J. Kacprowski, R. Gubrynowicz

AUTOMATYCZNE ROZPOZNAWANIE SAMOGŁOSEK POLSKICH METODĄ SEGMENTACJI WIDMA

22/1967

WARSZAWA

Na prawach rękopisu

Do użytku wewnętrznego

Zakład Badań Drgań IPPT PAN Nakład 175 egzemplarzy. Arkusz wyd. 2.9. Arkusz druk. 2.75. Oddano do drukarni w listopadzie 1967 r. Wydrukowano w styczniu 1968 r. Zam. 1045/o/67

W.D.N. Warszawa, ul. Śniadeckich 8

J.Kacprowski i R.Gubrynowicz
Pracownia Elektroakustyki
Zakładu Badania Drgań IPPT

AUTOMATYCZNE ROZPOZNAWANIE SAMOGŁOSEK POLSKICH METODĄ SEGMENTACJI WIDMA

1. Wprowadzenie

Jednym z głównych celów badań teoretycznych i doświadczalnych w dziedzinie analizy i syntezy mowy jest opracowanie metod i systemów automatycznego rozpoznawania sygnałów akustycznych, odpowiadających określonym elementom segmentalnym mowy na płaszczyźnie danego języka. Od rozwiązania tego problemu zależy w dużej mierze postęp techniczny w dziedzinie telekomunikacji, cybernetyki technicznej i automatycznego sterowania oraz dalsze rozszerzenie zastosowań maszyn informacyjno-logicznych i elektronicznych urządzeń liczących. We wszystkich tych przypadkach zagadnieniem podstawowym jest opracowanie koncepcyjne i realizacja techniczna układów wejściowych, których zadanie polega, ogólnie biorąc, na dostarczeniu maszynom lub urządzeniom informacji, programujących wykonywane przez nie czynności. Nośnikiem tych informacji mogą być w zasadzie sygnały o dowolnym charakterze fizycznym, występujące np. w postaci obrazów optycznych, symboli graficznych czy znaków odpowiednio wybranego kodu w systemie perforacji taśmy. Z punktu widzenia zastosowań praktycznych najbardziej dogodnym i operatywnym jest jednak niewątpliwie system przekazywania maszynom lub urządzeniom informacji wejściowych zakodowanych w postaci sygnału mowy, gdyż umożliwia on operatorowi, posługującemu się

danym urządzeniem, kierowanie jego pracą bezpośrednio za pomocą głosu, bez jakichkolwiek operacji pomocniczych.

Perspektywiczny przegląd nowych, lecz już realizowanych w skali laboratoryjno-badawczej, a częściowo i technicznej, zastosowań sygnału mowy w różnych dziedzinach nauki, techniki i przemysłu do celów przekazywania informacji lingwistycznych*/ w układzie transmisyjnym: człowiek-maszyna znaleźć można w pracach problemowych poświęconych temu zagadnieniu /patrz np. [1] [2]/. Najbardziej realne z technicznego punktu widzenia, a jednocześnie najbardziej uzasadnione aktualnymi kierunkami rozwojowymi postępu technicznego w dziedzinie telekomunikacji, cybernetyki i automatyki są zastosowania systemów i urządzeń do automatycznego rozpoznawania mowy w następujących przypadkach:

a/ w systemach telekomunikacyjnych o zwiększonej sprawności przekazywania informacji, działających na zasadzie dyskretnej analizy-syntezy i kompansji sygnału mowy [3],

b/ w maszynach piszących pod dyktando,

c/ przy wprowadzaniu do elektronicznych urządzeń liczących danych wejściowych w postaci elementów lingwistycznych /w najprostszym przypadku cyfr wymawianych głosem/,

d/ przy sterowaniu urządzeń i procesów technicznych głosem, zwłaszcza w tych przypadkach, kiedy zbiór możliwych poleceń przekazywanych maszynie jest ograniczony,

e/ w maszynach informacyjno-logicznych, przeznaczonych np. do bezpośredniego tłumaczenia tekstu słownego na język obcy.

Należy również podkreślić społeczne aspekty zagadnienia automatycznego rozpoznawania mowy, polegające na możliwościach zatrudnienia w przemyśle i produkcji ludzi o ograniczonym posługiwaniu się efektorami ruchowymi /ręce, nogi/ oraz na możliwościach przekazywania wiadomości lingwistycznych dźwiękowych ludziom głuchym.

- - - - - - - - - -

*/ To jest wiadomości, zakodowanych w postaci sygnału mowy.

2. Automatyczne rozpoznawanie elementów segmentalnych mowy na płaszczyźnie ogólnej teorii rozpoznawania

Z teoretycznego punktu widzenia zagadnienie automatycznego rozpoznawania mowy jest szczególnym przypadkiem problemu rozpoznawania dowolnych zjawisk, zdarzeń lub obiektów określonego rodzaju. Proces rozpoznawania - w najbardziej ogólnym znaczeniu tego pojęcia - polega na określaniu przynależności poszczególnych elementów e pewnego zbioru E

$$e \in E \qquad (1)$$

do zawartych w tym zbiorze N podzbiorów, czyli klas E_1, E_2 ... E_N,

$$\begin{aligned} E_1 &\subset E \\ E_2 &\subset E \\ &\text{- - - - -} \\ &\text{- - - - -} \\ E_N &\subset E \end{aligned} \qquad (2)$$

z których każda jest reprezentowana przez zespół znanych a priori i charakteryzujących ją elementów wzorcowych, posiadających istotne cechy dystynktywne decydujące o ich wzajemnym podobieństwie i powinowactwie w ramach danej klasy, a jednocześnie różniące je od elementów pozostałych klas zbioru.

Ponieważ każda klasa określona jest ex definitione przez zespół n charakterystycznych i mierzalnych cech informacyjnych, przeto dowolny element rozpatrywanego zbioru można przedstawić w postaci punktu w liniowej, n-wymiarowej przestrzeni euklidesowej, którego współrzędne x_1, x_2 ... x_n są ilościowymi miarami wybranych n cech charakterystycznych. Przy tak sformułowanej geometrycznej interpretacji zagadnienia, każda klasa E_k (k = 1, 2,.....N) przedstawia sobą obszar przestrzeni n-wymiarowej, utworzony przez zbiór M

punktów o współrzędnych x_{i1}, x_{i2} ... x_{in} $(i = 1, 2 \ldots M)$, z których każdy odpowiada jednemu z M reprezentujących daną klasę elementów wzorcowych. Pojawienie się nowego rozpoznawanego elementu jest w tym ujęciu równoważne zjawieniu się w rozpatrywanej przestrzeni nowego punktu o współrzędnych x'_1, x'_2 ... x'_n, a proces rozpoznawania sprowadza się do wyznaczenia tego spośród N obszarów przestrzeni, wewnątrz którego położony jest nowy punkt, odwzorowujący element rozpoznawany.

W jednej z poprzednich prac [4] przedyskutowano szczegółowo warunki fizycznej i matematycznej optymalizacji metody rozpoznawania, które, ogólnie biorąc, sprowadzają się do:

a/ wyboru optymalnego zespołu takich parametrów fizycznych, określających cechy informacyjne poszczególnych klas zbioru, aby zjawisko wzajemnego przenikania się ich obszarów nie występowało wogóle, tzn. aby iloczyn dwóch dowolnych klas E_k i E_l $(k, l = 1, 2 \ldots N, k \neq l)$ był zbiorem pustym Λ

$$E_k \cap E_l = \Lambda, \qquad (3)$$

lub przynajmniej, aby zjawisko to było zmniejszone do minimum,

b/ przyjęcia takiej matematycznej miary wyznaczania odległości między punktami w przestrzeni n-wymiarowej, aby średnia odległość między punktami odpowiadającymi elementom wzorcowym każdej klasy była jak najmniejsza, a jednocześnie jak najbardziej różniła się od średniej statystycznej odległości wzajemnej punktów, należących do dwóch różnych klas.

Warunek a/ dotyczy zatem wyboru najbardziej korzystnego i efektywnego układu współrzędnych, tj. oparcia definicji klas na takich parametrach fizycznych, których wartości w ramach określonej klasy różnią się między sobą w skali statystycznej nieznacznie, natomiast zmieniają się w sposób zasadniczy przy przejściu od jednej klasy do innej. Warunek b/ natomiast dotyczy matematycznej optymalizacji systemu,

i to zarówno w sensie odpowiedniej transformacji układu współrzędnych, zmierzającej do zmniejszenia obszarów poszczególnych klas, a więc tym samym zagęszczenia rozkładu punktów określających elementy wzorcowe tej klasy, jak i w sensie wyboru najbardziej korzystnej miary wyznaczania odległości między tymi punktami. W cytowanej poprzednio pracy [4] podano przykład interpretacji podanych wyżej założeń ogólnych na płaszczyźnie fonetyczno-akustycznej.

W konkretnym przypadku automatycznego rozpoznawania mowy rozpoznawanymi elementami zbioru mogą być w zasadzie dowolne, lecz z góry określone elementy segmentalne akustycznego sygnału mowy, na przykład rzędu fonemów, głosek, diad, morfemów, sylab, wyrazów lub fraz.

W oddzielnej pracy [5] , dotyczącej ogólnych podstaw teoretycznych procesu automatycznego rozpoznawania mowy, wykazano, że wybór rzędu rozpoznawanych elementów segmentalnych sygnału mowy jest problemem podstawowym, gdyż łączy się on ściśle z trzema zagadnieniami o istotnym znaczeniu technicznym i ekonomicznym,a mianowicie:

a/ liczebnością zbioru układu pamięci urządzenia rozpoznającego,

b/ efektywnością zastosowanego systemu kodowania,

c/ segmentacją sygnału mowy.

Nie rozstrzygając generalnie problemu, jaki rząd rozpoznawanych elementów segmentalnych sygnału mowy jest najkorzystniejszy z punktu widzenia optymalizacji procesu rozpoznawania, gdyż odpowiedź na to pytanie zależy przede wszystkim od przeznaczenia i zakresu zastosowań konkretnego urządzenia rozpoznającego, należy stwierdzić, że zagadnienie automatycznej identyfikacji elementów segmentalnych o rozciągłości fonematycznej, tj. fonemów lub głosek, stanowi w większości przypadków punkt wyjścia przy rozpoznawaniu elementów wyższego rzędu, o określonym znaczeniu fonetycznym /nap. diady/, słowotwórczym /morfemy, sylaby/, semantycznym /wyrazy/, artykulacyjnym /frazy/ czy gramatycznym /zdania/. Wiadomo, że

systemy rozpoznawania morfemów, sylab lub wyrazów, oparte na ustalaniu sekwencji tworzących je fonemów /głosek/ dają wyniki pozytywne. Na korzyść stosowania tych systemów przemawia przede wszystkim fakt, iż liczba fonemów jest niewielka i np. w języku polskim nie przekracza 40. Wynika stąd, że za pomocą 6-znakowego kodu binarnego można przekazywać informacje określające bezpośrednio wszystkie fonemy systemu języka polskiego, co przy prędkości przekazywania informacji równej 10 fonemów na sekundę wymaga stosowania kanału o przepustowości informacyjnej 60 bit/s, a więc około 500 razy mniejszej od przepustowości konwencjonalnych telefonicznych systemów transmisyjnych nawet miernej jakości. Drugą ważną zaletą systemów opartych na identyfikacji fonemów /głosek/ jest to, iż wymagają one stosowania układów pamięciowych o bardzo ograniczonej objętości, niewspółmiernie mniejszej od objętości pamięci systemów rozpoznawania sylab, a tym bardziej wyrazów. Ponadto systemy identyfikacji fonemów są stosunkowo proste z technicznego punktu widzenia, gdyż nie wymagają stosowania układów logicznych, odtwarzających lingwistyczną i statystyczną strukturę systemu danego języka i symulujących proces percepcji mowy przez człowieka na wyższych płaszczyznach rozpoznawiania, na przykład na płaszczyźnie lingwistycznej [5]. Z tych, między innymi, względów zagadnieniu automatycznej identyfikacji fonemów w systemach różnych języków poświęca się obecnie wiele uwagi w pracach badawczych ośrodków naukowych w poszczególnych krajach.

