tylko od długości tekstu, ale i od jego struktury fonetycznej. Z subiektywnych badań odsłuchowych przeprowadzonych przez A. G. Kakauridze i G. S. Ramiszwiliego [6], podczas których słuchacze identyfikowali głos nadawcy na postawie wypowiedzianej izolowanej głoski, wynika, że wprawdzie wszystkie dźwięki mowy zawierają informacje osobnicze, ale zawartość tych informacji w poszczególnych głoskach nie jest jednakowa. Przykłady odpowiednich danych liczbowych przytoczone są w tablicy 2.

Opierając się na powyższej tablicy można wyciągnąć kilka wniosków o ogólnym charakterze:

a/ Głoski dźwięczne dają lepszą rozpoznawalność głosu niż głoski bezdźwięczne. Porównanie głosek dźwięcznych [z], [ʒ] i [g] z ich odpowiednikami bezdźwięcznymi [s], [ʃ] i [k] wskazywałoby na ważną rolę składowych harmonicznych w procesie identyfikacji głosu.

b/ Ze wszystkich głosek dźwięcznych samogłoski niosą najwięcej informacji osobniczych. Wyjątek stanowi samogłoska [u], której obwiednia widma powyżej 900 Hz gwałtownie opada.

Z tych danych wynika więc, że przy ustalonej długości tekstu najwięcej informacji osobniczych będzie zawierała ta-

Tablica 2. Procentowa rozpoznawalność subiektywna głosu w zależności od wypowiedzianej głoski [6].

Głoska	Rozpoznawalność w %	Głoska	Rozpoznawalność w %	Głoska	Rozpoznawalność w %
[e]	90	[r]	78	[t]	54
[o]	86	[v]	76	[t͡s]	5o
[l]	84	[ʒ]	74	[x]	48
[a]	83	[m]	62	[s]	44
[i]	83	[g]	61	[ʃ]	37
[z]	79	[u]	60	[k]	30

ka wypowiedź, podczas której zostanie nadana możliwie jak największa liczba głosek dźwięcznych. Wówczas dla przyjętej metody pomiarowej i dla danego zbioru rozpoznawanych głosów można dobrać minimalną długość tekstu zapewniającą prawidłową identyfikację głosów.

W oparciu o wyniki cytowanych autorów [6] postanowiono we wstępnym etapie badań sprawdzić, czy dla ograniczonego zbioru głosów możliwa by była ich identyfikacja w oparciu o nadawany tekst złożony z sześciu samogłosek sylabicznych, przy zastosowaniu pasmowej analizy widmowej sygnału mowy. W wyniku badań przeprowadzonych dla $k = 15$ głosów męskich /przy czym każdy osobnik powtórzył tekst $i = 3$ razy/ stwierdzono, że wartości współrzędnych $x_{ji}^{(k)}$ są dla poszczególnych głosów wystarczająco zróżnicowane dla celów identyfikacji,

co w pewnym stopniu potwierdza wnioski z cytowanej już pracy. Jednakże w celu pełniejszego zbadania wpływu struktury fonetycznej tekstu na zróżnicowanie wartości współrzędnych $x_{ji}^{(k)}$ postanowiono przeanalizować wypowiedzi dwóch list wyrazowych. Obie składały się z pięciu wyrazów, przy czym w jednej z nich dominowały głoski dźwięczne /ok. 90%/, w drugiej zaś głoski bezdźwięczne /ok. 60%/. W tablicy 3 podano wartości niektórych współrzędnych $x_{ji}^{(k)}$ otrzymanych dla pięciu głosów wypowiadających obie listy wyrazowe trzykrotnie.

Z podanych tabel wynika, że dla listy wyrazowej A /o przewadze głosek dźwięcznych/ otrzymuje się bardziej zróżnicowane wartości parametrów $x_{ji}^{(k)}$ dla poszczególnych głosów, niż dla listy wyrazowej B /o przewadze głosek bezdźwięcznych/, co potwierdza tezę, że głoski dźwięczne niosą więcej informacji osobniczych niż głoski bezdźwięczne. Ponadto zwraca uwagę dość duży rozrzut wyników w obrębie jednego głosu, co jest spowodowane zbyt małą długością tekstów, wskutek czego wpływ chwilowych, przypadkowych zmian artykulacyjnych na ostateczne wartości parametrów $x_{ji}^{(k)}$ jest znaczny. Dokładniej zagadnieniem wpływu długości tekstu na rozrzut wartości parametrów zajmował się m.in. G. S. Ramiszwili [2]. W wyniku badań stwierdził on, że przy zastosowanej przez niego metodzie pomiaru przedziałów czasowych między kolejnymi przejściami przez zero, sygnał mowy można traktować jako proces stacjonarny, jeśli wypowiedź trwa dłużej niż dwie minuty /tekst dowolny/. W przypadku opisanej tu metody pasmowej

Tablica 3.

