Janusz Kacprowski

OGÓLNA KONCEPCJA SYSTEMU DWUSTRONNEJ KOMUNIKACJI AKUSTYCZNEJ «OPERATOR — MASZYNA ROBOCZA (POJAZD)» ZA POMOCĄ SYGNAŁU MOWY

40/1986

WARSZAWA 1986

ISSN 0208-5658

Praca wpłynęła do Redakcji dnia 22 października 1986 r.

Na prawach rękopisu

Instytut Podstawowych Problemów Techniki PAN

Nakład 140 egz. Ark.wyd. 1,9 Ark.druk. 2,25
Oddano do drukarni w listopadzie 1986 r.
Nr zamówienia 637/86

Warszawska Drukarnia Naukowa, Warszawa,
ul. Śniadeckich 8

Janusz Kacprowski
Zakład Akustyki Cybernetycznej
IPPT - PAN

OGÓLNA KONCEPCJA SYSTEMU DWUSTRONNEJ KOMUNIKACJI AKUSTYCZNEJ «OPERATOR - MASZYNA ROBOCZA (POJAZD)» ZA POMOCĄ SYGNAŁU MOWY

1. Wstęp

Lata 80-te obecnego stulecia są okresem intensywnego rozwoju automatyki i robotyki oraz ich zastosowań technicznych i przemysłowych. Zjawisko to występuje zwłaszcza w wysoko uprzemysłowionych krajach świata: Japonii, RFN, Szwecji, USA i ZSRR, które według orientacyjnych danych, zaczerpniętych ze źródeł amerykańskich (ROBOT INSTITUTE OF AMERICA, 1982), dysponowały już w roku 1982 łączną liczbą ponad 24.000 czynnych robotów przemysłowych różnych typów, co stanowiło około 90% globalnej populacji tych urządzeń w skali światowej, ocenianej wówczas na 26.500 sztuk. W świetle tempa rozwoju automatyzacji procesów produkcyjnych i przemysłowych, obserwowanego w latach 1982-1984, należy obecnie uznać te dane za znacznie zaniżone w stosunku do stanu aktualnego, tym bardziej, że nie obejmowały one niektórych krajów pozaeuropejskich. Według danych japońskich (YONEMOTO, 1984) całkowita liczba robotów przemysłowych, zainstalowanych i czynnych na świecie w roku 1984, wynosiła około 54.000, z czego 50% przypadało na Japonię.

W zestawieniu z tymi danymi liczba robotów, jakimi dysponuje Polska, jest rażąco mała. Niektórzy autorzy (MORECKI, BUĆ, 1981) oceniali populację robotów w Polsce w roku 1981 na około 350 egzemplarzy, natomiast inny autor (KACZMARCZYK, 1984) jest zdania, że z różnych względów nie przekraczała ona 200. Należy sądzić, że stan ten obecnie nie uległ znaczącej poprawie, co

świadomy o krytycznym opóźnieniu rozwoju automatyki i robotyki oraz ich zastosowań w Polsce w stosunku do czołówki światowej. Sytuacja taka jest groźna dla ogólnego, to jest technicznego i ekonomicznego stanu gospodarki narodowej, który w krajach rozwiniętych i uprzemysłowionych w znacznej mierze zależy od stopnia automatyzacji urządzeń technicznych i procesów technologicznych, zwłaszcza o charakterze przemysłowym i produkcyjnym, realizowanym za pomocą robotów, których zakres zastosowań, w miarę ich udoskonalania, obejmuje coraz liczniejsze gałęzie przemysłu i coraz bardziej skomplikowane funkcje, nie dające się w perspektywie najbliższych lat nawet przewidzieć. Spektakularnym wykładnikiem tych tendencji rozwojowych jest wydawnictwo zbiorowe (INDUSTRIAL ROBOT MAGAZINE, 1983), które obrazuje postęp w dziedzinie robotyki w skali światowej w bieżącym 10-leciu.

Jest jednak rzeczą oczywistą, że stopnia zaawansowania automatyzacji i robotyzacji w określonym kraju, gałęzi przemysłu czy w konkretnym zakładzie przemysłowym nie można ocenić wyłącznie licznością populacji zainstalowanych i czynnych robotów. Drugim, bardziej miarodajnym bo merytorycznym, kryterium oceny stanu automatyzacji i robotyzacji jest nowoczesność techniczna stosowanych do tego celu urządzeń, której miarą są m.in. ich możliwości manipulacyjne i lokomocyjne, mobilność, niezawodność oraz wielofunkcyjność i precyzja działania. Ostatnio coraz częściej stosowanym kryterium oceny robotów stają się ich możliwości intelektualne, polegające na samoprogramowalności oraz zdolności do podejmowania samodzielnych decyzji, w oparciu o analizę wytworzonego modelu otaczającego środowiska lub obiektu manipulacji, przy ewentualnej konsultacji z operatorem, przeprowadzanej w formie bezpośredniego dialogu w wybranym języku porozumiewania się.

Pod względem możliwości funkcjonalnych i intelektualnych robotów wyróżnia się obecnie ich trzy generacje, których definicje i określenia, dość zresztą płynne, a niekiedy nawet niejednoznaczne, występują w różnych pozycjach fachowej literatury przedmiotu, a w piśmiennictwie polskim są zestawione m.in. w cytowanej już poprzednio pracy (KACZMARCZYK, 1984), na którą można się w tym miejscu powołać, nie powtarzając zawartej w niej szczegółowej klasyfikacji. Do celów niniejszego opracowania wy-

starczą zaczerpnięte z niej następujące uproszczone określenia kolejnych generacji:

Generacja I: roboty sterowane w układzie otwartym, tj. bez sprzężenia zwrotnego, działające niezależnie od aktualnego stanu otaczającego środowiska lub obiektu manipulacji, wyposażone jedynie w czujniki proprioceptywne, służące do śledzenia stanów wewnętrznych samego robota, który w związku z tym określany jest jako układ o pozycjonowaniu wewnętrznym.

Generacja II: roboty wyposażone w czujniki zewnętrzne oraz odpowiednie układy sensoryczne, obecnie głównie dotykowe lub/i wizyjne, umożliwiające zbieranie informacji o aktualnym stanie otaczającego środowiska lub obiektu manipulacji i wykorzystanie ich do celów sterowania robotem, który w związku z tym jest określany jako układ o pozycjonowaniu zewnętrznym.

Generacja III: roboty wyposażone nie tylko w czujniki zewnętrzne i układy sensoryczne, ale i elementy sztucznej inteligencji, zdolne do samodzielnego podejmowania decyzji i korygowania programów działania, na podstawie analizy wytworzonego we własnym zakresie modelu otaczającego środowiska lub obiektu manipulacji, oraz posługujące się określonym językiem komunikacji z operatorem.

Obecnie w produkcji i zastosowaniach przemysłowych znajdują się głównie roboty I generacji, a zaczynają się pojawiać roboty II generacji. Roboty III generacji są we wstępnym stadium badań i opracowań. Roboty II i III generacji określane są niekiedy, zgodnie z dokumentem normalizacyjnym (ISO/TC 97, 1978), wspólnym terminem «roboty inteligentne» jako "roboty, które mogą same stanowić o swoim zachowaniu dzięki swoim możliwościom postrzegania i rozpoznania".

Rozwój robotyki i jej zastosowań w Polsce, w celu likwidacji zacofania naszego kraju w tej dziedzinie w stosunku do średniego poziomu światowego, został uznany za jeden z głównych i priorytetowych kierunków badań podstawowych i stosowanych oraz wdrożeń w planach 5-letnich 1986-90 i 1991-95. Zgodnie z postanowieniami Komitetu Nauki i Postępu Technicznego oraz wytycznymi, zawartymi w materiałach III Kongresu Nauki Polskiej, kierunek ten, na równi z automatyką, elektroniką, telekomunikacją, bioinżynierią oraz kilku innymi dyscyplinami nauk technicznych,

został umieszczony w programach badawczych centralnych i międzyresortowych, realizowanych w sposób kompleksowy przez placówki Polskiej Akademii Nauk, Ministerstwa Nauki i Szkolnictwa Wyższego oraz instytuty resortowe.

Szczególna rola przy realizacji tych zadań przypada Instytutowi Podstawowych Problemów Techniki PAN, który pełni rolę jednostki koordynującej program badań podstawowych "Układy ze sztuczną inteligencją do maszyn roboczych i pojazdów". Jednym z cząstkowych celów, określających kierunki badań podstawowych w ramach tego programu, jest rozwijanie metod i systemów komunikacji «człowiek - robot» w wybranych językach porozumiewania się i ich weryfikacja w układach doświadczalnych.

Opracowanie niniejsze jest raportem, który wskazuje na realne, choć dotychczas jeszcze nie wykorzystane możliwości zastosowania sygnału akustycznego, a zwłaszcza sygnału mowy, jako nośnika zakodowanych w nim informacji, do celów komunikacji i sterowania w hybrydowym układzie cybernetycznym «operator - maszyna robocza (pojazd)». Przedstawiona w obecnym raporcie ogólna koncepcja takiego systemu oparta jest na wynikach dotychczasowych badań podstawowych i stosowanych, teoretycznych i eksperymentalnych, w dziedzinie analizy, automatycznego rozpoznawania i syntezy mowy, prowadzonych w Pracowni Akustyki Mowy Zakładu Akustyki Cybernetycznej IPPT-PAN od roku 1960.