3. Cel i zakres pracy.

Celem niniejszej pracy jest przedstawienie przebiegu i wyników badań prowadzonych w Pracowni Elektroakustyki Zakładu Badania Drgań IPPT - PAN i dotyczących opracowania koncepcyjnego i weryfikacji doświadczalnej metody automatycznego rozpoznawania ograniczonego zbioru głosek polskich. Praca ma charakter teoretyczno-eksperymentalny i służy do sprawdze-

nia w skali laboratoryjnej, na ograniczonym materiale fonetycznym, założeń koncepcyjnych i ogólnych warunków technicznych urządzenia do automatycznego rozpoznawania skończonego zbioru elementów lingwistycznych wyższych rzędów /sylab, wyrazów/ pod kątem zastosowań do automatycznego sterowania urządzeń głosem.

Zbiór rozpoznawanych elementów celowo zawężono do sześciu polskich izolowanych samogłosek sylabicznych. Ograniczenie to, które zresztą nie zmniejsza wartości wyciąganych wniosków ogólnych, podyktowane zostało względami metodologicznymi, gdyż pozwoliło na przeprowadzenie rozważań teoretycznych i badań doświadczalnych przy stosunkowo niewielkiej liczbie parametrów.

Ogólna koncepcja opracowanej metody rozpoznawania oparta jest na przyjęciu jako cech informacyjno-dystynktywnych każdej klasy /grupy fonematycznej/ rozkładów energii akustycznej w widmie częstotliwości rozpoznawanych dźwięków mowy. Rozkłady energii wyznaczano przez podział /segmentację/ widma na n pasm, a następnie binarną reprezentację poziomów energetycznych w poszczególnych pasmach. Przeprowadzone badania obejmowały poszukiwanie optymalnego systemu segmentacji widma i wybór optymalnych progowych poziomów dyskryminacji w każdym pasmie, przy przyjęciu jako kryterium racjonalnego kompromisu między uzyskiwaną dokładnością metody rozpoznawania a techniczną rozbudową układu rozpoznającego. Opracowanie struktury sieci logicznej i jej minimalizację oparto na zastosowaniu ogólnych reguł i założeń algebry Boole'a. W oparciu o wyniki rozważań teoretycznych i badań eksperymentalnych opracowano model laboratoryjny urządzenia rozpoznającego, sprawdzono doświadczalnie dokładność jego działania, a otrzymane wyniki porównano z teoretyczną dokładnością metody rozpoznawiania, wyznaczoną na drodze analitycznej.

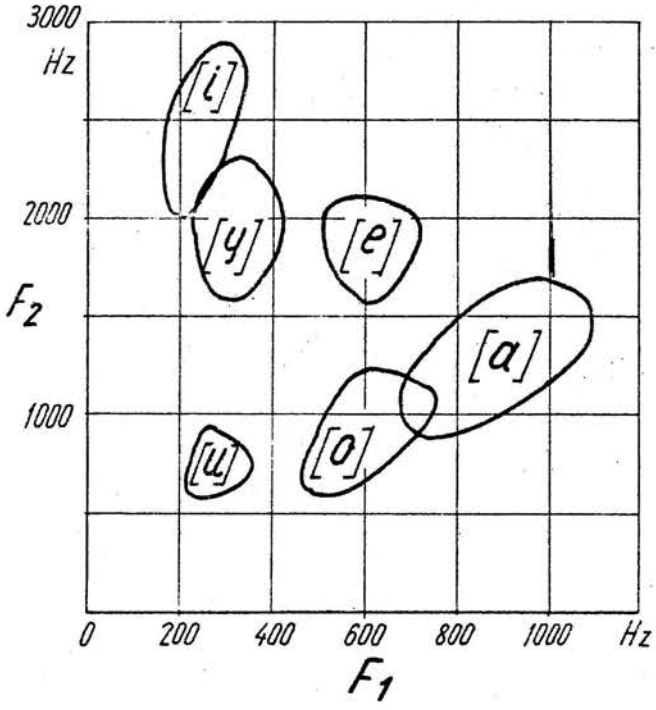
4. Akustyczne cechy dystynktywne samogłosek

Mechanizm wytwarzania dźwięków samogłoskowych oraz związana z tym ich struktura akustyczna były przedmiotem wielu prac z zakresu fonetyki akustycznej i są już dość dobrze znane. Wiadomo /patrz np. [6]/, że podstawową cechą wyróżniającą samogłoski od pozostałych dźwięków mowy jest charakter ich widma, które ma strukturę harmoniczną o wyraźnie zaznaczonych obszarach koncentracji energii. Te pasma częstotliwości, w których składowe harmoniczne widma przybierają wartości odpowiednio większe niż w pozostałych, nazywają się formantami: pierwszym F1, drugim F2 itd., a częstotliwości, odpowiadające wierzchołkom obwiedni widma w poszczególnych pasmach - częstotliwościami formantów: pierwszego F_1, drugiego F_2 itd.

W przypadku samogłosek, a także niektórych spółgłosek, w kształcie obwiedni widma zawarte są informacje, które stanowią istotne cechy dystynktywne umożliwiające identyfikację poszczególnych dźwięków mowy. Stwierdzono, że podstawowe cechy dystynktywne samogłosek określone są rozkładem kolejnych formantów w skali częstotliwości oraz wzajemnymi stosunkami ich amplitud /poziomów/, przy czym dominujące znaczenie mają w tym przypadku dwa najniższe formanty: pierwszy F1 i drugi F2. Formanty wyższe, przede wszystkim trzeci F3 i czwarty F4, wpływają wprawdzie w pewnym stopniu na jakość brzmienia danej samogłoski, jednak głównie charakteryzują one indywidualne /tj. osobnicze/ cechy głosu nadawcy wiadomości.

Na rys. 1 przedstawiono wykresy, obrazujące rozkłady częstotliwości dwóch najniższych formantów na płaszczyźnie F_1, F_2 w przypadku sześciu polskich izolowanych samogłosek sylabicznych: [a] [o] [u] [i] [y] [e], wymówionych pięciokrotnie przez 10 osób, 5 mężczyzn i 5 kobiet*/.

*/ Dane liczbowe uzyskano na podstawie wyników analiz spektrograficznych wykonanych w Pracowni Fonetyki Akustycznej Zakładu Badania Drgań IPPT.



Rys. 1 - Kontury obszarów formantowych samogłosek polskich w układzie współrzędnych F_1, F_2

Z wykresów widać, że przyjęcie jako parametrów fizycznych, określających cechy dystynktywne samogłosek, częstotliwości dwóch pierwszych formantów i oparcie definicji poszczególnych klas fonematycznych na rozkładach F_1 i F_2 w skali częstotliwości jest z punktu widzenia procesu automatycznego rozpoznawania samogłosek zupełnie realne, jakkolwiek - wskutek częściowego pokrywania się obszarów niektórych klas, jak np. klas [i] i [y] lub [a] i [o] - nie może zapewnić 100 % dokładności ich identyfikacji. Ponieważ, zgodnie z założeniem, zbiór rozpoznawanych elementów obejmuje jedynie samogłoski izolowane, można przyjąć, że ich struktura formantowa jest w czasie trwania wypowiedzi ustalona. Przy tym uproszczeniu proces rozpoznawania sprowadza się do wyznaczenia statycznych rozkładów energii w widmie, uśrednionych za

czas trwania wypowiedzi, i porównania ich z rozkładami wzorcowymi, określonymi w wyniku pomiarów statystycznych i zgromadzonymi w układzie pamięciowym.

5. Ogólna koncepcja metody

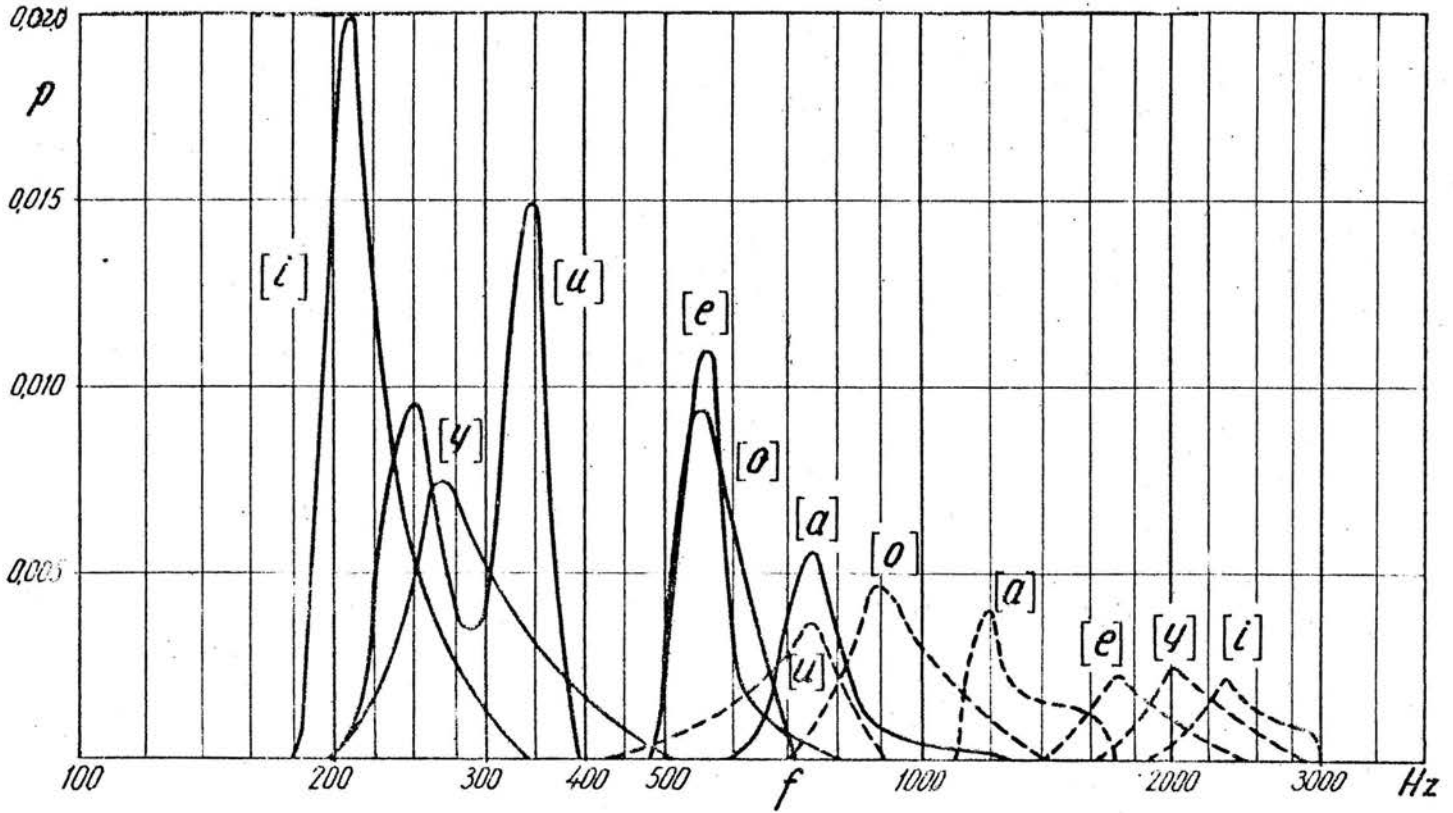
5.1. Wybór przedziałów segmentacji widma