A. Wartości wybranych współrzędnych $x_{ji}^{(k)}$ otrzymane dla listy wyrazowej A o przewadze głosek dźwięcznych.

Głos \ j (częst. śr. pasma)	1 /160Hz/	4 /315Hz/	6 /500Hz/	10 /1250Hz/	12 /2000Hz/	15 /4000Hz/
JRN	0,0800	0,0706	0,0752	0,0885	0,0325	0,0344
	0,0657	0,0759	0,0759	0,0923	0,0349	0,0188
	0,0742	0,0730	0,0895	0,0938	0,0259	0,0228
JK	0,0563	0,06o2	0,0678	0,0778	0,0020	0,0500
	0,0589	0,0642	0,0765	0,0749	0,0022	0,0437
	0,0571	0,0581	0,0781	0,0766	0,0022	0,0426
JM	0,0422	0,0267	0,0805	0,0894	0,0516	0,0367
	0,0388	0,0420	0,0965	0,0994	0,0559	0,0360
	0,0508	0,0288	0,0855	0,0954	0,0543	0,0261
RG	0,0087	0,0580	0,0790	0,0847	0,0280	0,0689
	0,0112	0,0481	0,0751	0,0849	0,0292	0,0724
	0,0082	0,0525	0,0756	0,0849	0,0263	0,0696
WM	0,0533	0,0852	0,1020	0,0946	0,0347	0,0044
	0,0648	0,0907	0,1020	0,0953	0,0395	0,0042
	0,0620	0,0815	0,0970	0,0876	0,0354	0,0040

Tablica 3.

B. Wartości wybranych współrzędnych $x_{ji}^{(k)}$ otrzymane dla listy wyrazowej B o przewadze głosek bezdźwięcznych.

Głos \ j (częst. śr. pasma)	1 /160Hz/	4 /315Hz/	6 /500Hz/	10 /1250Hz/	12 /2000Hz/	15 /4000Hz/
JNR	0,0734 0,0654 0,0802	0,0358 0,0215 0,0230	0,0583 0,0546 0,0664	0,0598 0,0518 0,0503	0,0760 0,0760 0,0674	0,1180 0,1250 0,0920
JK	0,0601 0,0584 0,0656	0,0281 0,0315 0,0292	0,0590 0,0645 0,0668	0,0454 0,0452 0,0448	0,0804 0,0764 0,0798	0,1290 0,1290 0,1255
JM	0,0634 0,0573 0,0700	0,0049 0,0017 0,0028	0,0731 0,0681 0,0740	0,0504 0,0468 0,0542	0,0886 0,0848 0,0826	0,1360 0,1401 0,1400
RG	0,0078 0,0140 0,0108	0,0440 0,0345 0,0361	0,0587 0,0590 0,0587	0,0503 0,0485 0,0460	0,0734 0,0687 0,0667	0,1340 0,1390 0,1370
WM	0,0303 0,0518 0,0477	0,0139 0,0096 0,0091	0,0675 0,0731 0,0727	0,0473 0,0476 0,0461	0,0797 0,0735 0,0823	0,1330 0,1370 0,1370

analizy widmowej, połączonej z transformacją współrzędnych, ze wstępnych pomiarów wydaje się, że do prawidłowego rozpoznawania głosów wystarczy dowolny tekst o długości przekraczającej jedną minutę. Problem ten, który został tutaj tylko zasygnalizowany, będzie jeszcze przedmiotem dokładniejszych badań.

4. Problem kryterium rozpoznawania

Z dotychczasowych rozważań wynika, że w rezultacie przeprowadzonej obróbki sygnału mowy dla każdej wypowiedzi otrzymuje się zestaw liczb, który opisuje położenie pewnego punktu w przestrzeni n-wymiarowej lub co jest jednoznaczne, przedstawia sobą współrzędne wektora łączącego ten punkt z początkiem układu współrzędnych. W ten sposób dla wielokrotnych wypowiedzi różnymi głosami, należącymi do rozpatrywanego zbioru, przy odpowiednim dobraniu układu współrzędnych otrzymuje się pewne zgrupowania punktów /wektorów/ odpowiadające poszczególnym głosom.