2. Założenia ogólne systemu

Tak ogólnie sformułowany przedmiot pracy wymaga wprowadzenia następujących uzasadnień, uściśleń i ograniczeń, określających w sposób jednoznaczny jego sens i znaczenie pojęciowe.

a). Ponieważ stosowane obecnie układy sterowania robotów inteligentnych II i III generacji, do których zaliczają się m.in. maszyny robocze i pojazdy, są z reguły układami komputerowymi, realizowanymi zazwyczaj w technice mikroprocesorowej i odpowiednio oprogramowanymi, przeto pod pojęciem : "dwustronna komunikacja «operator-robot» za pomocą sygnału mowy" należy w kontekście obecnej pracy rozumieć dwustronną komunikację między systemem komputerowym konkretnego robota i człowiekiem, wykonującym funkcje jego programowania, sterowania, kontroli i nadzoru. Tak pojmowana komunikacja człowiek-komputer odbywa się

w języku naturalnym za pomocą głosu i polega, mówiąc najogólniej, na wprowadzaniu do systemu informacji wejściowych w postaci instrukcji, rozkazów, nazw urządzeń, zbioru danych itp., zakodowanych w postaci akustycznego sygnału mowy naturalnej i wywołujących określoną reakcję systemu (tzw. "akustyczne wejście" systemu), oraz na odbiorze informacji wyjściowych, określających aktualne parametry stanu samego robota, jak i jego środowiska pracy lub obiektu jego manipulacji, przekazywanych sygnałem mowy syntetycznej (tzw. "akustyczne wyjście" systemu).

b). Zastąpienie konwencjonalnych układów wejściowych (klawiatura ze znakami alfanumerycznymi, pióro świetlne, czytnik) oraz wyjściowych (monitor, drukarka znakowa, perforator) systemu komputerowego urządzeniami peryferyjnymi akustycznego wejścia i akustycznego wyjścia uzasadnione jest względami ergonomicznymi, technicznymi i ekonomicznymi, gdyż:

- wykorzystuje naturalny językowy kod fonetyczny, najbardziej zrozumiały dla obsługi systemu i nie wymagający gruntownej znajomości wyspecjalizowanych języków programowania;

- stosuje narządy mowy i słuchu jako informacyjny kanał akustyczny, zastępujący narządy wzroku i ruchu, które alternatywnie mogą być wykorzystane do pełnienia innych funkcji obsługi systemu;

- daje operatorowi swobodę ruchu i położenia oraz umożliwia mu komunikację z systemem w zupełnej ciemności, co zapewnia ciągłość i pewność przepływu informacji w sytuacjach awaryjnych oraz w warunkach specjalnych, np. w stanie nieważkości lub przy znacznych przyspieszeniach;

- koszt akustycznych przetworników wejściowych (mikrofony) i wyjściowych (głośniki) jest znacznie mniejszy od odpowiednich konwencjonalnych urządzeń peryferyjnych systemu komputerowego.

c). Wspomnianego wyżej procesu wprowadzania do systemu komputerowego informacji w języku naturalnym za pomocą głosu nie należy utożsamiać z ogólnym pojęciem a u t o m a t y c z n e g o r o z p o z n a w a n i a m o w y (ARM), rozpatrywanej jako uporządkowany ciąg elementów segmentalnych fonetycznych i lingwistycznych, uwzględniający gramatyczną, fleksyjną oraz semantyczną i prozodyczną strukturę mowy ciągłej o rozciągłości co najmniej zdań. Generalne rozwiązanie tak ujęte-

go problemu ARM, choć stanowi on w skali światowej przedmiot i cel intensywnych badań, jest jeszcze niemożliwe, choćby ze względu na niedostateczną wiedzę w zakresie fonetyki, językoznawstwa oraz przetwarzania sygnału mowy. Natomiast istnieją już realne możliwości rozwiązań systemowych rozpoznawania ograniczonych zbiorów elementów fonetycznych i lingwistycznych, które w wybranych przypadkach mogą być stosowane przy realizacji akustycznego wejścia użytkowo ukierunkowanego systemu komputerowego. W obecnym etapie badań, prowadzonych w Pracowni Akustyki Mowy Zakładu Akustyki Cybernetycznej (ZAC) IPPT-PAN, proces rozpoznawania mowy obejmuje automatyczne rozpoznawanie ograniczonego zbioru kilkudziesięciu elementów leksykalnych w postaci wymawianych w izolacji wyrazów języka polskiego, stanowiących podstawowy słownik komunikacji człowiek-maszyna.

d). Podobne ograniczenia dotyczą procesu syntezy mowy, realizowanego w układzie elektronicznym, stanowiącym integralną część podsystemu akustycznego wyjścia systemu dwustronnej komunikacji człowiek-maszyna. Zgodnie z założeniem, synteza mowy na obecnym etapie prac badawczych ZAC polega na wytwarzaniu ograniczonego zbioru izolowanych wyrazów, z których każdy wyraża określoną reakcję przyjętego systemu komunikacji człowiek-komputer i stanowi uporządkowaną sekwencję elementów fonetycznych niższego rzędu, np. diad, morfemów, sylab itp., zgromadzonych w pamięci komputera. Takiej koncepcji syntezy nie należy zatem utożsamiać z ogólnie przyjętym pojęciem programowanej syntezy mowy ciągłej za pomocą syntezatorów artykulacyjnych, parametryczno-widmowych (formantowych, wokoderowych, sekwencyjnych) i innych, używanych w pracach badawczych, a niekiedy także w zastosowaniach praktycznych.

e). Logiczną konsekwencją uproszczeń i ograniczeń, sformułowanych w punktach (c) i (d), jest przyjęty na obecnym etapie pracy system rozpoznawania izolowanych wyrazów, oparty na kryteriach decyzyjnych uwzględniających stosunkowo nieliczny, ale wystarczający dla ograniczonego zbioru leksykalnego zespół 10 cech artykulacyjnych sygnału mowy, oraz system syntezy oparty na uproszczonych i wyspecjalizowanych regułach syntezy operujących zespołem około 180 elementów fonetyczno-lingwistycznych opisanych za pomocą parametrów fonetyczno-akustycznych i odpo-

wiednio dobranych do założonego zbioru wyrazów, będących produktem syntezy.

Wymienione w powyższych punktach ograniczenia zakresu pracy i jej ukierunkowanie na potencjalne zastosowania praktyczne do celów robotyki uzasadnione są zarówno ogólnym poziomem wiedzy i stanem badań w dziedzinie analizy, automatycznego rozpoznawania i syntezy mowy w kraju na tle poziomu światowego, jak i stanem posiadanych przez ZAC skromnych środków technicznych, a zwłaszcza wyposażenia komputerowego. Warunki te w decydujący sposób wpłynęły na przyjęty sposób analizy akustycznego sygnału mowy. Jednakże za podjęciem badań w zawężonym zakresie zadecydowała ostatecznie konieczność niezbędnego i uzasadnionego gospodarczo rozwijania metod i systemów służących do bezpośredniej komunikacji akustycznej człowiek-komputer w języku naturalnym za pomocą mowy, początkowo w celu umożliwienia szerokiemu gronu nie wykwalifikowanych użytkowników korzystania z ogólnodostępnych systemów komputerowych, a obecnie również do celów dwustronnej komunikacji w systemach cybernetycznych «operator-robot».

3. Podstawa teoretyczna i doświadczalna badań

Omawiany obecnie etap badań nie powstał w próżni, lecz stanowi konsekwentną kontynuację długofalowego programu badawczego prowadzonego w Pracowni Akustyki Mowy ZAC od roku 1960 w zakresie analizy, przetwarzania i syntezy mowy polskiej, którego dotychczasowe wyniki zostały ujęte w postaci ponad 50 prac naukowych. Główne kierunki prowadzonych w tym okresie badań obejmowały następujące zagadnienia: (a) Syntezę mowy, (b) Automatyczne rozpoznawanie wybranych elementów fonetycznych i lingwistycznych języka polskiego oraz (c) Akustyczne aspekty komunikacji człowiek-maszyna za pomocą mowy.

Program badawczy zainicjowano opracowaniem teoretycznych podstaw syntezy formantowej poszczególnych klas fonemów i połączeń fonematycznych języka polskiego, początkowo samogłosek (KACPROWSKI, 1962; KACPROWSKI, MIKIEL, 1963), spółgłosek nosowych (KACPROWSKI, 1963), a następnie sylab C-V (KACPROWSKI, MIKIEL, 1965; KACPROWSKI, 1965/66), ze szczególnym uwzględnieniem spółgłosek nosowych i zwartych. Weryfikację eksperymentalną za-

łożeń teoretycznych przeprowadzono w zbudowanym w tym celu w ZAC unikalnym parametrycznym syntezatorze formantowym mowy SYNFOR (KACPROWSKI, MIKIEL, 1968; KACPROWSKI, 1969), który przez wiele lat był cennym narzędziem badawczym, umożliwiającym również badanie parametrów fonetyczno-akustycznym mowy polskiej, m.in. metodą analizy przez syntezę.

W drugiej połowie lat 60-tych podjęto próby sformułowania teoretycznych podstaw procesu automatycznego rozpoznawania mowy (KACPROWSKI, 1967a), zawężając kolejno problem ogólny do poszczególnych klas fonemów, przede wszystkim samogłosek (KACPROWSKI, 1967b), metodą segmentacji widma (KACPROWSKI, GUBRYNOWICZ, 1967), a następnie polskich spółgłosek trących przy zastosowaniu metody przejść sygnału mowy przez zero (GUBRYNOWICZ i in., 1973). Niezależnie od kontynuowania prac nad rozpoznawaniem wybranych elementarnych segmentów mowy polskiej, podjęto pierwsze próby rozpoznawania izolowanych wyrazów, początkowo za pomocą układów wyłącznie analogowych (KACPROWSKI, MOTYLEWSKI, 1969), a następnie metodami cyfrowymi, głównie w oparciu o analizę rozkładów gęstości przejść przez zero. Metoda tej analizy została zweryfikowana na wybranych fonematycznych klasach dźwięków mowy, a nieco później została zastosowana do automatycznego rozpoznawania całych wyrazów wymawianych w izolacji. Eksperyment automatycznego rozpoznawania izolowanych wyrazów przeprowadzono na zamkniętym zbiorze dziesięciu cyfr od 0 do 9 wymawianych dowolnym głosem, uzyskując bardzo zachęcające wyniki (GUBRYNOWICZ 1974a).

Pozytywne wyniki prac badawczych w dziedzinie syntezy i rozpoznawania mowy zwróciły uwagę na możliwości wykorzystania tych procesów do przekazywania wiadomości w układzie informacyjnym maszyna-człowiek (KACPROWSKI, 1970) i stały się bodźcem do przeanalizowania akustycznych aspektów problemu komunikacji człowiek-komputer w języku naturalnym (KACPROWSKI, 1972) z zastosowaniem akustycznego wyjścia komputera w postaci parametrycznego syntezatora formantowego (KACPROWSKI, MIKIEL, 1971) i z zastosowaniem akustycznego wejścia (KACPROWSKI, 1974; GUBRYNOWICZ, 1974b).

Wyniki omówionych wyżej, z konieczności w największym skrócie, pionierskich w skali krajowej badań stały się podstawą

do opracowania założeń koncepcyjnych i organizacyjnych systemu dwustronnej komunikacji człowiek-komputer w języku naturalnym za pomocą głosu. Realizacja tego systemu stanowiła przedmiot i cel prac badawczych, prowadzonych w ZAC w latach 1976-1985 głównie pod kątem zastosowań w układach informatycznych ogólnego i specjalnego przeznaczenia (GUBRYNOWICZ, MIKIEL, 1980). Program ten będzie w latach 1986-1990 i następnych rozwijany i adaptowany do zastosowań w robotyce, przy wykorzystaniu dotychczasowych doświadczeń oraz rozwiązań koncepcyjnych i systemowych Zakładu Akustyki Cybernetycznej w zakresie komunikacji człowiek-komputer za pomocą sygnału mowy, z uwzględnieniem szczególnych warunków, jakie rozwiązaniom tym stawia specyficzny charakter integralnego członu w rozważanym systemie komunikacji, to jest konkretnego robota, jakim jest na przykład maszyna robocza lub pojazd, wyposażony w elementy sztucznej inteligencji. Celowi temu służy niniejsze opracowanie.