Przyjęcie jako jedynego kryterium rozpoznawania rozkładów energii w widmie sygnału akustycznego, a więc tym samym oparcie metody rozpoznawania na analizie widmowej badanego przebiegu, pociąga za sobą konieczność wyboru szerokości i liczby pasm częstotliwości, w których analiza ta będzie przeprowadzana. Można z góry stwierdzić, że podział widma na pasma o szerokości mniejszej od szerokości pasm formantowych byłby merytorycznie i technicznie nieuzasadniony, gdyż z jednej strony sprowadzałby się do wyznaczania rozkładu gęstości energii w obszarach poszczególnych formantów, co nie jest istotne z punktu widzenia procesu rozpoznawania, a z drugiej powodowałby zbyteczną rozbudowę układu. Do celów identyfikacji samogłosek warunkiem koniecznym i wystarczającym jest jedynie stwierdzenie występowania (+) lub niewystępowania (-) formantów F1 i F2 w ściśle określonych pasmach częstotliwości.

W pierwszym etapie prac przyjęto zasadę równomiernego /w skali logarytmicznej/ podziału całego zakresu częstotliwości, interesującego z punktu widzenia samogłosek 300 Hz - 7500 Hz na n = 8 pasm, bez wnikania w ich powiązania ze strukturą formantową rozpoznawanych dźwięków mowy [4][7]. Ograniczenie liczby parametrów do dwóch najniższych formantów F1 i F2 pozwala na zawężenie rozpatrywanego zakresu częstotliwości do obszaru od 180 Hz do 2800 Hz, gdyż w tych granicach zawarte są zmiany częstotliwości pierwszego i drugiego formantu (rys. 1) . Jednocześnie, uwzględniając prawdopodobieństwa występowania formantów samogłosek w skali

częstotliwości można wyodrębnić na płaszczyźnie F_1, F_2 pola prostokątne o bokach ΔF_1 i ΔF_2 tak dobranych, aby każde z nich miało, praktycznie biorąc, tylko jedno znaczenie fonetyczne, odpowiadające określonej samogłosce, i pokrywało cały obszar zmian częstotliwości jej formantów przy różnych sposobach artykulacji i różnych cechach osobniczych głosu.

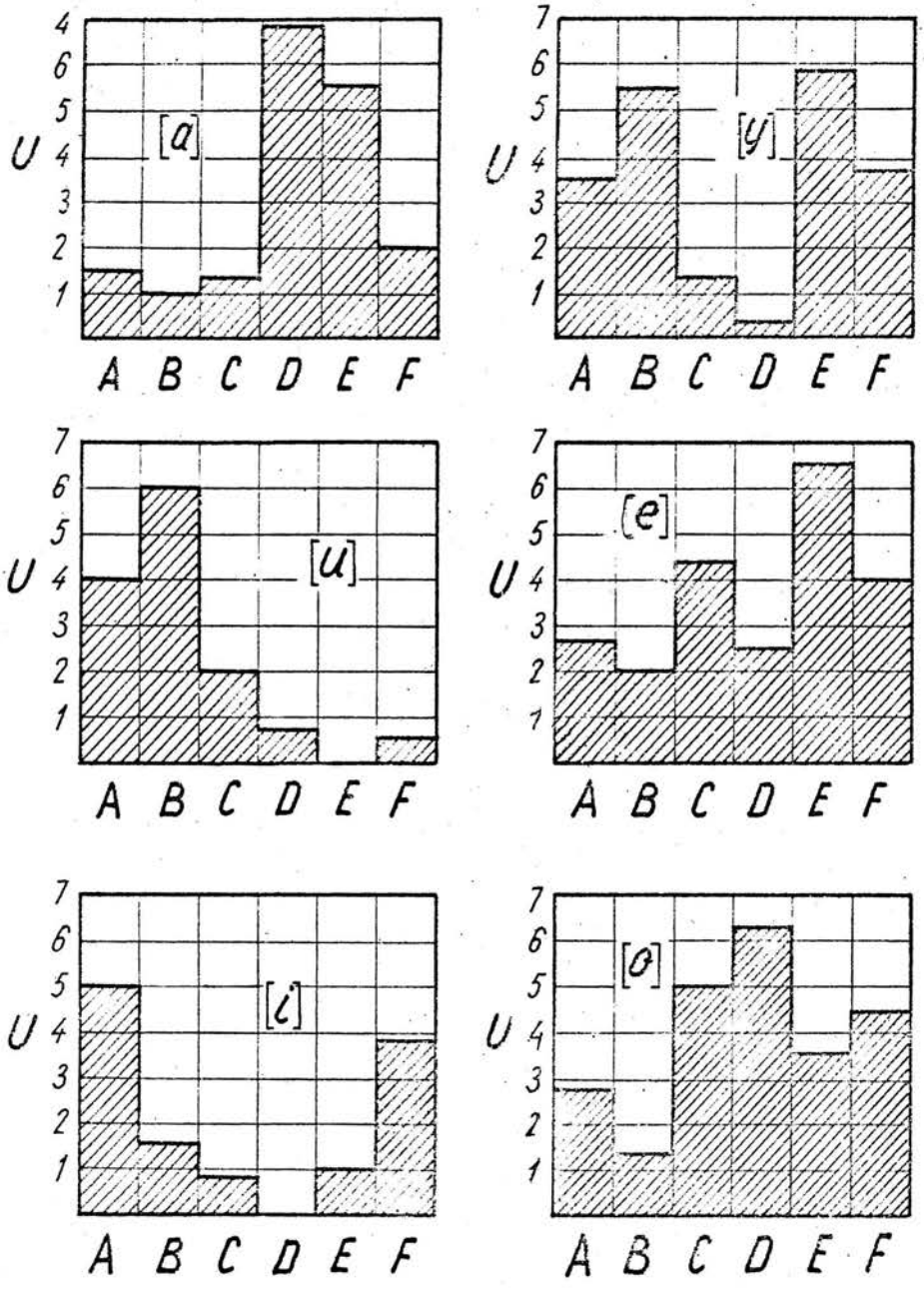
W tym celu, opierając się na danych liczbowych, które uprzednio posłużyły do wykonania wykresów na rys. 1, obliczono rozkłady gęstości prawdopodobieństwa występowania formantów F1 i F2 wszystkich rozpatrywanych samogłosek w skali częstotliwości. Wyniki obliczeń przedstawiono na rys. 2. Analizując otrzymane krzywe rozkładu wyznaczono charakterystyczne pasma częstotliwości, w których prawdopodobieństwo wystąpienia formantów pewnych grup samogłosek jest dostateczne duże ($p > 0,7$), zaś pozostałych grup możliwie małe ($p < 0,3$). Wyniki przeprowadzonej w ten sposób segmentacji widma oraz prawdopodobieństwa występowania formantów w przyjętych sześciu pasmach częstotliwości zawiera tablica 1. Segmentację częstotliwościową zrealizowano w układzie technicznym w postaci zestawu filtrów środkowoprzepustowych, których konstrukcja jest przedmiotem oddzielnej pracy [7]. Częstotliwości graniczne poszczególnych filtrów, mierzone na poziomie - 3 dB w stosunku do wierzchołka charakterystyki tłumienności, są równe częstotliwościom granicznym odpowiednich segmentów widma, podanych w tablicy 1. Na rys. 3 przedstawiono przykładowo wyniki kwantyzacji widma sześciu samogłosek wymówionych głosem męskim, uzyskane za pomocą wspomnianego zestawu filtrów oznaczonych kolejnymi symbolami literowymi od A do F. Rzędne wykresów przedstawiają względne poziomy wyprostowanego napięcia wyjściowego poszczególnych kanałów analizatora, scałkowane w układzie RC o stałej czasowej rzędu 10 ms.



Rys. 2. Rozkłady gęstości prawdopodobieństwa występowania formantów F1 /linia ciągła/ i F2 /linia przerywana/ w skali częstotliwości

Tablica 1. Prawdopodobieństwo występowania formantów F1 i F2 w poszczególnych pasmach częstotliwości

Symbol pasma	Zakres częstotliwości (Hz)	Prawdopodobieństwo występowania formantów F1 i F2 w pasmach											
		F1						F2					
		[a]	[o]	[u]	[i]	[y]	[e]	[a]	[o]	[u]	[i]	[y]	[e]
A	170– 240	–	–	0,19	0,81	0,05	–	–	–	–	–	–	–
B	240– 450	–	–	0,81	0,19	0,94	–	–	–	–	–	–	–
C	490– 600	–	0,81	–	–	–	0,88	–	–	0,22	–	–	–
D	720–1100	0,75	–	–	–	–	–	0,04	0,93	0,30	–	–	–
E	1500–1900	–	–	–	–	–	–	0,17	–	–	–	0,62	0,71
F	1900–2550	–	–	–	–	–	–	–	–	–	0,76	0,38	0,25



Rys. 3. Skwantowane rozkłady widmowe samogłosek otrzymane dla jednego głosu męskiego

5.2. Binarna reprezentacja poziomów widma w pasmach

Jak wynika z rys. 3, wystąpienie w jednym z sześciu pasm częstotliwości /kanałów analizatora/ formantów pewnych samogłosek objawia się wyższym poziomem energetycznym widma, niż w przypadku samogłosek, których formanty leżą poza tym pasmem. Stwierdzenie pojawienia się (+) lub braku (-) formantu wymaga ustalenia progu dyskryminacji poziomu, którego przekroczenie sygnalizowałoby istnienie formantu w rozpatrywanym pasmie częstotliwości.

W tym celu, opierając się na danych liczbowych zestawionych w tablicy 1, wyodrębniono w każdym pasmie częstotliwości dwie grupy samogłosek, a mianowicie:

1/ grupę, w której prawdopodobieństwo występowania formantów w tym pasmie jest duże $(p > 0,7)$,

2/ grupę, w której to prawdopodobieństwo jest małe $(p < 0,3)$.