Ogólnie biorąc każdy proces rozpoznawania musi być poprzedzony etapem uczenia, w czasie którego dla danego zbioru głosów /t.j. klas rozpatrywanego zbioru/ na podstawie skończonej ilości wypowiedzi /realizacji/ uzyskuje się informacje o rozmieszczeniu obszarów odpowiadających poszczególnym klasom i ich strukturze statystycznej. Uśredniając współrzędne wektora wszystkich realizacji danej klasy otrzymuje się dla niej realizację wzorcową, która może być przed-

stawiona również jako punkt w przestrzeni n-wymiarowej lub jako wektor. Problem uczenia i tworzenia realizacji wzorcowych dla poszczególnych klas stanowi osobne, bardzo szerokie zagadnienie, wykraczające poza ramy niniejszej pracy. Syntetyczne ujęcie tego problemu znaleźć można w pracy N. Nilsona [7].

W geometrycznych metodach rozpoznawania, to znaczy takich, w których każdą realizację przedstawia się w postaci pewnego zestawu współrzędnych w przestrzeni n-wymiarowej, mogą być stosowane różne kryteria decyzyjne. Na przykład proces rozpoznawania może polegać na wyliczaniu odległości między punktem /końcem wektora/ odpowiadającemu analizowanej, nieznanej wypowiedzi, a punktami /końcami wektorów/ odpowiadającymi wzorcom poszczególnych klas, według następującego wzoru:

$$D_k = \sqrt{\sum_{j=1}^{n} \left(x_j - x_j^{(k)}\right)^2} \quad , \qquad /3/$$

gdzie x_j jest współrzędną wypowiedzi nadanej nieznanym głosem, $x_j^{(k)}$ jest współrzędną wypowiedzi wzorcowej k-tego głosu, a następnie na zbadaniu, dla której klasy głosów spełniony jest warunek D_k = minimum. Do tej właśnie klasy nastąpi zaliczenie badanego głosu. Inny sposób rozpoznawania traktuje n-wymiarową przestrzeń jako przestrzeń wektorową. Wówczas proces identyfikacji polega na wyliczeniu np. cosinusów kątów γ_k między wektorem badanym W, a kolejnymi wektorami wzorcowymi $W^{(k)}$ według wzoru:

$$\cos \varphi_k = \frac{W \cdot W^{(k)}}{|W| \cdot |W^{(k)}|} = Z_K \qquad /4/$$

Decyzja zaliczenia badanego głosu do jednej z klas wzorcowych następuje dla tej klasy, dla której spełniony jest warunek z_k = maksimum.

Na obecnym etapie badań trudno jest przesądzić które z tych dwóch kryteriów jest lepsze. Z czysto matematycznego punktu widzenia różnica między nimi jest niewielka. Jednakże w praktyce, w zależności od przyjętej metody pomiarowej i zastosowanego przekształcenia układu współrzędnych, kryteria te nie będą sobie równoważne. Problem ten będzie jeszcze przedmiotem dalszych badań.

5. Optymalizacja zespołu parametrów wchodzących do kryterium rozpoznawania

Z opisu metody, podanego w rozdziale 2. 2 wynika, że we wstępnym etapie badań przyjęto podział widma częstotliwościowego na pasma tercjowe, zgodnie z normami IEC. Podział ten byłby optymalny, gdyby cechy osobnicze rozkładały się w paśmie zajętym przez sygnał mowy w sposób równomierny. Jednakże wielu badaczy stwierdziło, że warunek ten w rzeczywistości nie jest spełniony [2] [9]. Istnieją pewne pasma częstotliwościowe zawierające więcej informacji osobniczych niż inne.

Minimalizacja ilości parametrów uzyskana przez wyelimino-

wanie tych pasm, które niosą mało informacji osobniczych, jest istotnym zagadnieniem, ponieważ od jego rozwiązania zależy rozbudowa układowa urządzenia rozpoznającego oraz czas trwania i niezawodność procesu identyfikacji. Im większa jest liczba parametrów, które należy uwzględnić w kryterium rozpoznawania, tym układ będzie bardziej rozbudowany i tym dłuższy będzie czas rozpoznawania.