4. Organizacja systemu dwustronnej komunikacji «człowiek - komputer»

4.1. Ogólna charakterystyka posiadanych środków technicznych do realizacji systemu doświadczalnego.

Zasadniczym elementem doświadczalnego systemu dwustronnej komunikacji człowiek-komputer, do którego mogą być wprowadzane i z którego są jednocześnie wyprowadzane informacje w postaci akustycznego sygnału mowy, był w ZAC do roku 1980 minikomputer MERA-304 o bardzo małej mocy obliczeniowej i pamięci operacyjnej 8 kB. Od roku 1980 prace były prowadzone z wykorzystaniem minikomputera MERA-400 z pamięcią 64 kB, a od roku 1983 także przy użyciu mikrokomputera ZX 81, wyposażonego w pamięć 64 kB. Obecnie Zakład dysponuje ponadto mikrokomputerem Sinclair Spectrum Plus 48 K. Tym nie mniej w początkowym okresie prac badawczych 1976-1980 ograniczenia, wynikające ze stosowania minikomputera MERA-304, rzutowały w sposób decydujący na tworzenie i wybór rozwiązań koncepcyjnych i technicznach systemu.

Przyjęto, że akustyczny sygnał mowy, przed wprowadzeniem go do komputera, musi być poddany, za pomocą odpowiednich układów analogowych, znacznym przekształceniom, umożliwiającym wy-

dobycie z niego najistotniejszych o nim informacji, na podstakórych można następnie dokonać jego transformacji na ciąg dyskretnych znaków. Zakłada się przy tym, że każdy z tych znaków reprezentuje jeden i tylko jeden element zbioru symboli, stosowanych do opisu rozpoznawanego sygnału. Opis ten jest realizowany automatycznie przez komputer w oparciu o określony zestaw wybranych parametrów fizycznych, mierzonych i przekształcanych do postaci cyfrowej przez zbudowany w Pracowni Akustyki Mowy ZAC wyspecjalizowane urządzenie zwane Parametrycznym Analizatorem Mowy (PAM), współpracujące w systemie on-line z minikomputerem (GUBRYNOWICZ, 1979). Dzięki takiemu rozwiązaniu do minimum zmniejszono proces przetwarzania sygnału w samym komputerze, co spowodowało, że proces rozpoznawania może się odbywać w czasie tylko kilkakrotnie dłuższym od czasu rzeczywistego, mimo bardzo ograniczonej mocy obliczeniowej minikomputera.

Za podstawę opisu sygnału mowy w postaci ciągu wybranych elementów segmentalnych mowy przyjęto opis artykulacyjny o uniwersalnym charakterze, w zasadzie nie podporządkowany przyjętemu słownikowi rozpoznawanych wyrazów. Szersze omówienie tego zagadnienia jest podane w punkcie 5.1 niniejszego opracowania.

Ograniczenia te rzutowały również na sposób realizacji akustycznego wyjścia systemu, w którym proces generacji sygnału mowy odbywa się za pomocą parametrycznego syntezatora formantowego, sterowanego z minikomputera zgodnie z regułami syntezy, zapisanymi w jego pamięci. Reguły te opisują przebieg zmian w czasie wybranych parametrów fizycznych syntezy, niezbędnych do wygenerowania odpowiedniego akustycznego sygnału mowy za pomocą syntezatora SYNKOM-6b, omówionego w punkcie 6.2. Za podstawę opisu sygnału mowy przy jego syntezie przyjęto klasyczny opis fonetyczno-akustyczny, który stworzył podstawy badań prowadzonych od roku 1960 nad procesem syntezy mowy polskiej i który umożliwił najpełniejsze wykorzystanie uzyskanych dotychczas wyników przy realizacji wyjścia akustycznego.

Oba urządzenia, wejście i wyjście akustyczne, zostały tak opracowane, że od strony jednostki centralnej symulują one odpowiednio czytnik i perforator. Uprościło to znacznie współpracę tych urządzeń z jednostką centralną i zredukowało w znacznym stopniu prace nad oprogramowaniem systemu komunikacji, a jedno-

cześnie nadało przyjętym rozwiązaniom technicznym pewną uniwersalność, umożliwiając podłączanie urządzeń akustycznego wejścia i wyjścia do dowolnych systemów komputerowych.

4.2. Wybór słownika komunikacji człowiek-maszyna.

Przy projektowaniu urządzeń akustycznego wejścia i wyjścia dowolnego systemu komputerowego o konkretnym przeznaczeniu użytkowym należy ściśle określić zadania, których wykonywanie urządzenia te mają umożliwiać. Od tego zależą bowiem rozmiary i skład leksykalny słownika rozpoznawanych i generowanych wyrazów oraz stosowane metody ich identyfikacji i syntezy.

W opracowywanym systemie doświadczalnym o nie określonym a priori przeznaczeniu użytkowym założono, że wejście akustyczne ma umożliwiać sterowanie za pomocą mowy pracą jednostki centralnej komputera oraz przepływem informacji i danych między nią i urządzeniami zewnętrznymi. W zasadzie przewiduje się, że sterowanie to będzie się odbywać za pomocą instrukcji, wypowiadanych w formie izolowanych wyrazów i utworzonych z nich krótkich zdań o ściśle zdefiniowanej uproszczonej strukturze składniowej. Zasady doboru słownika i ograniczeń nakładanych na syntaktyczną strukturę instrukcji zostały opracowane już wcześniej (GUBRYNOWICZ, 1974b). Przyjęto wówczas pewien słownik wyrazów, umożliwiający sterowanie maszyny cyfrowej za pomocą mowy, niezależnie od konkretnego przeznaczenia wejścia akustycznego, i zapewniający dość duże możliwości komunikacji operatora z systemem komputerowym o rozbudowanej konfiguracji, przy odpowiednim jego oprogramowaniu. Słownik ten początkowo składał się z 39 wyrazów, lecz w toku prowadzonych badań rozbudowano go do 60 wyrazów, którymi są cyfry od 0 do 9 i zbiór komend, określających wykonywane funkcje oraz nazwy (hasła) związane z obsługą urządzeń zewnętrznych. Przyjmuje się ponadto możliwość wymiany słowników oraz ich dalszą rozbudowę w przypadku zmiany przeznaczenia lub konfiguracji systemu.

Zakłada się, że przy sterowaniu systemu komputerowego sygnałem mowy instrukcje wejściowe będą formowane w postaci ciągu wyrazów wymawianych w izolacji, to jest ograniczonych obustronnie krótkimi przerwami o czasie trwania ok. stukilkudziesięciu milisekund, a operator będzie miał możliwość kontrolowania po-

prawności rozpoznawania dzięki powtarzaniu zidentyfikowanych wyrazów przez wyjście akustyczne, generujące sygnał mowy syntetycznej. Oczywistym warunkiem użyteczności systemu komunikacji człowiek-komputer, lub mówiąc najogólniej człowiek-maszyna, jest niezależność poprawności rozpoznawania od cech indywidualnych głosu operatora, jednak przy ścisłym przestrzeganiu przez niego reguł posługiwania się systemem, zarówno w znaczeniu poprawnej i znormalizowanej artykulacji, zachowania określonego poziomu, tempa i intonacji wypowiedzi, jak i właściwym użytkowaniu przetworników wejściowych (mikrofonów) w określonych warunkach akustycznych otoczenia.

W docelowym systemie komunikacji operator-maszyna robocza (pojazd), skład leksykalny i objętość słownika rozpoznawanych wyrazów zależeć będą od założonego przeznaczenia użytkowego konkretnego robota, jego typu, rodzaju, akustycznych warunków pracy oraz zastosowanych w nim rozwiązań technicznych. W związku z tym słownik wyrazów (komend, instrukcji, nazw itp.), rozpoznawanych przez akustyczne wejście systemu, musi być ustalony w ścisłym porozumieniu z jego projektantami, konstruktorami i potencjalnymi użytkownikami. To samo dotyczy słownika wyrazów, generowanych przez akustyczne wyjście systemu w postaci mowy syntetycznej i tworzących określone informacje wyjściowe, dotyczące aktualnych parametrów stanu zarówno samego robota, jak i jego środowiska pracy oraz obiektu manipulacji. Jak już wspomniano, w słowniku wyjściowym systemu muszą być ponadto zawarte wyrazy słownika wejściowego, ze względu na konieczność powtarzania zwrotnego informacji wejściowych do celów bezpieczeństwa i kontroli poprawności działania systemu. Jako regułę generalną należy przyjąć zasadę, że rozkazy i instrukcje wejściowe oraz informacje wyjściowe, które dotyczą stanów i sytuacji awaryjnych, nie mogą tworzyć sekwencji dłuższych niż dwu- lub co najwyżej trzywyrazowe (EPPLE i in., 1984).

5. Organizacja wejścia akustycznego

5.1. Podstawy artykulacyjnego opisu sygnału mowy.

Jednym z podstawowych problemów w złożonym procesie automatycznego rozpoznawania mowy jest wybór właściwego sposobu opisu sygnału mowy, opartego na określonym zespole mierzalnych

parametrów fizycznych sygnału. Ponieważ najmniejszą lingwistyczną jednostką znaczeniową mowy jest fonem, przy czym liczba fonemów dla poszczególnych języków europejskich zawiera się w granicach od 30 do 40, przeto istnieje naturalna tendencja do stosowania opisu fonetyczno-akustycznego, zgodnie z którym rozpoznawany sygnał byłby przekształcony do postaci ciągu fonemów. Jednak ogromna złożoność procesu wytwarzania mowy powoduje, że na ogół nie jest możliwe jednoznaczne przyporządkowanie określonym segmentom sygnału akustycznego nazw fonemów. Przyczyną tego jest m.in. fakt, że zmierzone parametry fizyczne sygnału są bardziej jednoznacznie zależne od konfiguracji artykulacyjnej narządu mowy, niż od struktury fonetycznej sygnału mowy, która jest związana z wyższymi poziomami przetwarzania informacji lingwistycznych przez człowieka. Z tego powodu przy opracowywaniu systemu akustycznego wejścia komputera w ZAC przyjęto opis artykulacyjny sygnału mowy jako podstawowy, lecz ograniczony do takiego zakresu, jaki jest niezbędny do rozpoznawania założonego zbioru wyrazów.

W przyjętym przez Sloata (SLOAT i in., 1978) opisie artykulacyjnym dźwięków mowy użyto następujące kryteria:

(a) Sposób artykulacji - określający jak i w jakim stopniu poszczególne efektory narządu mowy (język, podniebienie miękkie, wargi itd.) uczestniczą w wytwarzaniu danego dźwięku;

(b) Rodzaj pobudzania - określający typ źródła pobudzającego do drgań powietrze, znajdujące się w ponadkrtaniowej części kanału głosowego;

(c) Miejsce artykulacji - określające w pierwszym przybliżeniu położenie punktu w jamie ustnej, w którym następuje maksymalne przewężenie toru głosowego.