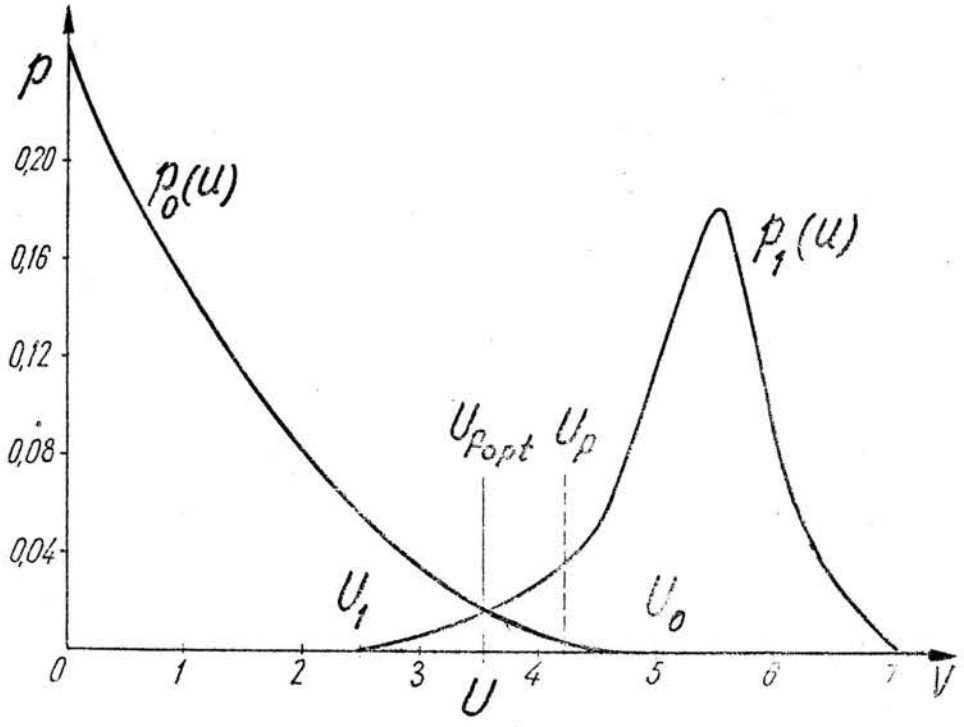
Następnie, posługując się materiałem fonetycznym w postaci sześciu samogłosek wypowiedzianych dwukrotnie piętnastoma typowymi głosami męskimi, wyznaczono statystyczne rozkłady poziomów widma /w skali napięć/ dla każdej z tych grup oddzielnie. Na rys. 4 przedstawiono przykładowo takie rozkłady otrzymane dla kanału D obejmującego pasmo częstotliwości od 720 Hz do 1100 Hz. W tym przypadku krzywa oznaczona $p_1(u)$ przedstawia empiryczną funkcję rozkładu prawdopodobieństwa poziomów widma samogłosek [a] i [o], a krzywa $p_0(u)$ - analogiczną funkcję pozostałych samogłosek rozpatrywanego zbioru, to jest [u][i][y][e]. Ponieważ obie krzywe częściowo zachodzą na siebie, wyznaczenie wartości progowej, która pozwalałaby na bezbłędny rozdział obu grup samogłosek przez binarną dyskryminację poziomu widma w tym pasmie jest teoretycznie niemożliwe. Zadanie sprowadza się zatem do wyznaczenia optymalnej wartości progowej, przy której przyjęciu błąd polegający na zaliczaniu samogłosek z

grupy opisanej rozkładem $p_0(u)$ do grupy opisanej rozkładem $p_1(u)$, lub odwrotnie, jest najmniejszy. Zadanie to można rozwiązać metodą analityczną w sposób następujący.

Niech napięcie progowe w danym kanale ma wartość U_p /rys. 4/. Prawdopodobieństwo zaliczenia samogłoski z grupy opisanej rozkładem $p_0(u)$ do grupy samogłosek opisanych rozkładem $p_1(u)$ wynosi

$$Q_0 = \int_{U_p}^{\infty} p_0(u)\,du = \int_{U_p}^{U_0} p_0(u)\,du \,, \qquad (4)$$

gdzie U_0 oznacza poziom, powyżej którego wartość funkcji $p_0(u)$ jest praktycznie równa zeru.



Rys. 4. Rozkład gęstości prawdopodobieństwa poziomów sygnałów w skali napięć w kanale D

Analogicznie, prawdopodobieństwo zaliczenia samogłoski z grupy opisanej rozkładem $p_1(u)$ do grupy samogłosek opisanej rozkładem $p_0(u)$ jest równe

$$Q_1 = \int_0^{U_p} p_1(u)\,du = \int_{U_1}^{U_p} p_1(u)\,du\,, \qquad (5)$$

gdzie U_1 oznacza poziom, poniżej którego wartość funkcji $p_1(u)$ jest praktycznie równa zeru.

Całkowite prawdopodobieństwo błędu decyzji w rozpatrywanym kanale jest sumą obu prawdopodobieństw:

$$Q_c = Q_0 + Q_1\,. \qquad (6)$$

Można wykazać w sposób elementarny, że suma ta osiąga wartość minimalną, gdy przy przyjętej wartości progowej U_p spełniona jest równość

$$p_0(U_p) = p_1(U_p)\,. \qquad (7)$$

Wynika stąd, że optymalną wartość progową binarnej dyskryminacji poziomu widma w każdym kanale określa punkt przecięcia się wyznaczonych empirycznie dla tego kanału rozkładów prawdopodobieństw $p_0(u)$ i $p_1(u)$.

Posługując się tą metodą kolejno dla wszystkich kanałów i oznaczając pojawienie się formantów określonej samogłoski w poszczególnych kanałach symbolami A, B, C, D, E, F, a brak formantów symbolami $\bar{A}$, $\bar{B}$, $\bar{C}$, $\bar{D}$, $\bar{E}$, $\bar{F}$, można uzyskać binarną reprezentację rozkładu widmowego każdej rozpoznawanej samogłoski rozpatrywanego zbioru w postaci iloczynu logicznego typu

$$m = x_A \cdot x_B \cdot x_C \cdot x_D \cdot x_E \cdot x_F, \qquad (8)$$

w którym każda z sześciu zmiennych x_A, x_B ... x_F może przyjmować jedną z dwóch wzajemnie się wykluczających wartości $A\bar{A}$, $B\bar{B}$, ... $F\bar{F}$.

6. Struktura sieci logicznej układu rozpoznającego

6.1. Rola sieci logicznej w procesie automatycznego rozpoznawania

Z technicznego punktu widzenia proces automatycznego rozpoznawania elementów e dowolnego zbioru E sprowadza się do:

a/porównania każdego nowego elementu z elementami wzorcowymi, zarejestrowanymi w układzie pamięciowym urządzenia rozpoznającego i reprezentującymi poszczególne podzbiory /klasy/, na które zbiór E został a priori podzielony,

b/ wydania decyzji, w stosunku do której z założonych klas zbioru E rozpoznawany element wykazuje największe podobieństwo w przyjętym układzie współrzędnych.

Rolę układu, wykonującego te dwie funkcje, pełni sieć logiczna, której strukturę można określić w wyniku pomiarów statystycznych, obejmujących możliwie obszerny zbiór elementów rozpoznawanych.

6.2. Ogólne postaci funkcji Boole'a rozpoznawanych klas samogłosek.

Każdej samogłosce wypowiedzianej dowolnym głosem odpowiada właściwy dla niej rozkład widma, określony, zgodnie z poprzednimi założeniami, iloczynem logicznym (8) , który w ujęciu algebry Boole'a [8] formalnie odpowiada minimalnemu iloczynowi /mintermowi/ sześciu zmiennych binarnych. Tej samej samogłosce, lecz wymówionej bądź przez tego samego osobnika w inny sposób, bądź - przez innego osobnika o innych właściwościach indywidualnych głosu, odpowiadać mogą oczywiście inne rozkłady formantów w skali częstotliwości, czyli inne iloczyny logiczne sześciu zmiennych określających

binarne poziomy widma w poszczególnych pasmach częstotliwości. W ogólnym przypadku zatem każda z sześciu klas fonematycznych obejmujących sześć samogłosek sylabicznych rozpatrywanego zbioru określona jest funkcją Boole'a w postaci sumy logicznej typu

$$F_{[x]} = \sum m_{[x]j} \quad , \qquad (9)$$

gdzie indeks [x] jest symbolem danej klasy zbioru, a indeks j oznacza numery porządkowe tych mintermów, które w wyniku badań statystycznych zostały uznane za reprezentatywne dla klasy [x]. W rozpatrywanym tu przypadku funkcji sześciu zmiennych indeks j jest określony przez liczby całkowite i dodatnie zawarte w przedziale od j = 0 , czemu odpowiada

$m_0 = \bar{A}\,\bar{B}\,\bar{C}\,\bar{D}\,\bar{E}\,\bar{F}$, do j = 63 $(m_{63} = A\,B\,C\,D\,E\,F)$.

W rezultacie, przy takim ujęciu zagadnienia, każda z sześciu klas fonematycznych zbioru obejmującego samogłoski [a][o][u] [i][y][e] może być określona funkcją /wielomianem/ Boole'a, oznaczoną odpowiednio jako $F_{[a]}$, $F_{[o]}$, $F_{[u]}$, $F_{[i]}$, $F_{[y]}$, $F_{[e]}$ i reprezentującą teoretycznie wszystkie możliwe rozkłady formantów danej samogłoski, odpowiadające różnym warunkom jej artykulacji i różnym cechom osobniczym głosu.

W rozpatrywanym tu konkretnie przypadku, sieć logiczna ma postać układu o n = 6 wejściach, do których wprowadzane są sygnały binarne $A\bar{A}$, $B\bar{B}$, ... $F\bar{F}$ z poszczególnych kanałów pasmowych analizatora, oraz o N = 6 wyjściach, na których występują sygnały binarne 0 lub 1, oznaczające przynależność rozpoznawanej samogłoski do jednej z sześciu grup fonematycznych. Wystąpienie sygnału 1 na wyjściu przyporządkowanemu klasie [x] uwarunkowane jest pojawieniem się na wejściu układu zespołu sygnałów binarnych, odpowiadającego dowolnemu iloczynowi logicznemu $m_{[x]j}$ (8), który wchodzi w skład wielomianu Boole'a $F_{[x]}$ (9), opisującego cechy dystynktywne tej klasy głosek. Wyznaczenie postaci wielomianów $F_{[a]}$, $F_{[o]}$, $F_{[u]}$, $F_{[i]}$, $F_{[y]}$, $F_{[e]}$ musi być dokonane metodą empiryczną,

w wyniku pomiarów statystycznych.