W celu porównania pasm tercjowych pod kątem zawartości informacji osobniczych zastosowano metodę analizy wariancyjnej parametrów [8]. Zasadniczą ideą tej metody jest założenie, że wybrany do kryterium rozpoznawania parametr winien posiadać dwie następujące własności:

a/ przy wielokrotnym powtarzaniu tekstu przez tę samą osobę rozrzut jego wartości powinien być możliwie mały,

b/ dla różnych osób powinien przybierać możliwie dużo zróżnicowane wartości.

Jak wiadomo, jedną z miar rozrzutu jest wariancja. Wtedy współczynnikiem obejmującym obie własności parametru może być na przykład stosunek wariancji wartości średnich, otrzymanych dla poszczególnych osób do średniej /ze wszystkich q głosów wchodzących do zbioru rozpoznawanych głosów/ wariancji parametru w obrębie jednego głosu zgodnie ze wzorem:

$$F_j = \frac{\frac{s}{q-1}\sum_{k=1}^{q}\left(X_j^{(k)} - U_j\right)^2}{\frac{1}{(s-1)q}\sum_{k=1}^{q}\sum_{i=1}^{s}\left(X_j^{(k)} - x_{ji}^{(k)}\right)^2} \qquad /5/$$

gdzie:

$X_j^{(k)} = \frac{1}{s}\sum_{i=1}^{s} x_{ji}^{(k)}$ jest wartością średnią j-tego parametru dla k-tego głosu /k=1,2,...,q/ z s powtórzeń;

$U_j = \frac{1}{q}\sum_{k=1}^{q} X_j^{(k)}$ jest wartością średnią j-tego parametru /j=1,2,...,n/ dla wszystkich q głosów.

Przy takim zdefiniowaniu współczynnika F_j, im większa jest jego wartość dla danego parametru, tym więcej jest informacji osobniczych w paśmie, które on reprezentuje.

Obliczenia współczynnika F_j dla wszystkich pasm częstotliwościowych przeprowadzono dla dwóch rodzajów nadawanego tekstu:

a/ dla tekstu złożonego z sześciu izolowanych samogłosek,

b/ dla tekstu "gazetowego" o takiej długości aby można było przyjąć, że jest on fonetycznie zrównoważony. Wyniki obliczeń dla obu tekstów wypowiedzianych 15-oma głosami przedstawiono w Tablicy 4.

Analizując powyższą tablicę można wyciągnąć kilka wniosków:

a/ Dla tekstu gazetowego współczynniki F_j są większe niż dla tekstu samogłoskowego. Wynika to po pierwsze z faktu,

Tablica 4. Wartości współczynników F_j dla poszczególnych pasm częstotliwości, obliczone według wzoru /5/ dla dwóch tekstów: samogłoskowego i gazetowego.

$F_{śr.}$ [Hz] / Tekst	160	200	250	315	400	500	630	800
samogłoskowy	39	6	21	58	20	16	9	8
gazetowy	64	38	118	66	58	40	46	86

$F_{śr.}$ [Hz] / Tekst	1000	1250	1600	2000	2500	3150	4000
samogłoskowy	4	11	11	14	14	23	11
gazetowy	68	35	47	33	24	58	47

że tekst samogłoskowy jest bardzo krótki i w czasie jego wypowiedzi zostaje przekazanych znacznie mniej informacji osobniczych, niż w przypadku nadawania tekstu gazetowego, a ponadto wpływ przypadkowych zmian artykulacyjnych na wartość parametrów jest znacznie większy dla tekstów krótkich.

b/ Dla obu tekstów dolne pasma częstotliwości widma mowy, odpowiadające mniej więcej zakresowi pierwszego formantu dźwięków samogłoskowych i niższych harmonicznych tonu krtaniowego, zawierają najwięcej informacji osobniczych. Nie jest to całkiem zgodne z wynikami otrzymanymi na drodze subiektywnych badań odsłuchowych dokonanych przez L. Dukiewicz dla mowy polskiej [9], a także z wynikami otrzymanymi

dla mowy rosyjskiej [2], z których uzyskano, że pasmo od 700 Hz do 3500 Hz zapewnia rozpoznawalność głosów większą niż 90%. W tym zakresie tylko pasmo 2500 Hz dla tekstu samogłoskowego oraz pasma 800 Hz, 1000 Hz i 3150 Hz dla tekstu gazetowego mają względnie duże współczynniki F_j. Być może rozbieżności między przytoczonymi wynikami są spowodowane specyfiką opisanej metody pomiarowej.