Kryteria te stanowią tylko część pełnego opisu artykulacyjnego, stosowanego w tradycyjnej fonetyce, jednak są one podstawowe i w obecnym stanie badań prowadzonych w ZAC wydaje się, że będą wystarczające do ogólnej i zarazem jednoznacznej klasyfikacji wybranych grup dźwięków mowy, z których zbudowane są wyrazy typowego uproszczonego słownika komunikacji człowiek-maszyna.

Ze względu na sposób artykulacji dźwięki mowy można podzielić na dwie zasadnicze klasy: r e s o n a n t y i o b s t r u e n t y (BIEDRZYCKI, 1978). Resonanty, do których zalicza-

ją się przede wszystkim samogłoski, spółgłoski nosowe, boczne i płynne, charakteryzują się z artykulacyjnego punktu widzenia tym, że przepływ powietrza przez jamę ustną i/lub nosową odbywa się bez przeszkód, a powstające przy tym ewentualne zawirowania powietrza są przypadkowe i nie stanowią cechy charakterystycznej tych dźwięków. Natomiast przy wymawianiu dźwięków typu obstruent, swobodny przepływ powietrza z głośni do wylotu ust ulega zaburzeniu wskutek chwilowego zamknięcia drogi przepływu lub powstania w określonym miejscu toru głosowego tak znacznego przewężenia, że powstają w jego okolicy zawirowania powietrza, będące źródłem szumu. Te zasadnicze różnice w sposobie artykulacji mają wyraźne odbicie w akustycznej strukturze analizowanego sygnału mowy. Rezonanty odznaczają się dużym, w porównaniu z obstruentami, poziomem energii globalnej, wyrażonym relatywnie wysokim poziomem obwiedni sygnału, jak również koncentracją energii w dolnym zakresie widma częstotliwości akustycznych, poniżej 1000 Hz.

W następnym kroku klasyfikacji, uwzględniającej sposób artykulacji, obie wymienione wyżej klasy dźwięków mowy mogą być podzielone na dalsze podgrupy. Klasa rezonantów dzieli się na cztery podgrupy, obejmujące odpowiednio dźwięki samogłoskowe, nosowe, boczne oraz płynne. Samogłoski ustne (i, y, e, a, o, u) od pozostałych rezonantów wyróżnia to, że przy ich wymawianiu powietrze przepływa od głośni przez jamę ustną w miarę swobodnie, przy ustalonej konfiguracji toru głosowego gardłowo - ustnego. Przy artykulacji spółgłosek nosowych (m, n, ɲ, ŋ) jama ustna jest całkowicie zamknięta i powietrze przepływa swobodnie przez kanał nosowy, przy ustalonej konfiguracji toru głosowego, która różni się dla poszczególnych spółgłosek nosowych jedynie ukształtowaniem i objętością akustycznie czynnego obszaru wnęki jamy ustnej. Spółgłoski boczne (l, ɫ) odznaczają się tym, że przy ich artykulacji przednia lub środkowa część języka jest uniesiona do góry i przylega do podniebienia twardego, blokując jamę ustną wzdłuż jej linii środkowej, a powietrze opływa z boków powstałą wskutek tego przeszkodę, przy ustalonej konfiguracji wnęki jamy ustnej. Konfiguracja ta zmienia się jedynie podczas wymawiania spółgłoski płynnej (j), kiedy następuje ruch języka do lub z położenia, charakterystycznego dla sąsiadującej

z nią dominującej samogłoski, gdyż ten typ głosek można wymówić jedynie w co najmniej jednostronnym kontekscie samogłoskowym.

Również dźwięki typu obstruent można podzielić ze względu na sposób artykulacji, oparty na analizie przepływu powietrza przez jamę ustną, na trzy zasadnicze grupy spółgłoskowe: zwarte, trące i zwarto-trące. Jeżeli podczas procesu wymawiania głoski następuje zamknięcie jamy ustnej i chwilowe zablokowanie strumienia powietrza, poprzedzające właściwy proces artykulacji, który występuje przy otwartej jamie ustnej, to dźwięki wytwarzane w ten sposób zaliczane są do klasy zwartych. Natomiast w wypadku gdy w jamie ustnej tworzy się przewężenie w postaci tak wąskiej szczeliny, że przy ciągłym przepływie powietrza przez nią powstają zawirowania będące źródłem szumu, to wymawiane dźwięki należą do klasy trących. Trzecia grupa głosek łączy w sobie cechy artykulacyjne obu poprzednich klas, to znaczy najpierw występuje pełne zamknięcie jamy ustnej, a następnie stosunkowo niewielkie jej otwarcie, tworzące wąską szczelinę w poprzednim miejscu zwarcia, w której powstają zawirowania przepływającego przez nią powietrza podobnie jak w przypadku dźwięków trących. Wytwarzane w ten sposób dźwięki należą do klasy zwarto-trących.

W przedstawionym podziale klasyfikacyjnym dźwięków mowy na klasy rezonantów i obstruentów osobną grupę stanowią dźwięki typu /r/. Grupa ta jest trudna do jednoznacznego opisu, ponieważ ma kilka wariantów artykulacyjnych, zależnych silnie od zadanego kontekstu samogłoskowego i spółgłoskowego. Dźwięki typu /r/ są wymawiane przy cofnięciu się tylnej części języka do ścianki nagłośni, w wyniku czego powstaje przewężenie, przy jednoczesnym uniesieniu środkowej części języka ku podniebieniu twardemu. W zależności od kontekstu, dźwięki typu /r/ mogą być wymawiane jako dźwięki klasy obstruent (tzw. wariant uderzeniowy) lub mogą być polisegmentalne i składać się z krótkich występujących naprzemian segmentów typu rezonant i obstruent (tzw. wariant drżący).

Odmienny sposób artykulacji poszczególnych klas w obrębie rezonantów, a zwłaszcza obstruentów, znajduje swoje odbicie w stosunkowo łatwo mierzalnych parametrach fizycznych akustycznego sygnału mowy. I tak na przykład w przypadku wymawiania gło-

sek zwartych i zwarto-trących znamienne jest występowanie poprzedzającej je bezpośrednio pauzy w sygnale o czasie trwania 20 do 120 ms, po której następuje skokowy wzrost poziomu energii globalnej sygnału, w przypadku głosek zwarto-trących skoncentrowanej głównie w górnym zakresie częstotliwości widma, powyżej 4 kHz. Relatywnie wysoki poziom energii w górnym zakresie częstotliwości widma jest również typowy dla głosek trących, z tą jednak różnicą, że mają one bardziej stacjonarny charakter niż wyżej wymienione dwie klasy głosek: zwartych i zwarto-trących.

W przypadku rezonantów odmienna artykulacja poszczególnych klas dźwięków znajduje również swoje odbicie w fizycznej strukturze sygnału, ale już w mniej wyraźny i jednoznaczny sposób, niż w przypadku obstruentów. Wszystkie rezonanty mają na ogół wyraźną strukturę formantową i dopiero przy dokładniejszej analizie widmowej można dostrzec istotne zachodzące między nimi różnice. I tak samogłoski mają najbardziej stabilną strukturę widmową o wyraźnych maksimach obwiedni widma, tworzących tzw. formanty, lecz bez charakterystycznych dla pozostałych rezonantów minimów, to jest antyformantów, które odpowiadają zerom funkcji transmitancji toru głosowego. Rozróżnienie spółgłosek bocznych od nosowych może być dokonane w oparciu o szczegółową analizę struktury formantowej, obejmującą również sąsiadujące z tymi spółgłoskami samogłoski. Należy podkreślić, że rozróżnianie poszczególnych dźwięków mowy typu rezonant jest dość złożone i wymaga bardzo szczegółowej analizy sygnału, uwzględniającej również efekt koartykulacji, który polega na zmianie warunków, tj. sposobu i miejsca artykulacji, a tym samym również struktury widmowej niektórych dźwięków mowy pod wpływem oddziaływania na nie dźwięków z nimi sąsiadujących. Wskutek tego zjawiska wyniki analizy dyskryminacyjnej sygnału nie zawsze jeszcze są w pełni zadowalające.

W oparciu o <u>rodzaj pobudzenia</u> kanału głosowego można wyróżnić trzy podstawowe klasy dźwięków mowy: o pobudzaniu krtaniowym, ponadkrtaniowym i mieszanym. W przypadku mowy polskiej pobudzanie krtaniowe jest dźwięczne, gdyż strumień powietrza wypływający z płuc jest modulowany przez drgające fałdy głosowe, w wyniki czego widmo sygnału pobudzenia, wytwarzanego przez

źródło krtaniowe, ma strukturę dyskretną (harmoniczną) o częstotliwości podstawowej F_0 równej częstotliwości drgań fałdów głosowych. W przypadku pobudzania ponadkrtaniowego kanał głosowy jest pobudzany sygnałem szumowym (aperiodycznym), powstającym wskutek zawirowań strumienia powietrza w przewężeniu utworzonym w części ponadkrtaniowej toru głosowego. Ponadto istnieje oddzielna grupa dźwięków mowy o pobudzeniu mieszanym, w którym występują jednocześnie oba rodzaje pobudzania.

Podział dźwięków mowy ze względu na rodzaj pobudzania dotyczy wyłącznie obstruentów, ponieważ rezonanty są z założenia przebiegami dźwięcznymi. Natomiast wszystkie dźwięki typu obstruent, ze względu na akustyczne cechy źródła pobudzającego kanał głosowy, można podzielić na dźwięczne, bezdźwięczne (szumowe) oraz dźwięczno-szumowe.

Ostatnim parametremartykulacyjnym, stosowanym przy opisie dźwięków mowy, jest miejsce artykulacji. Ze względu na miejsce położenia przewężenia, utworzonego między odpowiednią częścią języka i górną powierzchnią jamy ustnej, lub miejsce, w którym występuje całkowite zablokowanie strumienia powietrza, dźwięki mowy można podzielić, idąc od wylotu ust wgłąb kanału głosowego, na dwuwargowe, zębowo-wargowe, zębowe, zazębowe, dziąsłowe, podniebienne (inaczej zwane środkowo-językowymi) i tylnojęzykowe. Podział ten jest bardzo szczegółowy i w zasadzie dla poszczególnych głosek położenie miejsca przewężenia kanału głosowego lub blokady strumienia powietrza może ulegać pewnym zmianom, w zależności od miejsca artykulacji sąsiednich dźwięków, wskutek wspomnianego poprzednio zjawiska koartykulacji.