Posługując się jak poprzednio materiałem fonetycznym obejmującym sześć samogłosek wymówionych dwukrotnie przez 15 mężczyzn, wyznaczono empirycznie wielomiany Boole'a (9) opisujące rozpatrywane klasy fonematyczne. Wielomiany te mają postaci następujące:

$$F_{[a]} = \bar{A}\,\bar{B}\,\bar{C}\,D\,\bar{E}\,\bar{F} + \bar{A}\,B\,\bar{C}\,D\,\bar{E}\,F + \bar{A}\,B\,\bar{C}\,D\,E\,\bar{F} + \bar{A}\,\bar{B}\,\bar{C}\,D\,E\,F + \bar{A}\,\bar{B}\,C\,D\,\bar{E}\,\bar{F} + \bar{A}\,\bar{B}\,C\,D\,E\,F + A\,\bar{B}\,\bar{C}\,D\,E\,\bar{F} + A\,\bar{B}\,\bar{C}\,D\,E\,F = \quad (10)$$
$$= m_4 + m_5 + m_6 + m_7 + m_{12} + m_{15} + m_{38} + m_{39} ,$$

$$F_{[o]} = \bar{A}\,\bar{B}\,\bar{C}\,D\,\bar{E}\,\bar{F} + \bar{A}\,\bar{B}\,C\,\bar{D}\,\bar{E}\,\bar{F} + \bar{A}\,\bar{B}\,C\,\bar{D}\,\bar{E}\,F + \bar{A}\,\bar{B}\,C\,D\,\bar{E}\,\bar{F} + \bar{A}\,\bar{B}\,C\,D\,\bar{E}\,F + \bar{A}\,B\,C\,D\,E\,F + A\,\bar{B}\,C\,D\,\bar{E}\,F + A\,\bar{B}\,C\,D\,E\,F = \quad (11)$$
$$= m_4 + m_8 + m_9 + m_{12} + m_{13} + m_{31} + m_{45} + m_{47} ,$$

$$F_{[u]} = \bar{A}\,B\,\bar{C}\,\bar{D}\,\bar{E}\,\bar{F} + A\,B\,\bar{C}\,\bar{D}\,\bar{E}\,\bar{F} + A\,B\,\bar{C}\,\bar{D}\,\bar{E}\,F + A\,B\,\bar{C}\,D\,\bar{E}\,\bar{F} + A\,B\,C\,\bar{D}\,\bar{E}\,\bar{F} + A\,B\,C\,D\,\bar{E}\,\bar{F} = \quad (12)$$
$$= m_{16} + m_{48} + m_{49} + m_{52} + m_{56} + m_{60} ,$$

$$F_{[i]} = \bar{A}\,B\,\bar{C}\,\bar{D}\,\bar{E}\,\bar{F} + \bar{A}\,B\,\bar{C}\,\bar{D}\,\bar{E}\,F + A\,\bar{B}\,\bar{C}\,\bar{D}\,\bar{E}\,F + A\,B\,\bar{C}\,\bar{D}\,\bar{E}\,\bar{F} + A\,B\,\bar{C}\,\bar{D}\,\bar{E}\,F = \quad (13)$$
$$= m_{16} + m_{17} + m_{33} + m_{48} + m_{49} ,$$

$$F_{[y]} = \bar{A}\,B\,\bar{C}\,\bar{D}\,\bar{E}\,F + \bar{A}\,B\,\bar{C}\,\bar{D}\,E\,F + \bar{A}\,B\,C\,\bar{D}\,E\,F + A\,\bar{B}\,\bar{C}\,\bar{D}\,E\,\bar{F} + A\,\bar{B}\,\bar{C}\,\bar{D}\,E\,F + A\,B\,\bar{C}\,\bar{D}\,\bar{E}\,\bar{F} + A\,B\,\bar{C}\,\bar{D}\,\bar{E}\,F + A\,B\,\bar{C}\,\bar{D}\,E\,\bar{F} + A\,B\,\bar{C}\,\bar{D}\,E\,F + A\,B\,C\,\bar{D}\,E\,F = \quad (14)$$

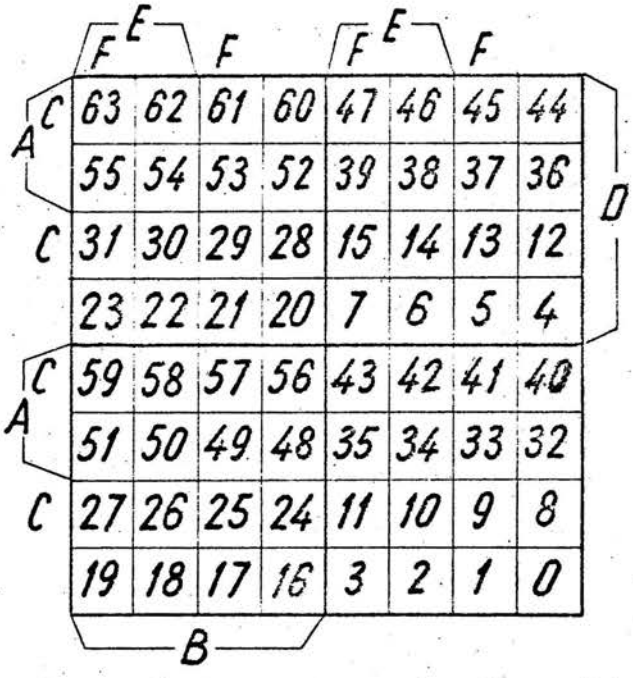
$$= m_{17} + m_{19} + m_{27} + m_{34} + m_{35} +$$
$$+ m_{48} + m_{49} + m_{50} + m_{51} + m_{59} ,$$

$$F_{[e]} = \bar{A}\,\bar{B}\,C\,\bar{D}\,\bar{E}\,\bar{F} + \bar{A}\,\bar{B}\,C\,\bar{D}\,E\,\bar{F} + \bar{A}\,\bar{B}\,C\,\bar{D}\,E\,F +$$
$$+ \bar{A}\,\bar{B}\,C\,D\,E\,F + A\,\bar{B}\,C\,\bar{D}\,E\,F + A\,B\,C\,\bar{D}\,E\,F = \quad (15)$$
$$= m_8 + m_{10} + m_{11} + m_{15} + m_{43} + m_{59}$$

Wielomiany Boole'a $F_{[a]}$, $F_{[o]}$, $F_{[u]}$, $F_{[i]}$, $F_{[y]}$, $F_{[e]}$ wyrażone wzorami (10) do (15) można traktować jako rozkłady widmowe samogłosek "wzorcowych" reprezentujących poszczególne klasy rozpoznawanego zbioru, zarejestrowane następnie w układzie pamięciowym urządzenia. Rozkłady te stanowią jednocześnie podstawę do opracowania struktury sieci logicznej układu decyzji. W tym celu zastosowana została metoda analizy graficznej w oparciu o macierz Karnaugha [9]. Przykład takiej macierzy dla n = 6 zmiennych binarnych A, B, C, D, E, F podano na rys. 5, gdzie, zgodnie z przyjętą konwencją, każda kratka wykresu oznaczona jest liczbą, która w dziesiętnym systemie liczenia odpowiada liczbie dwójkowej, symbolizującej rozpatrywany iloczyn logiczny /minterm/.

Funkcje $F_{[a]}$, $F_{[o]}$, $F_{[u]}$, $F_{[i]}$, $F_{[y]}$, $F_{[e]}$ wyrażone wzorami (10) do (15) zostały naniesione na macierz Karnaugha na rys. 6 przez umieszczenie w poszczególnych kratkach symbolu literowego odpowiedniej klasy samogłosek oraz liczby oznaczającej ilość występujących identycznych rozkładów widmowych, określonych przez minterm przyporządkowany danej kratce.

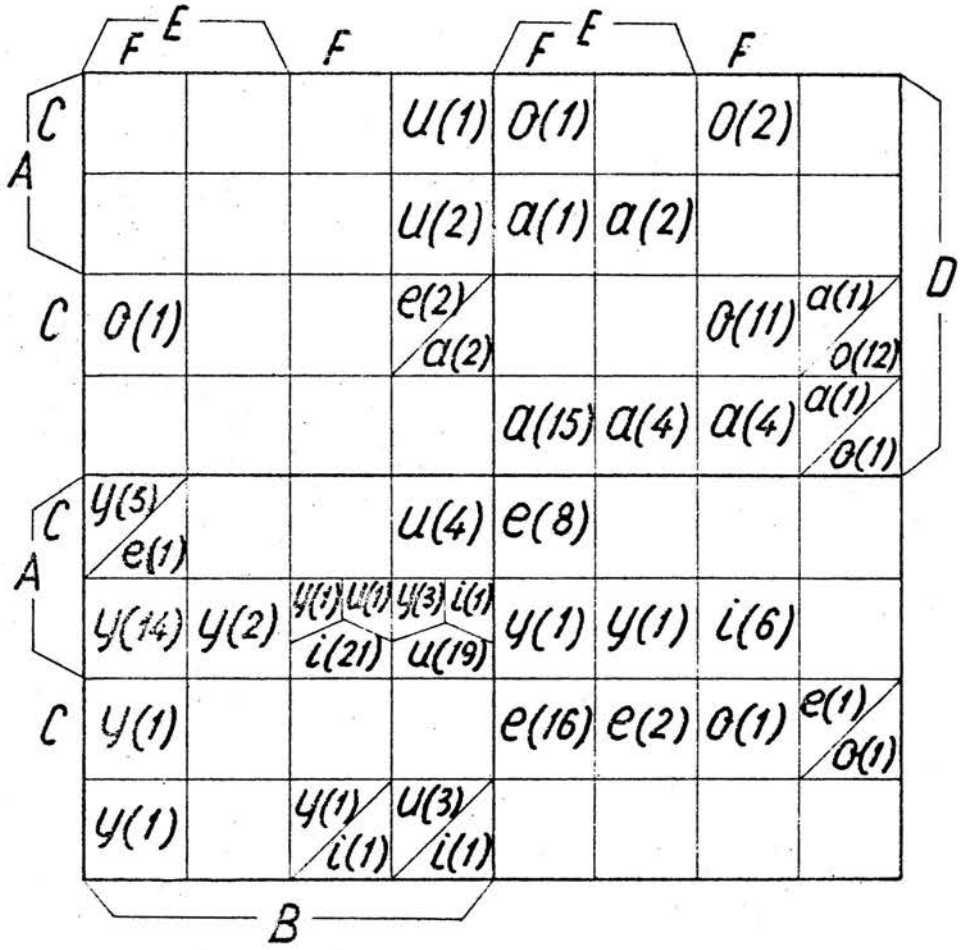
Wspomniane w rozdziale 4 zjawisko wzajemnego przenikania się dwóch lub, ogólnie biorąc, kilku różnych klas zbioru występuje w macierzy Karnaugha na rys. 6 w przypadku tych samogłosek, które są reprezentowane przez identyczne rozkłady widmowe oznaczone wspólnym mintermem. Typowym przykładem mogą tu być samogłoski [u] [i] [y], którym odpowiadają, między innymi, minterny $m_{49} = A\,B\,\bar{C}\,\bar{D}\,\bar{E}\,F$ oraz $m_{48} = A\,B\,\bar{C}\,\bar{D}\,\bar{E}\,\bar{F}$.