Mając dla określonego tekstu uszeregowane parametry $x_j^{(k)}$ według wielkości odpowiadającym im współczynnikom F_j, pozostaje jeszcze określenie liczby pasm częstotliwości, które należy uwzględnić w procesie rozpoznawania. Pozornie wydawać by się mogło, że liczba parametrów powinna być możliwie duża. Jednak pomijając już wcześniejsze zastrzeżenia, uwzględnienie parametrów o małym współczynniku F_j pogorszy jakość rozpoznawania. Z drugiej strony zbyt mała ilość parametrów wskutek dużych strat informacji uniemożliwi efektywne rozpoznawanie głosów.

Niestety na tym wstępnym etapie badań jest niemożliwe zadecydowanie, ile pasm należy uwzględnić w procesie rozpoznawania. Dopiero przeprowadzone na dużym materiale statystycznym eksperymenty, mające na celu ustalenie zależności efektywności rozpoznawania od ilości parametrów i od im odpowiadających współczynników F_j, pozwolą definitywnie rozstrzygnąć ten problem. Wyniki tych badań, uzyskane przy pomocy maszyny cyfrowej pozwolą na opracowanie wstępnych założeń dla technicznego układu rozpoznawania głosów.

Wykaz literatury.

[1] LADEFOGED, P., BROADBENT, D. Information conveyed by vowels, Journ. Acoust. Soc. Am. 29 /1957/, nr 1, 98-106.

[2] RAMISZWILI, G. S. Ob awtomaticzeskom uznawanii gołosow, Izw. AN ZSRR Tiechniczeskaja Kibiernietika, 1966 nr 5, 87-92.

[3] TILLMANN, H. G. Automatische Identifikation von Sprechern, NTZ 1967 nr 12, 706-713.

[4] GLENN, J. W., KLEINER, W. Speaker identification based on nasal phonation, Journ. Acoust. Soc. Am. 43 /1968/, nr 2, 368-372.

[5] KERSTA, L. G. Voiceprint identification, Nature 196 /1962/, 4861, 1253-1257.

[6] KAKAURIDZE, A. G., RAMISZWILI, G. S. O roli zwukow reczi w uznawanii gołosow, Elementy wyczyslitielnoj tiechniki i maszinnyj pierewod /sbornik trudow/, Tbilisi 1964.

[7] NILSON, N. Learning machines, New York 1965.

[8] PRUZANSKY, S., MATHEWS, M. V. Talker-recognition procedure based on analysis of variance, Journ. Acoust. Soc. Am. 36 /1964/, nr 11, 2041-2047.

[9] DUKIEWICZ, L. Frequency-band dependence of speaker identification, Speech analysis and synthesis v.2 /w druku/.

SPECTRAL ANALYSIS OF THE SPEECH SIGNAL AS A CUE FOR THE EVALUATION OF THE INDIVIDUAL SPEAKER'S VOICE FEATURES

Summary

In the speech signal three kinds of information are involved, viz. the linguistic, socio-linguistic and personal information, the latter being the only cue for the identification of individual speaker's voice features. Two methods are hitherto used to eliminate the influence of the phonetic and linguistic content of the utterance on the personal information carried by the speech signal.

One of them consists in determining the personal features in terms of the individual characteristic distortion of a test word uttered by a given speaker. The second method is based on the analysis of some statistical parameters of the speech signal which are independent of the linguistic content of the utterance. These parameters should be averaged for an arbitrary, sufficiently long utterance.

The aim of the present work is to prove whether it is possible to adopt the above mentioned method of spectral analysis for the classification of voices uttering a definite text. The method consist in spectral analysis of the speech signal in 1/3 octave bands and in binary level discrimination in individual frequency bands. The results of the analysis are expressed in terms of n numerical values

corresponding to total time periods in which the threshold levels in n individual frequency bands have been excceded.

The influency of the phonetic and linguistic content of the utterance on the accuracy of voice classification has been investigated. As speech material four types of texts were used throughout the work, viz. six Polish syllabic vowels spoken continuosly, a text from a daily newspaper and two texts arbitrary chosen, containg most voiced sounds and most unvoiced fricative sounds, respectively.

The problem of the suitable choice of optimal frequency bands, that is bands in which most personal information is contained, has been discussed in detail. The possibility of adopting the proposed method of analysis to automatic voice classification has been considered.