Zmiany położenia miejsca artykulacji znajdują swoje odbicie przede wszystkim w strukturze widmowej dźwięków mowy. I tak na przykład w przypadku wymawiania głosek typu rezonant wpływ miejsca artykulacji wyraża się zmianami częstotliwości pierwszego (F1) i drugiego (F2) formantu. Zwłaszcza zmiany częstotliwości drugiego formantu, wywołane zmianą miejsca artykulacji, są szczególnie widoczne. Przy przesuwaniu się miejsca artykulacji od tyłu ku przodowi toru głosowego następuje szybki wzrost częstotliwości drugiego formantu przy jednoczesnej tendencji malejącej częstotliwości pierwszego formantu. W przypadku obstruentów, zwłaszcza głosek trących, zmienia się widmo w górnym

zakresie częstotliwości przy zmianie miejsca przewężenia. Gdy przesuwa się ono ku przodowi, drugi moment widma M_2, który określa położenie środka ciężkości widma, wzrasta. Moment ten można wyznaczyć bardzo łatwo, stosując metodę przejść przez zero (GUBRYNOWICZ, 1974a).

Szczegółowe informacje dotyczące wytwarzania i artykulacji dźwięków mowy oraz ich cech dystynktywnych, wyrażonych w terminach charakterystycznych parametrów fonetyczno-akustycznych, znaleźć można w pracy (JASSEM, 1973).

Dokładne określenie położenia miejsca artykulacji nie jest w praktyce pomiarowej proste, a ponadto, ze względu na efekt koartykulacji, położenie to nie jest w zadanych granicach niezmienne. Z tego powodu, w przyjętym w ZAC opisie artykulacyjnym dźwięków mowy, dogodniej jest ograniczyć się do przybliżonego określenia ułożenia masy języka, które jest związane w pewien sposób z miejscem artykulacji w przypadku obstruentów, a jednocześnie opisuje bardziej precyzyjnie artykulację rezonantów. Stosując do tego rodzaju opisu trzy miejsca ułożenia masy języka: przednie, środkowe i tylne, można utworzyć wspólny opis miejsc artykulacji dla rezonantów i obstruentów, będący jednocześnie w zgodzie ze współczesną klasyfikacją artykulacyjno - akustyczną (WIERZCHOWSKA, 1980). Przyjęto zatem podział dźwięków mowy ze względu na ułożenie masy języka na trzy klasy, gdyż na podstawie zmian częstotliwości drugiego momentu widma M_2 i zmian częstotliwości drugiego formantu F2 można określić położenie masy języka w jednej z trzech stref toru głosowego.

Tak więc w oparciu o opracowany w ZAC system opisu artykulacyjnego dźwięków mowy, można je klasyfikować w sposób wielostopniowy, a choć w obecnym etapie badań jest on trzystopniowy, ograniczony do sposobu i miejsca artykulacji oraz rodzaju pobudzenia, to w miarę potrzeby będzie go można rozszerzać, uwzględniając w opisie dodatkowe cechy artykulacyjne, takie jak labializacja, palatyzacja, iloczas itp.

5.2. Parametryczny Analizator Mowy.

W oparciu o przedstawione w poprzednim punkcie podstawy przyjętego opisu artykulacyjnego dźwięków mowy, opracowano i wykonano w ZAC urządzenie nazwane Parametrycznym Analizatorem

Mowy (PAM), umożliwiające prosty i szybki pomiar wybranych parametrów fizycznych analizowanego sygnału, stanowiących podstawę tego opisu.

W poszczególnych kanałach analizatora mierzone są następujące parametry: a. ogólny poziom energii sygnału AO, b. poziom energii w dolnym zakresie częstotliwości, poniżej 800 Hz LP, c. poziom energii w górnym zakresie częstotliwości, powyżej 5 kHz HP, d. gęstość przejść przez zero ϑ, e. częstotliwość F_1 pierwszego formantu F1, f. częstotliwość F_2 drugiego formantu F2. Ponadto mierzony jest okres tonu krtaniowego T_o, który jest wykorzystywany do synchronizacji pomiaru wyżej wymienionych parametrów z impulsami tonu krtaniowego w przypadku analizy głosek dźwięcznych i dźwięczno-szumowych. Istnieje także możliwość pomiaru zespołu parametrów ze stałą kwantyzacją czasową co 10 ms. W przypadku dźwięków mowy o pobudzaniu wyłącznie szumowym czas kwantyzacji jest stały i równy 5 ms. Okres tonu krtaniowego T_o jest mierzony w sposób ciągły, zgodnie z metodą zastosowaną w urządzeniu zwanym INTONOGRAF-77 (MIKIEL i in., 1977).

Pomiary poziomów w dolnym LP i w górnym HP zakresie widma oraz w pełnym zakresie częstotliwości LO są realizowane za pomocą oryginalnego konwertera analogowo-cyfrowego (MIKIEL i in., 1975). Zastosowanie miary logarytmicznej do opisu zmian poziomu w poszczególnych zakresach częstotliwości umożliwia uzyskanie dużej skali dynamiki opisu, równej około 45 dB, przy użyciu do kodowania słowa o długości tylko 8 bitów.

Oryginalna metoda pomiaru częstotliwości dwóch pierwszych formantów F1 i F2 (WIĘŹLAK, 1979) polega na odfiltrowaniu zakresów częstotliwości obejmujących lokalne maksima obwiedni, które odpowiadają pierwszemu i drugiemu formantowi. Filtracja ta odbywa się w sposób dynamiczny za pomocą zestawu dwóch filtrów nadążnych, górno- i dolnoprzepustowego. Zakres przestrajania się filtru górnoprzepustowego w przypadku pomiaru pierwszego formantu jest uzależniony od częstotliwości tonu krtaniowego w taki sposób, aby pierwsza harmoniczna w widmie sygnału znalazła się poza zakresem przenoszenia tego filtru. Analogiczny filtr w torze pomiaru drugiego formantu ma automatycznie ograniczony zakres przestrajania w taki sposób, aby pierwszy formant leżał stale poniżej jego częstotliwości granicznej. Rozwiązanie

takie umożliwia zmniejszenie wzajemnego oddziaływania formantów na dokładność pomiaru ich częstotliwości metodą przejść przez zero i pozwala na ich dokładne rozgraniczenie nawet w przypadku, gdy są bardzo blisko siebie położone.

5.3. Algorytm wstępnej klasyfikacji rozpoznawanych wyrazów.

Po przesłaniu danych z Parametrycznego Analizatora Mowy do minikomputera następuje ich wstępna obróbka, której głównym celem jest wygładzenie przebiegów czasowych nagłych zmian mierzonych parametrów i wyeliminowanie przypadkowych błędów pomiarowych. Proces wygładzania realizowany jest za pomocą procedury tzw. trzypunktowej mediany, która eliminuje z wygładzanego przebiegu pojedyncze skokowe zmiany. W ramach wstępnej obróbki sygnału dokonywana jest również analizą, której celem jest wykrycie i zasygnalizowanie ewentualnych błędów, powstałych przy transmisji danych do minikomputera. Dotycza to zwłaszcza błędów, powstałych przy przełączaniu kanałów przez multiplekser. W przypadku wykrycia błędów tego typu odpowiednie cykle pomiarowe są usuwane z przebiegów jeszcze przed ich wygładzeniem.

W początkowym etapie prac badawczych nad opracowaniem algorytmu klasyfikacji skoncentrowano się na opisie dwóch podstawowych klas dźwięków: rezonant i obstruent (WIĘŹLAK i GUBRYNOWICZ, 1980). Do charakterystyki dźwięków mowy użyto relacji rozmytych (ZADEH, 1973), które z powodzeniem stosowano uprzednio do automatycznego rozpoznawania spółgłosek nosowych w mowie ciągłej (DE MORI i in., 1979).

Kolejnym etapem badań było sformułowanie opisu elementarnych segmentów sygnału w oparciu o cztery klasy dźwięków: rezonanty RS, głoski trące SS, obstruenty OC i zwarcia bezdźwięczne UC (WIĘŹLAK i GUBRYNOWICZ, 1981). Relacje rozmyte przypisują parametrom fizycznym każdego elementarnego segmentu opis, związany z czterema wyżej wymienionymi klasami. Relacje te stwierdzają, jaka jest m o ż l i w o ś ć, że dany segment elementarny przynależy do jednej lub kilku z czterech klas: RS, SS, OC i UC. Liczbowo m o ż l i w o ś ć ta jest określona współczynnikiem przynależności μ danego segmentu do odpowiedniej klasy (ZADEH, 1978).

Algorytm opisu segmentów elementarnych nie uwzględniał położenia segmentu w wyrazie, a jedynie chwilowe zmiany wartości trzech

parametrów widmowo-amplitudowych: AO, LP i HP. W następnym etapie wprowadzono zatem algorytm grupowania segmentów elementarnych w dłuższe segmenty, przy czym sposób analizy segmentów był zależny od ich położenia w wypowiedzi. Za pomocą procedury grupowania zrealizowano dwa następne etapy klasyfikacji, mianowicie:

- tworzono duże segmenty należące do czterech opisanych poprzednio klas RS, SS, OC i UC;
- analizując sekwencje opisów segmentów elementarnych, wyodrębniono pięć dalszych klas dźwięków mowy, tj. głoski /r/ (JR), głoski trące dźwięczne (VF), dwa typy głosek zwartych, zawierających silny (PS) lub słaby (PL) element szumowy, a także długie grupy rezonantów (RC).

Procedury te zaprojektowano początkowo w formie wieloprzebiegowego translatora skończonego (WIĘŹLAK i GUBRYNOWICZ, 1983), jednak ze względu na niejednoznaczności, występujące w opisie niektórych klas, zdecydowano się na użycie w translatorze rozmytego opisu wyjściowego, który każdemu wyjściowemu segmentowi przypisuje współczynnik przynależności do jednej lub kilku z dziewięciu wymienionych wyżej klas. Wynikiem działania procedury grupowania była sekwencja etykiet opisów wyjściowych wraz z odpowiednimi współczynnikami przynależności.

Działanie algorytmu segmentacji i klasyfikacji zweryfikowano doświadczalnie na wybranym uprzednio i wspomnianym w punkcie 4.2 obecnego opracowania zbiorze 60 wyrazów, umożliwiających sterowanie maszyny cyfrowej za pomocą mowy, niezależnie od jej konkretnego przeznaczenia użytkowego (GUBRYNOWICZ, 1974b). Podczas obecnej weryfikacji algorytmu wyrazy te wymawiało w izolacji 11 osób, w tym jedna kobieta, dysponujące wymową warszawskiej polszczyzny kulturalnej, posługując się stylem formalnym lub konwersacyjnym (BIEDRZYCKI, 1978). W oparciu o analizę wymienionych poprzednio trzech parametrów widmowo-amplitudowych w przypadku dwóch głosów męskich wyznaczono funkcje przynależności, występujące w relacjach rozmytych. Poprawność opisu czterech klas, sprawdzona na całym zbiorze uczącym, wyniosła 90%, tzn. tyle segmentów typu RS, SS, OC i UC zostało poprawnie opisanych.