Rys. 5. Ogólna postać macierzy Karnaugha dla sześciu zmiennych binarnych

Ponieważ jednak minterm m_{49} występuje w przypadku 21 głosek [i], a tylko jednej głoski [y] i jednej [u], przeto został uznany za typowy dla klasy [i]. Z podobnych względów minterm m_{48} przyporządkowany został formalnie klasie głosek [u]. W tych przypadkach natomiast, kiedy wspólny minterm opisuje jednakowo liczne rozkłady widmowe samogłosek dwóch różnych klas /np. m_4, m_8 lub m_{17}/, może on być uznany za równorzędnie reprezentatywny dla obu rozpatrywanych klas zbioru.

W konsekwencji takich uproszczeń należy się liczyć z nieuniknionymi przekłamaniami układu rozpoznającego, który w pewnych przypadkach będzie mylnie rozpoznawać poszczególne samogłoski. Wynikający stąd błąd systematyczny ocenić można orientacyjnie na około 7,8 %, gdyż, jak wynika z analizy macierzy na rys. 6, w skali statystycznej z ogólnej liczby 180 samogłosek 14 może być mylnie zidentyfikowanych.



Rys. 6. Macierz Karnaugha z rys. 5 z naniesionymi rozkładami widmowymi samogłosek, określonymi wzorami (10) do (15)

6.3. Uproszczenie postaci funkcji Boole'a

W cytowanej poprzednio pracy [4] wykazano, że dokładne odtworzenie w strukturze sieci logicznej pełnych wielomianów Boole'a, w rozpatrywanym tu przypadku opisanych wzorami (10) do (15), nie jest niezbędnym warunkiem prawidłowego działania układu decyzji, a powodowałoby jedynie nadmierną rozbudowę urządzenia. Realizacja techniczna tych wielomianów wymagałaby w konkretnym przypadku zastosowania 301 elementów półprzewodnikowych /diod lub tranzystorów/, co byłoby technicznie i ekonomicznie nieuzasadnione. Z tego względu uproszczenie postaci funkcji Boole'a staje się podstawowym zagadnieniem technicznej realizacji procesu automatycznego rozpoznawania.

Sprecyzowany w rozdziale 6.2. warunek zadziałania układu decyzji można rozszerzyć zakładając, że sygnał 1 na wyjściu [x] sieci logicznej musi wystąpić w każdym przypadku pojawienia się na jej wejściu funkcji $F_{[x]}$, natomiast nie może wystąpić przy pojawieniu się funkcji $F_{[w]}$, $F_{[z]}$... itd., określających cechy dystynktywne pozostałych klas [w], [z]... itd. rozpatrywanego zbioru. Przy takim ujęciu zagadnienia, koniecznymi i wystarczającymi warunkami wystąpienia sygnału 1 na jednym z wyjść, przyporządkowanych poszczególnym klasom głosek [a] [o] [u] [i] [y] [e], jest pojawienie się na wejściu sieci logicznej zespołu sześciu sygnałów binarnych, określonych odpowiednio jedną z sześciu następujących funkcji:

$$F^{*}_{[a]} = F_{[a]}\ \overline{F}_{[o]}\ \overline{F}_{[u]}\ \overline{F}_{[i]}\ \overline{F}_{[y]}\ \overline{F}_{[e]}\ , \qquad (16)$$

$$F^{*}_{[o]} = \overline{F}_{[a]}\ F_{[o]}\ \overline{F}_{[u]}\ \overline{F}_{[i]}\ \overline{F}_{[y]}\ \overline{F}_{[e]}\ , \qquad (17)$$

$$F^{*}_{[u]} = \overline{F}_{[a]}\ \overline{F}_{[o]}\ F_{[u]}\ \overline{F}_{[i]}\ \overline{F}_{[y]}\ \overline{F}_{[e]}\ , \qquad (18)$$

$$F^{*}_{[i]} = \overline{F}_{[a]}\ \overline{F}_{[o]}\ \overline{F}_{[u]}\ F_{[i]}\ \overline{F}_{[y]}\ \overline{F}_{[e]}\ , \qquad (19)$$

$$F^{*}_{[y]} = \overline{F}_{[a]}\ \overline{F}_{[o]}\ \overline{F}_{[u]}\ \overline{F}_{[i]}\ F_{[y]}\ \overline{F}_{[e]}\ , \tag{20}$$

$$F^{*}_{[e]} = \overline{F}_{[a]}\ \overline{F}_{[o]}\ \overline{F}_{[u]}\ \overline{F}_{[i]}\ \overline{F}_{[y]}\ F_{[e]}\ . \tag{21}$$

Przy matematycznej syntezie funkcji $F^{*}_{[a]}$, $F^{*}_{[o]}$, $F^{*}_{[u]}$, $F^{*}_{[i]}$, $F^{*}_{[y]}$, $F^{*}_{[e]}$ można korzystać z klasycznych aksjomatów, reguł i metod ogólnej teorii układów logicznych [9], jednak w wielu przypadkach dogodniej jest sformułować ich postaci, mając przede wszystkim na względzie uproszczenie struktury sieci logicznej i wynikającą stąd minimalizację i optymalizację realizujących ją układów.

Analizując szczegółowo macierz Karnaugha na rys. 6 można stwierdzić, że warunki (16) do (21) spełniają na przykład funkcje:

$$F^{*}_{[a]} = \overline{B}\ \overline{C}\ D\ , \tag{22}$$

$$F^{*}_{[o]} = \overline{B}\ C\ D\ \overline{E}\ , \tag{23}$$

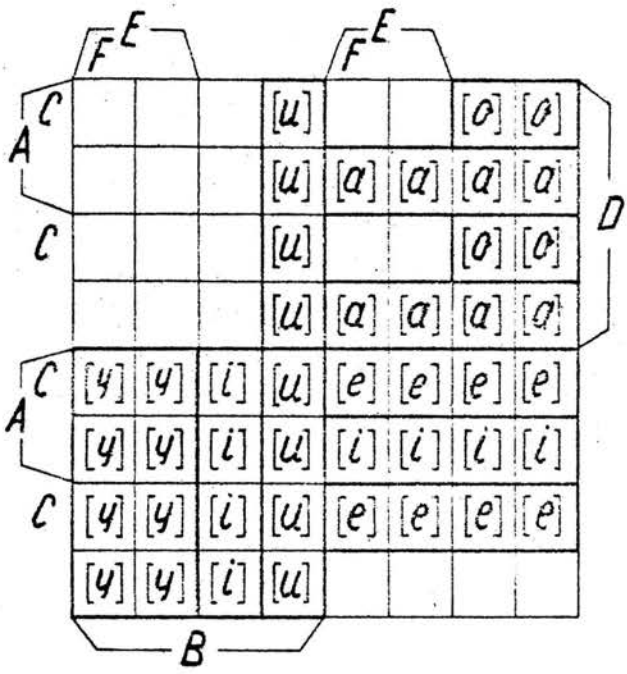
$$F^{*}_{[u]} = B\ \overline{E}\ \overline{F}\ , \tag{24}$$

$$F^{*}_{[i]} = B\ \overline{D}\ \overline{E}\ F + A\ \overline{B}\ \overline{C}\ \overline{D}\ , \tag{25}$$

$$F^{*}_{[y]} = B\ \overline{D}\ E\ , \tag{26}$$

$$F^{*}_{[e]} = \overline{B}\ C\ \overline{D}\ . \tag{27}$$

Funkcje te zostały naniesione na macierz Karnaugha na rys. 7. Z porównania rys. 6 i 7 wynika, że przekształcenie funkcji $F_{[a]}$, $F_{[o]}$, ... $F_{[e]}$, określonych wzorami (10) do (15), do postaci $F^{*}_{[a]}$, $F^{*}_{[o]}$, ... $F^{*}_{[e]}$ według wzorów (22) do (27) osiągnięto przez zwiększenie obszarów poszczególnych klas głosek, a mianowicie - włączenie do nich dodatkowych komórek /mintermów/, przylegających do komórek, tworzących pierwotnie utworzone obszary opisane wzorami (10) do (15). Dzięki temu przekształceniu zwiększyło się prawdopodobieństwo po-

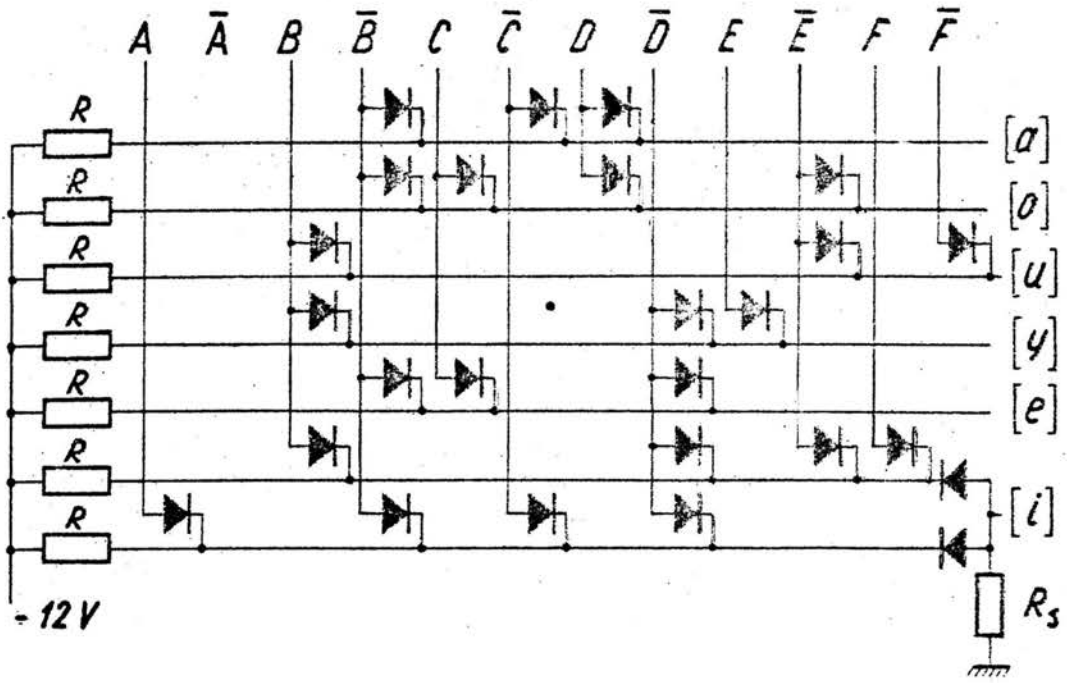


Rys. 7. Macierz Karnaugha z uproszczonymi funkcjami Boole'a, określonymi wzorami (22) do (27)

prawnego rozpoznania tych indywidualnych samogłosek, których rozkłady widmowe nie zostały uwzględnione w poszczególnych składnikach wielomianów $F_{[a]}$, $F_{[o]}$ ... $F_{[e]}$, ale których cechy fonetyczno-akustyczne wykazują istotne podobieństwo do cech określonej klasy samogłosek. Zwiększenie obszarów poszczególnych klas przynosi ponadto dodatkową korzyść w postaci zwiększenia odporności układu rozpoznającego na wpływ zakłóceń.