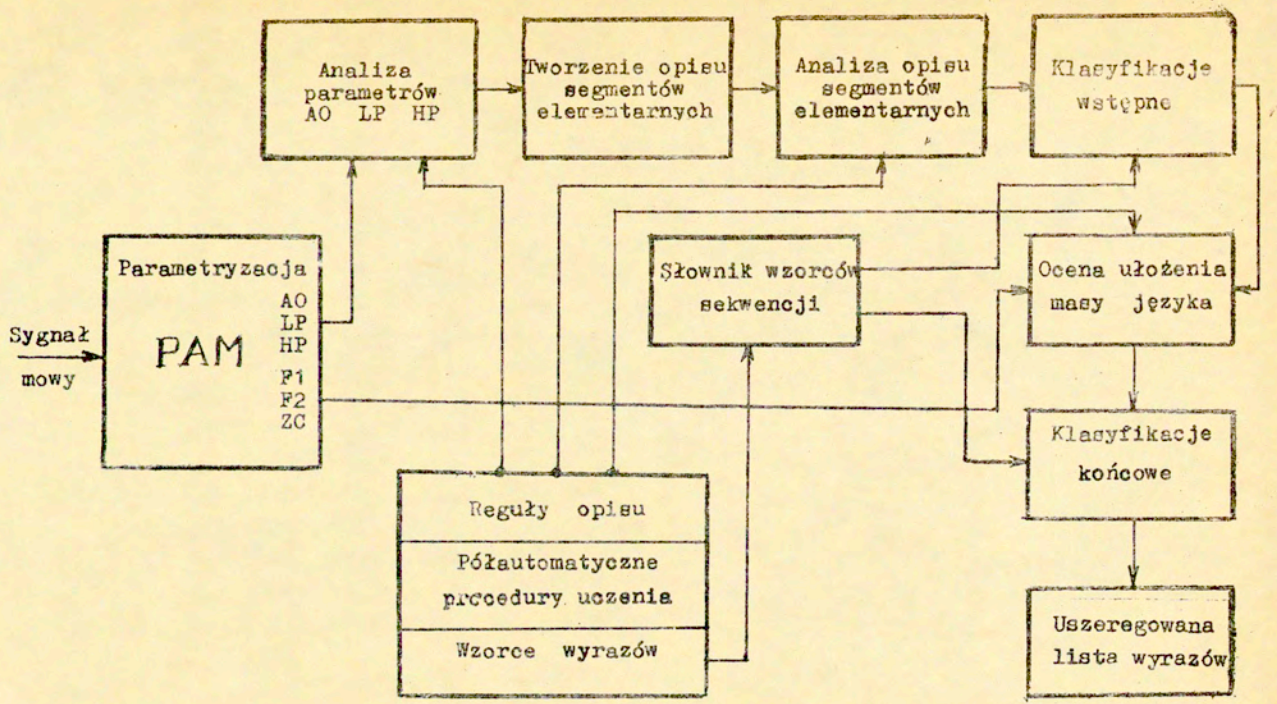
Algorytm kontekstowo zależny tworzono i sprawdzono w opar-

ciu o większy zbiór uczący, zawierający prawie wszystkie wypowiedzi. Na tym etapie badań dokładność na 9 klas mieściła się w granicach od 95% dla segmentów RS, SS, UC, PS do 70% dla PL. Błędy wynikały głównie z niedokładności metody parametryzacji (skala dynamiki opisu analizowanego sygnału wynosiła w tym przypadku około 35 dB) i stosunkowo małej liczby analizowanych parametrów, przy nie wykorzystaniu potencjalnych możliwości analizatora PAM (patrz punkt 5.2). Należy jednak pamiętać, że głównym celem tego etapu badań było wydobycie z sygnału maksymalnej ilości informacji, tj. liczby opisywanych klas, przy minimalnej liczbie analizowanych parametrów.

Niezależnie od tych prac prowadzono wstępne badania nad analizą położenia masy języka w czasie artykulacji, w oparciu o przebiegi formantowe i średnią gęstość przejść przez zero. Badając artykulację segmentów płynnych i samogłoskowych wykorzystano cyfrowy model symulacyjny toru głosowego gardłowo-ustnego, zrealizowany na minikomputerze MERA-400 (NOWAKOWSKA, 1983).

Wstępna weryfikacja doświadczalna opracowanej metody segmentacji i klasyfikacji grupowej dźwięków mowy w oparciu o parametryzację sygnału i jego opis artykulacyjny dała wyniki pozytywne i pozwala wnioskować, że metoda ta może znaleźć skuteczne zastosowanie w szczególnym przypadku akustycznego wejścia systemu komunikacji człowiek-komputer, przystosowanym do sterowania i nadzoru maszyn roboczych i pojazdów, głosem operatora.

Pełny i docelowy model systemu rozpoznawania izolowanych wyrazów według proponowanej metody, będącej przedmiotem aktualnych badań, przedstawiono na rys.1. W skład jego wchodzą, oprócz wspomnianych wyżej algorytmów, procedury automatycznego tworzenia wzorców sekwencji segmentów i ich porównywania na danym etapie opisu. Ważne jest to, że system jako taki nie stawia ograniczeń dotyczących wyboru i wymienności słownika komunikacji oraz jego rozszerzania, w celu adaptowania go do konkretnych zastosowań użytkowych. Ponadto system rozpoznawania, oparty na opisie artykulacyjnym dźwięków mowy, stosunkowo łatwo może być w pewnych granicach uniezależniony od cech osobniczych głosu operatora, a jego koncepcja ogólna pozwala w prosty sposób przejść od rozpoznawania izolowanych wypowiedzi (wyrazów) do rozpoznawania mowy ciągłej, wykorzystując do tego celu dłuższe segmenty sygnału, na przykład sylaby, pseudo-sylaby lub morfemy.



Rys.1. Uproszczony schemat blokowy modelu systemu rozpoznawania wyrazów (WIĘŹLAK i GUBRYNOWICZ, 1983)

6. Organizacja wyjścia akustycznego

6.1. Podstawy przyjętego systemu syntezy mowy.

W znanych w literaturze systemach syntezy mowy, stosowanych przy realizacji wyjścia akustycznego, dominują obecnie rozwiązania dwojakiego rodzaju. Pierwsze z nich polega na zarejestrowaniu wybranych elementów segmentalnych mowy naturalnej oraz stosowaniu opracowanych uprzednio reguł łączenia ze sobą tych elementów w celu uzyskania dłuższych segmentów sygnału mowy o założonym znaczeniu lingwistycznym i semantycznym, niosących określone informacje. System ten, zwany syntezą kompilacyjną, zaczyna dominować w dostępnych na rynku układach wyjściowych ze względu na stosunkowo łatwą realizację techniczną, ma jednak ograniczone możliwości zastosowań, a generowany sygnał mowy syntetycznej jest na ogół niskiej jakości. Drugi sposób syntezy polega na generowaniu sygnału mowy, wytworzonego w całości w układach elektronicznych, bądź przez modelowanie naturalnego układu artykulacyjnego organu mowy człowieka (syntezatory artykulacyjne), bądź przez modelowanie widma sygnału (np. syntezatory formantowe).

Jak już wspomniano w rozdziale 3 obecnego opracowania, ze względu na wieloletnie doświadczenia ZAC w dziedzinie formantowej syntezy mowy, umożliwiającej jednocześnie najbardziej ekonomiczną organizację wyjścia akustycznego, ten sposób syntezy przyjęto jako podstawowy przy realizacji doświadczalnego systemu komunikacji słownej komputer-człowiek. Analogicznie jak w przypadku wejścia akustycznego, w którym zasadniczy proces analizy sygnału mowy odbywa się za pomocą rozbudowanego układu analogowego, wyposażonego w układy kodujące wyniki pomiarów do postaci cyfrowej, zasadniczy proces syntezy formantowej jest realizowany za pomocą wyspecjalizowanego urządzenia zewnętrznego, natomiast z komputera wysyłane są tylko sygnały cyfrowe sterujące tym procesem, zgodnie z uprzednio ustalonymi regułami syntezy wybranych elementów segmentalnych mowy, występujących w przyjętym zbiorze wyrazów. Elementami tymi są morfemy i diady, których liczba niezbędna do syntezy np. wszystkich 60 wyrazów przyjętego słownika doświadczalnego systemu dwustronnej komunikacji człowiek-komputer wynosi odpowiednio 24 i 143. Proces syntezy

elementów segmentalnych jest realizowany w sposób automatyczny, zgodnie z wyznaczonymi wcześniej dla poszczególnych elementów funkcjami sterującymi, zapisanymi w pamięci minikomputera. Natomiast sam proces łączenia morfemów i diad może odbywać się pod kontrolą operatora, bądź odpowiedniego programu nadzorującego, organizującego wyprowadzania informacji przez wyjście akustyczne dwiema możliwymi, alternatywnie stosowanymi metodami, mianowicie:

(a) odpowiednio do rodzaju informacji, wprowadzanej do systemu komputerowego przez wejście akustyczne, jak to ma miejsce w opracowywanym układzie doświadczalnym człowiek-komputer i będzie miało miejsce w systemie docelowym komunikacji operator-robot przy powtarzaniu zwrotnym informacji wejściowych w postaci mowy syntetycznej do celów bezpieczeństwa i kontroli;

(b) z reguły w systemie komunikacji operator-robot do celów informowania obsługi o aktualnych parametrach stanu robota oraz jego środowiska pracy i obiektu jego manipulacji, jak również przy akustycznym sygnalizowaniu stanów awaryjnych za pomocą mowy syntetycznej.

Jest rzeczą oczywistą, że przypadek (b) może być zrealizowany tylko przy zastosowaniu czujników wewnętrznych i zewnętrznych oraz odpowiednich układów sensorycznych, w jakie wyposażone są z reguły roboty II i wyższych generacji, posiadające elementy sztucznej inteligencji.

6.2. Syntezator mowy SYNKOM-6b.

Przyjęte rozwiązanie konstrukcyjne syntezatora jest kompromisem między złożonością i rozbudową zastosowanych w nim układów analogowych, sterowanych napięciowo, takich jak modulatory, filtry formantowe itp., a ilością informacji niezbędnych do wygenerowania sygnału mowy, które należy wyprowadzić z minikomputera. Im w większym stopniu proces syntezy mowy będzie realizowany w maszynie cyfrowej, tym większą ilość informacji należy z niej wyprowadzić, a sam proces będzie trwał dłużej, wielokrotnie przekraczając rzeczywisty czas trwania generowanego sygnału. W przypadku zastosowania minikomputera MERA-304 pełne zamodelowanie syntezy mowy w postaci cyfrowej jest wręcz niemożliwe.