Jednak najbardziej istotna korzyść wynikająca z uproszczenia pierwotnych postaci funkcji $F_{[a]}$, $F_{[o]}$ ... $F_{[e]}$ polega na związanym z tym uproszczeniu struktury sieci logicznej. Można łatwo wykazać, że do technicznej realizacji funkcji $F^{*}_{[a]}$, $F^{*}_{[o]}$ ... $F^{*}_{[e]}$ określonych wzorami (22) do (27) wystarcza zastosowanie 26 elementów półprzewodnikowych

/diod lub tranzystorów/, a więc przeszło 11 razy mniej, niż do realizacji pełnych wielomianów Boole'a (10) do (15). Na rys. 8 podano schemat ideowy sieci logicznej spełniającej funkcje $F^*_{[a]}$, $F^*_{[o]}$, $F^*_{[u]}$, $F^*_{[i]}$, $F^*_{[y]}$, $F^*_{[e]}$ i zrealizowanej za pomocą diod półprzewodnikowych.



Rys. 8. Schemat sieci logicznej układu rozpoznającego

Wprowadzone uproszczenie nie pozostaje bez wpływu na ogólną dokładność rozpoznawania. Z analizy macierzy Karnaugha na rys. 7 wynika, że wskutek dokonanych przekształceń obszarów poszczególnych klas błąd systematyczny metody przy założonej liczbie elementów wzorcowych wzrósł z 7,8 % do 8,7 %, przy jednoczesnym braku decyzji w przypadku sześciu samogłosek na ogólną ich liczbę 180. Zastosowane przekształ-

cenie nie jest oczywiście jedynym możliwym i przyjęte zostało jako racjonalny kompromis między osiąganą dokładnością rozpoznawania, a rozbudową struktury sieci logicznej.

Przewidywaną teoretycznie, na podstawie analizy tablicy Karnaugha z rys. 7, macierz pomyłek /ang. confusion matrix/ dla przypadku 15 głosów męskich przedstawiono w tablicy 2.

Tablica 2. Przewidywana rozpoznawalność samogłosek wypowiedzianych głosami męskimi /w procentach/

Samogłoska nadana	Samogłoska rozpoznawana						Brak decyzji
	[a]	[o]	[u]	[i]	[y]	[e]	
[a]	90	3	-	-	-	-	7
[o]	3	83	-	-	-	7	7
[u]	-	-	97	3	-	-	-
[i]	-	-	7	93	-	-	-
[y]	-	-	10	13	77	-	-
[e]	-	-	-	-	3	90	7

Posługując się następnie rzeczywistym materiałem fonetycznym w postaci utrwalonych na taśmie magnetycznej sześciu samogłosek wypowiedzianych dwukrotnie przez 15 mężczyzn, otrzymano w wyniku badań eksperymentalnych macierz pomyłek podaną w tablicy 3. Można stwierdzić zgodność między przewidywaną teoretycznie i zmierzoną doświadczalnie liczbą przekłamań urządzenia rozpoznającego. Występujące różnice są wynikiem błędów przypadkowych, spowodowanych głównie wpływem zakłóceń.

Tablica 3. Rozpoznawalność samogłosek wypowiedzianych głosami męskimi, zmierzona w układzie doświadczalnym /w procentach/

Samogłoska nadana	Samogłoska rozpoznawana						Brak decyzji
	[a]	[o]	[u]	[i]	[y]	[e]	
[a]	90	7	-	-	-	-	3
[o]	3	84	-	3	-	-	10
[u]	-	-	91	3	-	-	6
[i]	-	-	3	94	-	-	3
[y]	-	-	6	17	71	6	-
[e]	3	-	-	-	3	88	6

7. Ogólna ocena dokładności metody rozpoznawania

Funkcje Boole'a (10) do (15) oraz (22) do (27), na podstawie których opracowano strukturę sieci logicznej przedstawionej na rys. 8, były wyznaczone w oparciu o ograniczony materiał fonetyczny, obejmujący samogłoski wypowiedziane przez określoną liczbę wybranych głosów męskich. Z tego względu sieć logiczną można traktować jako optymalną jedynie przy rozpoznawaniu samogłosek wymówionych typowym głosem męskim. Należy z kolei sprawdzić, jakiej dokładności rozpoznawania można oczekiwać przy założonej strukturze sieci logicznej w przypadku dowolnego głosu, to jest zarówno męskiego, jak żeńskiego.

Prawdopodobieństwo poprawnego rozpoznania dowolnej samogłoski jest równe iloczynowi prawdopodobieństw przybrania właściwych wartości, to jest 1 lub 0, przez zmienne binarne

x_A, x_B ... x_F, wchodzące w skład uproszczonej funkcji Boole'a, która reprezentuje rozpatrywaną klasę głosek. Z przyjętego systemu dyskryminacji poziomu widma w poszczególnych pasmach częstotliwości /patrz rozdział 5.2./ wynika, że prawdopodobieństwo przybrania przez każdą ze zmiennych x_A, x_B ... x_F wartości 1 lub 0 jest równe odpowiednio prawdopodobieństwu wystąpienia lub niewystąpienia formantu w rozpatrywanym kanale. Korzystając zatem z danych liczbowych zestawionych w tablicy 1, która dotyczy zarówno głosów męskich, jak i żeńskich, i uwzględniając postaci funkcji $F^*_{[a]}$, $F^*_{[o]}$, ... $F^*_{[e]}$ określone wzorami (22) do (27), można obliczyć prawdopodobieństwa $P_{[a]}$, $P_{[o]}$, ... $P_{[e]}$ poprawnego rozpoznania poszczególnych samogłosek rozpatrywanego zbioru w sposób następujący:

$$P_{[a]} = P(x_B = \overline{B}) \cdot P(x_C = \overline{C}) \cdot P(x_D = D) =$$

$$= 1{,}0 \cdot 1{,}0 \cdot 0{,}79 = 0{,}79 \,, \qquad (28)$$

$$P_{[o]} = P(x_B = \overline{B}) \cdot P(x_C = C) \cdot P(x_D = D) \cdot P(x_E = \overline{E}) =$$

$$= 1{;}0 \cdot 0{,}81 \cdot 0{,}93 \cdot 1{,}0 = 0{,}75 \,, \qquad (29)$$

$$P_{[u]} = P(x_B = B) \cdot P(x_E = \overline{E}) \cdot P(x_F = \overline{F}) =$$

$$= 0{,}81 \cdot 1{,}0 \cdot 1{,}0 = 0{,}81 \,, \qquad (30)$$

$$P_{[i]} = P(x_B = B) \cdot P(x_D = \overline{D}) \cdot P(x_E = \overline{E}) \cdot P(x_F = F) +$$

$$+ P(x_A = A) \cdot P(x_B = \overline{B}) \cdot P(x_C = \overline{C}) \cdot P(x_D = \overline{D}) =$$

$$= 0{,}19 \cdot 1{,}0 \cdot 1{,}0 \cdot 0{,}76 + 0{,}81 \cdot 0{,}81 \cdot 1{,}0 \cdot 1{,}0 =$$

$$= 0{,}144 + 0{,}656 = 0{,}80 \,, \qquad (31)$$

$$P_{[y]} = P(x_B = B) \cdot P(x_D = \overline{D}) \cdot P(x_E = E) =$$

$$= 0,94 \cdot 1,0 \cdot 0,62 = 0,58 \quad , \qquad (32)$$

$$P_{[e]} = P(x_B = \overline{B}) \cdot P(x_C = C) \cdot P(x_D = \overline{D}) =$$

$$= 1,0 \cdot 0,88 \cdot 1,0 = 0,88 \quad . \qquad (33)$$

Zgodnie z przewidywaniami, dokładność rozpoznawania samogłosek, wymawianych dowolnym głosem, męskim lub żeńskim, jest odpowiednio mniejsza niż w przypadku typowego głosu męskiego, do którego została dostosowana sieć logiczna układu decyzji. Dokładność tę można oczywiście zwiększyć, przyjmując jako podstawę do opracowania struktury sieci logicznej wielomiany Boole'a wyznaczone w oparciu o rozszerzony materiał fonetyczny, obejmujący samogłoski wypowiedziane również głosami żeńskimi. Takie rozwiązanie byłoby konieczne, gdyby opracowane urządzenie było samoistne i miało za zadanie identyfikację wyłącznie izolowanych dźwięków samogłoskowych. Ponieważ jednak, zgodnie z założeniem, referowana praca ma charakter teoretyczno-eksperymentalny i służy do sprawdzenia w skali laboratoryjnej ogólnej koncepcji i warunków technicznych metody automatycznego rozpoznawania elementów lingwistycznych wyższych rzędów, przeto rozszerzanie materiału fonetycznego, stanowiącego bezpośredni przedmiot badań, nie wydaje się na obecnym etapie zaawansowania prac celowe, tym bardziej, że otrzymane wyniki ilościowe nie zmniejszają wartości wyciąganych wniosków ogólnych o charakterze metodologicznym.

Istotne zastrzeżenia budzić może stosunkowo duża wartość błędu przy rozpoznawaniu samogłoski [y], której prawdopodobieństwo identyfikacji w przypadku głosów męskich wynosi 71 %, a w przypadku dowolnego głosu - 58,2 %. Przyczyną tego jest specyficzny charakter struktury formantowej klasy głosek [y], której formant F1 leży w obszarze formantów F1 gło-

sek [i] oraz [u], a formant F2 - w zakresie częstotliwości formantów F2 głosek [i] oraz [e] /por. rys. 1 i 2/, co utrudnia realizację techniczną procesu jej automatycznego rozpoznawania wyłącznie na płaszczyźnie akustycznej, w oderwaniu od struktury lingwistycznej określonego języka. Należy podkreślić, że również w procesie percepcji mowy przez człowieka rozpoznawalność izolowanych głosek [y] jest znacznie mniejsza niż pozostałych dźwięków samogłoskowych [10].

Zestawienie zbiorcze dokładności rozpoznawania samogłosek w przyjętej metodzie segmentacji widma i dyskryminacji poziomu w pasmach, w oparciu o założoną strukturę sieci logicznej podano w tablicy 4.

Tablica 4. Przewidywana rozpoznawalność doświadczalnego urządzenia do automatycznego rozpoznawania samogłosek wypowiedzianych dowolnym głosem /w procentach/

Samogłoska nadana	Samogłoska rozpoznawana						Brak decyzji
	[a]	[o]	[u]	[i]	[y]	[e]	
[a]	79	-	-	-	-	-	21
[o]	18	75	-	-	-	6	1
[u]	-	-	81	3	-	-	16
[i]	-	-	19	80	-	-	1
[y]	-	-	22	12	58	-	2
[e]	-	-	-	-	-	88	12

8. Realizacja techniczna układu eksperymentalnego do rozpoznawania samogłosek

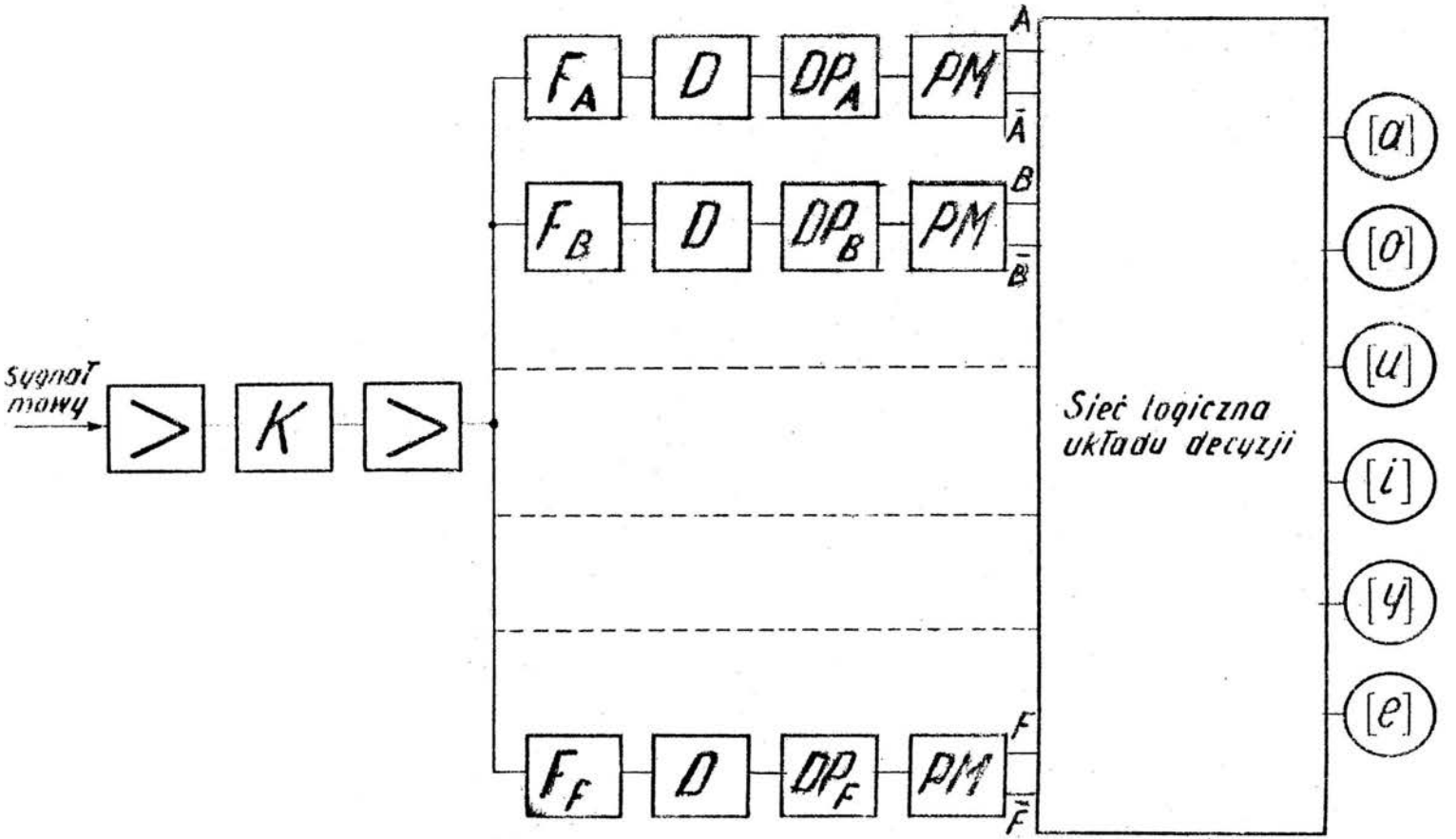
Na podstawie przeprowadzonych rozważań teoretycznych potwierdzonych wynikami wstępnych badań doświadczalnych w układach laboratoryjnych opracowano układ eksperymentalny do automatycznego rozpoznawania izolowanych polskich samogłosek sylabicznych.

Schemat blokowy układu podany jest na rys. 9.

Kompresor K w torze mikrofonowym służy do normalizacji sygnału mowy w sensie uniezależnienia napięcia wyjściowego filtrów analizujących F_A, F_B ... F_F od poziomu nadawania. Kompresor zapewnia stałość napięcia wyjściowego w granicach 3 dB przy zmianach poziomu nadawania o 20 dB. Stała czasowa kompresora jest rzędu 150 ms.

Detektory kanałowe D są zbudowane w układzie prostowników jednopołówkowych /diody krzemowe/ z podwajaniem napięcia. Napięcia wyjściowe detektorów kanałowych są scałkowane w układach RC o stałych czasowych rzędu 10 ms. Rolę układów progowych w poszczególnych kanałach analizatora pełnią indywidualne dyskryminatory poziomów DP_A, DP_B ... DP_F w układzie przerzutnika jednostawowego Schmitta, wyzwalające przerzutniki monostabilne PM o dwóch wyjściach, na których występują napięciowe sygnały binarne $A\bar{A}$, $B\bar{B}$, ... $F\bar{F}$, doprowadzane do końcówek wejściowych sieci logicznej układu decyzji. Sygnały wyjściowe układu decyzji uruchamianją optyczne wskaźniki klas rozpoznawanych samogłosek.

Opracowany układ zamyka wstępny etap prac w zakresie automatycznego rozpoznawania elementów lingwistycznych wyższego rzędu, np. sylab lub wyrazów.



Rys. 9. Schemat blokowy doświadczalnego urządzenia do automatycznego rozpoznawania samogłosek

WYKAZ LITERATURY

1. Olson, H.F., Processing of sound. Proc. IRE. t.50 /1962/, nr 5, ss. 599-600.
2. Kacprowski J., Zastosowania analizy i syntezy mowy w telekomunikacji i automatyce. Rozprawy Elektrotechniczne, t.11/1965/, nr 2, ss. 479-491.
3. Kacprowski J., Speech compression by means of analysis-synthesis methods. Proc. Vibr.Probl., t.5/1964/, nr 3, ss. 193-207.
4. Kacprowski J., Teoretyczne podstawy metody automatycznego rozpoznawania samogłosek. Archiwum Akustyki, t.2 /1967/, nr 3, ss. 227-253.
5. Kacprowski J., Teoretyczne podstawy procesu automatycznego rozpoznawania mowy. Archiwum Akustyki, t.2/1967/, nr. 2, ss. 123-151.
6. Fant G., Acoustic Theory of Speech Production, s-Gravenhage, 1960.
7. Gubrynowicz R., Kwantyzacja częstotliwościowa widma do celów automatycznego rozpoznawania samogłosek. Archiwum Akustyki, t. 2 /1967/, nr. 3, ss. 255-266.
8. Mostowski A.W., Algebry Boole'a i ich zastosowania. PWN, Warszawa 1964.
9. Phister M., Logical design of digital computers. Wyd. John Wiley and Sons, New York 1958, rozdz. 1 - 4.
10. Pollack I., Picket C.K., Information of elementary multidimensional auditory displays. Journ. Acoust. Soc. Amer., t. 26/1954/, ss. 155-158.

AUTOMATIC RECOGNITION OF POLISH VOWELS IN TERMS OF SPECTRUM SEGMENTATION

Summary

The subject of the present paper is the general discussion of some theoretical problems concerning a simple method used for the automatic recognition of isolated speech sounds. The principle of the method consists in dividing the short-time spectrum of the input speech signal into n frequency bands and in evaluating the spectrum levels in successive bands with simple amplitude discriminators. To simplify the theoretical considerations, a few constraints and limitations have been introduced concerning both the investigated speech material /six Polish syllabic vowels/ and the number of distinctive features describing each class of vowels /n = 6 frequency bands, a = 2 spectrum levels in each band, maximum amount of information $I_{max} = a^n = 2^6 = 64$ bits per vowel/.

Taking as the starting point the formal rules of the classical theory of the recognition of membership in classes, the authors discuss the general conditions which should be fulfilled by a device which has to recognize isolated speech sounds independently of speakers voice quality and articulation. These conditions have been then adapted to the particular case of the limited ensemble of speech sounds consisting of six Polish isolated syllabic vowels uttered by male speakers. Special attention is paid to the problem of the most effective method of spectrum segmentation, that is to the choice of the appropriate frequency bands and the individual threshold levels in each band.

The theoretical considerations are based on the general rules of Boolean algebra which constitutes the most convenient mathematical means for the description, synthesis and design of the logical nets forming the decision circuit of the recognizer. The Boolean functions describing the distinctive features of the considered vowel classes have been then reduced to simplify the internal structure of the decision circuit built of semiconductor diods. The choice of the appropriate spectrum patterns to be recorded in the memory of the recognizer is done by the graphical method based on the analysis of Karnaugh's matrices. The theoretically forseen as well as experimentally measured confusion matrices in recognition of six Polish vowels are given to prove the accuracy of the method.

In the last chapter the general block diagram of an experimental device for automatic recognition of isolated Polish vowels is briefly described. The device will be used in further research work on automatic recognition of speech entities of higher orders /e.g. syllables and words/ with applications to automatic control of machines by voice.

LA RECONNAISSANCE AUTOMATIQUE DE VOYELLES DE POLONAIS A L'AIDE DE LA SEGMENTATION DU SPECTRE.

R e s u m é

Le travail présent expose certains problèmes théoriques d'une méthode simplifiée pour la reconnaissance automatique de sons isolés de la parole. Cette méthode est basée sur l'analyse du spectre à l'aide d'un banc de filtres passbande à la sortie desquels on estime après la détection le niveau d'amplitude du chaque signal. Cette estimation est réalisée à l'aide de discriminateurs d'amplitudes.

Dans le but de simplifier les considérations théoriques on a introduit quelques limitations concernantes autant le nombre de phonèmes /six voyelles seulement/ que le nombre de paramètres qui caractérisent chaque classe de voyelles /n = 6 bandes des fréquences, a = 2 niveaux de discrimination d'amplitude, d'où le maximum de quantité d'information est $I_{max} = 2^6 = 64$ bits par voyelle/.

Comme point de départ on a adopté les principes généraux de la théorie classique de la reconnaissance des éléments dans les classes. Les auteurs ont exposé les conditions générales pour un système de reconnaissance qui identifie les sons de la parole indépendamment du locuteur et de l'articulation. En titre d'exemple on a précisé les conditions pour la reconnaissance des voyelles isolées, prononcées uniquement par des hommes.

Une attention particulière était portée sur la méthode optimum de la division du spectre et sur le choix de niveau de discrimination d'amplitude dans chaque bande de fréquences.

Les considérations théoriques sont basées sur les principes de l'algèbre de Boole qui est la plus appropriée pour

la description, synthèse et conception du circuit logique dans le bloc de décision.

Le choix de la distribution spectrale qui doit être enregistrée dans la mémoire du système est basé sur la méthode graphique d'analyse de matrice de Karnaugh.

Dans la dernière partie on a présenté les résultats théoriques et expérimentales de la reconnaissance sous la forme de matrices de confusions /confusion matrix/ qui donnent une idée sur la précision de la méthode. Enfin on a donné le schéma de principe du système utilisé, suivi d'une courte description.