Opracowany i wykonany w ZAC syntezator SYNKOM-6b jest analogowym układem elektronicznym, w którym zastosowano podukłady, odtwarzające w domenie przebiegów elektrycznych funkcje: generacyjną i modulacyjną, odpowiednio źródła krtaniowego i toru głosowego. W syntezatorze można wyróżnić trzy człony funkcjonalne: kanał samogłoskowy i nosowy, które są pobudzane sygnałem harmonicznym, wytwarzanym przez generator tonu krtaniowego przy syntezie wszystkich głosek dźwięcznych, oraz kanał szumowy pobudzany z generatora szumu, który jest stosowany do syntezy spółgłosek trących i zwarto-trących bezdźwięcznych. Sygnały pobudzające są w każdym kanale odpowiednio formowane za pomocą specjalnych filtrów stałych i przestrajanych. W wyniku przeprowadzonych badań ustalono, że podczas dynamicznego procesu syntezy wystarczy przestrajać tylko niektóre z tych filtrów, dzięki czemu uzyskuje się dużą oszczędność w opisie sygnału, a co za tym idzie osiąga się znaczne uproszczenie samego sterowania tym procesem (KACPROWSKI i MIKIEL, 1968).

Sygnał wyjściowy kanału nosowego zależy od poziomu na wyjściu modulatora amplitudy A_N i od położenia trzech biegunów funkcji transmitancji, zamodelowanej za pomocą trzech połączonych ze sobą kaskadowo filtrów N1, N2 i N3. Najbardziej wielostronną rolę w procesie syntezy mowy pełni kanał zwany umownie samogłoskowym, który uczestniczy nie tylko w procesie generacji dźwięków samogłoskowych V, ale także diad typu V-V, V-N, N-V oraz C-V i V-C, w których segmenty spółgłoskowe mają strukturę dźwięczną lub dźwięczno-szumową. Wzajemne stosunki poziomów pobudzenia harmonicznego i szumowego są sterowane za pomocą odpowiednich modulatorów amplitudy A_O i A_C. Funkcja transmitancji tego kanału określana jest przez trzy przestrajane w czasie filtry formantowe F1, F2 i F3 oraz jeden filtr stały F4. Zwiększenie naturalności brzmienia samogłosek nazalizowanych i diad typu V-N i N-V zrealizowano przez dwustopniową skokową zmianę szerokości pasm filtrów formantowych B_N w kanale nosowym. Natomiast charakterystyczna dla spółgłosek płynnych zmiana szerokości pasma drugiego formantu jest zrealizowana przez dołączanie do tego filtru odpowiedniego układu, co odbywa się przez podanie sygnału binarnego B_R. Spółgłoski trące i zwarto-trące bezdźwięczne są realizowane w kanale szumowym za pomocą para-

metrów A_C, K1 i K2, a przy syntezie ich dźwięcznych odpowiedników wykorzystywany jest dodatkowo kanał samogłoskowy o pobudzaniu harmonicznym. Oddzielny filtr formantowy zastosowano do syntezy spółgłosek trących /f/ i /v/, których widma w zakresie wyższych częstotliwości znacznie różnią się od widm pozostałych spółgłosek trących w ich obu wariantach: bezdźwięcznym i dźwięcznym.

6.3. Organizacja współpracy syntezatora z minikomputerem MERA-304.

Jak wspomniano w poprzednim punkcie, w trybie pracy syntezatora SYNKOM-6b jako urządzenia akustycznego wyjścia, minikomputer MERA-304 spełnia tylko rolę bloku sterującego procesem syntezy, który realizowany jest w urządzeniu wyjściowym. Reguły syntezy, ustalone w wyniku badań i eksperymentów, są przetransformowane na zbiór funkcji sterujących, zgromadzonych w pamięci minikomputera. Sterowanie przez minikomputer procesem syntezy zgodnie z zadanym programem polega na wysyłaniu w określonych momentach czasu do Cyfrowej Jednostki Sterującej (CJS) kolejnych bloków informacyjnych o długości od 1 do 6 bajtów, w których są zakodowane stany poszczególnych układów syntezatora.

Przesyłanie sygnałów sterujących odbywa się w sposób sekwencyjny pod nadzorem programu, który wykonuje następujące działania:

- rozpoznanie wyrazu zadanego z klawiatury przez operatora;
- przepisanie do pola roboczego pamięci zbioru adresów programów sterujących kolejnych morfemów i diad oraz wzorców intonacyjnych, uzależnionych od pozycji i funkcji danego wyrazu we frazie;
- kontrola przesyłania informacji do jednostki sterującej syntezatora.

Wytworzenie zbioru wzorcowych przebiegów sterujących procesem syntezy morfemów i diad, stanowiących fonologiczne elementy składowe generowanego sygnału mowy w postaci izolowanych wyrazów, stanowi odrębny przedmiot badań, realizowanych w oparciu o analizy spektrograficzne i intonograficzne. Analizy te są wykonywane za pomocą analizatora kanałowego równoczesnego i intonografu, współpracujących on-line z minikoputerem przy użyciu

odpowiedniego oprogramowania. Wiele parametrów syntezy można wyznaczyć w podany wyżej sposób, jednakże parametry określające współpracę między poszczególnymi kanałami syntezatora, zwłaszcza między kanałem samogłoskowym i nosowym, a także wyznaczające stosunki poziomów pobudzania krtaniowego i szumowego w kanale samogłoskowym, mogą być ustalone wyłącznie na drodze syntezy eksperymentalnej metodą kolejnych prób i powtórzeń. Do tego celu konieczne jest zapewnienie możliwości nieograniczonej zmiany parametrów syntezy, sterowanej ręcznie lub półautomatycznie, oraz łatwości wprowadzania poprawek, przy jednoczesnym odsłuchu odpowiednich segmentów generowanego sygnału mowy syntetycznej.

7. Wnioski

Wyniki uzyskane w toku przeprowadzonych badań stwarzają podstawy teoretyczne i doświadczalne do realizacji technicznej systemu dwustronnej komunikacji akustycznej człowiek-maszyna w języku naturalnym, w oparciu o ograniczone, użytkowo ukierunkowane i wymienne zbiory wyrazów.

Dążenie do uzyskania optymalnych rozwiązań w tworzonym systemie spowodowało, że od strony wejścia akustycznego sygnał mowy jest opisywany w odmienny sposób, niż od strony wyjścia. Powodem tego jest różna metodyka, zastosowana do przekształcania sygnału mowy po stronie wejścia i jego formowania po stronie wyjścia akustycznego. Zastosowany system syntezy formantowej, oparty na wieloletnich doświadczeniach własnych ZAC, jest - jak się wydaje - w obecnych warunkach optymalny, bowiem wymaga reguł syntezy opisanych za pomocą tylko niewielu parametrów. Jednakże nie wszystkie z nich można w łatwy i jednoznaczny sposób określić w wyniku analizy rzeczywistego sygnału mowy naturalnej. Niektóre parametry syntezy można dobrać wyłącznie na podstawie prób eksperymentalnych, połączonych ze słuchową oceną generowanych sygnałów syntetycznych.

Do wstępnego rozpoznawania izolowanych wyrazów opracowano algorytm dwustopniowej klasyfikacji dźwięków mowy, opartej na ich opisie artykulacyjnym, przy zastosowaniu do tego celu relacji rozmytych. Zrealizowane oprogramowanie zostało częściowo zweryfikowane doświadczalnie na przyjętym zbiorze wyrazów, wypowiedzianych sześcioma głosami. Uzyskane pozytywne wyniki są

zachęcające i przemawiają za kontynuowaniem prac nad stosowaniem w technicznych systemach automatycznego rozpoznawania mowy artykulacyjnego opisu sygnału mowy przy użyciu relacji rozmytych.

Przyjęty sposób fonetyczno-akustycznego opisu sygnału mowy po stronie wyjścia akustycznego, polegający na przedstawieniu sygnału w postaci uporządkowanego ciągu morfemów i diad, z których są zbudowane syntetyczne wyrazy, umożliwia - przy zastosowanej formantowej metodzie syntezy - zoptymalizowanie tego procesu przez zmniejszenie liczby jego parametrów, to jest sygnałów sterujących. Zastosowanie parametrycznego syntezatora mowy, zrealizowanego w systemie cyfrowo sterowanego układu analogowego, umożliwia zminimalizowanie liczby niezbędnych działań wykonywanych przez minikomputer, przy zachowaniu stosunkowo wysokiej naturalności i zrozumiałości mowy syntetycznej oraz przekazywaniu informacji prozodycznych. Przeprowadzona analizą objętości informacyjnej sygnałów sterujących syntezator oraz sygnałów wprowadzanych do maszyny cyfrowej z Parametrycznego Analizatora Mowy wykazała, że możliwe jest zastąpienie minikomputera trzeciej generacji, do której należy MERA-304, przez mikrokomputer powszechnego zastosowania o słowie 8-bitowym i pamięci operacyjnej rzędu 48 kB składający się z kilku elementów o wysokiej skali integracji, i stanowiący podstawowy człon systemu akustycznego wejścia/wyjścia. Zmiana procesora da znaczne zmniejszenie nakładów pracy, związanych z opracowaniem i wdrożeniem oprogramowania systemu dwustronnej komunikacji człowiek-komputer za pomocą mowy.

Przystosowanie opracowywanego obecnie doświadczalnego systemu dwustronnej komunikacji akustycznej człowiek-komputer, o nie określonym a priori przeznaczeniu użytkowym, do konkretnego przypadku dwustronnej komunikacji słownej w rzeczywistym układzie cybernetycznym «operator - maszyna robocza (pojazd)» wymagać będzie oczywiście rozwiązania szeregu problemów natury badawczej i technicznej, wynikających ze specyfiki przeznaczenia i warunków pracy systemu docelowego. Zanim jednak parametry techniczne konkretnego robota, przeznaczonego do określonych zastosowań, zostaną sprecyzowane przez jego konstruktorów w porozumieniu z jego potencjalnymi użytkownikami, można już z góry

przewidzieć i wymienić kilka zagadnień, głównie natury akustycznej, które nie występowały lub były świadomie pomijane przy opracowywaniu systemu doświadczalnego, a które są istotne w układzie rzeczywistym, pracującym w warunkach przemysłowych.

Fizycznym wykładnikiem trudności tych warunków jest niekorzystny stosunek sygnału do szumu S/N, wywołany wysokim poziomem zakłóceń akustycznych (hałasów) w środowisku pracy robotów przemysłowych, w przeciwieństwie do sterylnych warunków akustycznych w specjalnych pomieszczeniach laboratoryjnych. Następnie pasmo przenoszenia elektronicznych układów transmisyjnych i przetwórczych obejmuje zakres 20 Hz - 20 kHz, podczas gdy w warunkach eksploatacji przemysłowej jest ograniczone niekiedy nawet do pasma telefonicznego 300 Hz - 3400 Hz. Ponadto systemem akustycznego wejścia robota z reguły nie będzie się posługiwać wyłącznie jeden jego operator, do którego głosu system ten będzie przystosowany, lecz różni operatorzy, o różnych cechach osobniczych głosu i różnych nawykach artykulacyjnych, co nawet przy rygorystycznym przestrzeganiu znormalizowanych warunków formułowania i artykulacji wypowiedzi stwarza układowi trudne warunki poprawnego działania.

Reasumując należy stwierdzić, że opracowanie użytkowego systemu dwustronnej komunikacji operator-robot za pomocą mowy , przystosowanego do pracy w warunkach przemysłowych, wymaga podjęcia dodatkowych prac badawczych, teoretycznych i doświadczalnych, których celem będzie zapewnienie poprawnego działania systemu w niekorzystnych warunkach zewnętrznych, wyrażonych małym stosunkiem sygnału do szumu S/N i ograniczoną szerokością pasma częstotliwości sygnału, przy jednoczesnym osiągnięciu małej wrażliwości układu automatycznego rozpoznawania mowy (wyrazów) na zmiany cech osobniczych głosu i nawyki artykulacyjne operatora.

Wydaje się, że szczegółowe i wszechstronne zbadanie i zoptymalizowanie działania układu akustycznego wejścia/wyjścia użytkowego systemu komunikacji ≪operator - maszyna robocza (pojazd)≫ w różnych warunkach zewnętrznych będzie mogło być najskuteczniej i najbardziej sprawnie przeprowadzone metodami symulacji komputerowej.

BIBLIOGRAFIA

1. BIEDRZYCKI L. (1978): Fonologia angielskich i polskich rezonantów. PWN, Warszawa.

2. DE MORI R., GUBRYNOWICZ R., LAFACE P. (1979): Inference of a knowledge source for the recognition of nasals in continuous speech. IEEE Trans. on Acoustics, Speech and Signal Processing, ASSP-27, 5, 538-550.

3. EPPLE W., DILLMANN R., WANHORST H. (1984): Speech communication enhances man-robot interfaces. Proc. 14th Intern. Symp. on Industrial Robots and 7th Intern. Conf. on Industrial Robot Technology, Gothenburg, Sweden, (ed. N. Martensson), 475-486.

4. GUBRYNOWICZ R., KACPROWSKI J., MIKIEL W., SKALSKI W. (1973): Klasyfikacja spółgłosek trących metodą przejść przez zero. Prace IPPT PAN, 40/1973.

5. GUBRYNOWICZ R. (1974a): Zastosowanie metody przejść przez zero do analizy sygnału mowy i automatycznego rozpoznawania ograniczonego zbioru wyrazów. Prace IPPT PAN, 71/1974.

6. GUBRYNOWICZ R. (1974b): Automatyczne rozpoznawanie mowy w komunikacji człowiek - elektroniczna maszyna cyfrowa. Prace IPPT PAN, 71/1974.

7. GUBRYNOWICZ R., DE MORI R., LAFACE P. (1978): La description au niveau acoustique des consonnes nasales prononcees dans un discours continu. Actes des 9emes Journees d'Etudes sur la Parole, Lannion, 287-296.

8. GUBRYNOWICZ R. (1979): Wejście akustyczne do minikomputera serii MERA-300. Mat. XXVI Otw. Seminarium z Akustyki PTA, Wrocław--Oławnica, 97-100.

9. GUBRYNOWICZ R., MIKIEL W. (1980): Analiza i synteza mowy w komunikacji człowiek-maszyna. W: Akustyka mowy i diagnostyka akustyczna (red. J. Kacprowski), 45-74, IPPT PAN, W-wa.

10. INDUSTRIAL ROBOT MAGAZINE (1983): Special Tenth Anniversary Issue on Decade of Robotics. Springer Verlag and IFS Publications Ltd.

11. ISO/TC 97 (1978): Study on standardization of terms and symbols relating to industrial robots in Japan. Document SC 8, N 434.

12. JASSEM W. (1973): Podstawy fonetyki akustycznej. PWN, W-wa.

13. KACPROWSKI J. (1962): Teoretyczne podstawy syntezy samogłosek polskich. Rozpr. Elektrotechn., VIII, 1, 127-203.

14. KACPROWSKI J. (1963): An approach to the synthesis of Polish nasal consonants by means of terminal-analog speech synthesizer. Proc. of Vibration Problems, 4, 3(16), 235-254.

15. KACPROWSKI J., MIKIEL W. (1963): Preliminary synthesis of Polish vowels by means of recurrently impulsed formant filters. Proc. of Vibration Problems, 4, 1(14), 27-41.

16. KACPROWSKI J., MIKIEL W. (1965): Synthesis of Polish C-V syllables by means of terminal-analog speech synthesizer. Proc. of 5th Intern. Congress on Acoustics, Liege, Report A-14.

17. KACPROWSKI J. (1965/66): Simplified rules for parametric synthesis of nasal and stop consonants in C-V syllables by means of terminal-analog speech synthesizer. Acustica, 16, 6, 356-364.

18. KACPROWSKI J. (1967a): Teoretyczne podstawy procesu automatycznego rozpoznawania mowy. Archiwum Akustyki, 2, 2, 123-151.

19. KACPROWSKI J. (1967b): Teoretyczne podstawy metody automatycznego rozpoznawania samogłosek. Archiwum Akustyki, 2,3, 227-254.

20. KACPROWSKI J., GUBRYNOWICZ R. (1967): Automatyczne rozpoznawanie samogłosek polskich metodą segmentacji widma. Prace IPPT PAN, 22/1967.

21. KACPROWSKI J., MIKIEL W. (1968): Recent experiments in parametric synthesis of Polish speech sounds. Proc. of 6th Inter.Congress on Acoustics, Tokyo, Report B-5-11.

22. KACPROWSKI J. (1969): Podstawy teoretyczne i realizacja techniczna formantowego syntezatora mowy SYNFOR II. Archiwum Akustyki, 4, 2, 199-220.

23. KACPROWSKI J., MOTYLEWSKI J. (1969): Automatyczne rozpoznawanie wyrazów metodą segmentacji widma sygnału mowy. Prace IPPT PAN, 29/1969.

24. KACPROWSKI J. (1970): Zastosowanie procesu syntezy mowy do do przekazywania wiadomości w układzie informacyjnym maszyna - człowiek. Archiwum Akustyki, 5, 1, 53-86.

25. KACPROWSKI J., MIKIEL W. (1971): The terminal-analog speech synthesizer as acoustic output of a computer. Proc. of 7th Intern. Congress on Acoustics, Budapest, Vol.3, Rpt.23-C-4.

26. KACPROWSKI J. (1972): Akustyczne aspekty problemu komunikacji człowiek - komputer w języku naturalnym. Archiwum Akustyki, 7, 3, 201-212.

27. KACPROWSKI J. (1974): Sygnał akustyczny w procesach sterowania i diagnostyki. Archiwum Akustyki, 9, 4, 376-388.

28. KACZMARCZYK A. (1984): Roboty przemysłowe lat osiemdziesiątych. Wydawnictwa Komunikacji i Łączności, Warszawa.

29. MIKIEL W., DRZEWIECKI J., JAKUBOWICZ J., KUPCZYK I., KUPCZYK K. (1975): Logarytmiczny konwerter analogowo-cyfrowy. Mat. XXII Otw. Seminarium z Akustyki PTA, Wrocław - Świeradów Zdrój, t.I, 136-140.

30. MIKIEL W., GUBRYNOWICZ R., HAGMAJER W. (1977): Intonograf - system do pomiaru i wizualizacji intensywności i melodii mowy. Mat. XXIV Otw. Seminarium z Akustyki PTA, Gdańsk - Władysławowo, t.I, 120-123.

31. MORECKI A., BUĆ J. (1981): Development of robotics in Poland 1977-1980. Proc. of 11th Int. Symposium on Industrial Robots, Tokyo, Japan.

32. NOWAKOWSKA W. (1983): Model symulacyjny toru głosowego gardłowo-ustnego. Prace IPPT PAN, 41/1983.

33. ROBOT INSTITUTE OF AMERICA (1982): RIA Worldwide robotics survey and directory 1982.

34. SLOAT C., TAYLOR S.H., HOARD J. (1978): Introduction to phonology. Prentice-Hall, Englewood Cliffs, N.J.

35. WIERZCHOWSKA B. (1980): Fonetyka i fonologia języka polskiego. Ossolineum, Wrocław.

36. WIĘŹLAK W.W. (1979): Bieżący pomiar częstotliwości formantowych w sygnale mowy. Mat. XXVI Otw. Seminarium z Akustyki PTA, Wrocław - Oleśnica, 189-192.

37. WIĘŹLAK W.W., GUBRYNOWICZ R. (1980): Wstępna klasyfikacja sygnału mowy w systemie rozpoznawania ograniczonego zbioru wyrazów. Mat. XXVII Otw. Seminarium z Akustyki PTA, Warszawy - Puławy, 192-195.

38. WIĘŹLAK W.W., GUBRYNOWICZ R. (1981): Representation of speech in terms of articulatory elements in an Isolated Word Recognition System. Proc. of the Fourth F.A.S.E. Symposium on Speech Acoustics, Venezia, 301-304.

39. WIĘŹlak W.W., GUBRYNOWICZ R. (1982): Articulatory description of speech signal in Isolated Word Recognizer. Proc. of ICASSP-82, Paris, 529-533.

40. WIĘŹLAK W.W., GUBRYNOWICZ R. (1983): The rules of articulatory description for isolated word recognition. Proc. 11th I.C.A. Toulouse Satellite Symposium, Toulouse, 77.

41. YONEMOTO K. (1984): Robotization and its future outlook in Japan. Proc.14th Intern. Conf. on Industrial Robot Technology. Gothenburg, Sweden, (ed. N. Markensson), 733-738.

42. ZADEH L.A. (1973): The concept of a linguistic variable and its application to approximate reasoning. American Elsevier, New York.

43. ZADEH L.A. (1978): Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets and Systems, 1, 1, 3-28.

STRESZCZENIE

W pracy przedstawiono ogólną koncepcję systemu dwustronnej komunikacji akustycznej w cybernetycznie pojmowanym układzie nadzoru, sterowania i dialogu «człowiek-maszyna» w języku naturalnym, za pomocą sygnału mowy. Koncepcję tę oparto na wieloletnich doświadczeniach i dotychczasowych osiągnięciach Zakładu Akustyki Cybernetycznej IPPT-PAN w dziedzinie analizy, automatycznego rozpoznawania i syntezy mowy, ze szczególnym uwzględnieniem fonetycznej i lingwistycznej specyfiki języka polskiego. Omówiono ogólne założenia i przyjęte metody realizacji urządzeń akustycznego wejścia i akustycznego wyjścia systemu dwustronnej komunikacji akustycznej «operator-robot» w oparciu o ograniczone, użytkowo ukierunkowane zbiory izolowanych wyrazów, traktowanych jako elementy fonetycznego kodu językowego.

Do automatycznego rozpoznawania izolowanych wyrazów po stronie akustycznego wejścia systemu opracowano algorytm wielostopniowej klasyfikacji dźwięków mowy, oparty na ich opisie artykulacyjnym, przy zastosowaniu do tego celu relacji rozmytych. Proces syntezy izolowanych wyrazów po stronie akustycznego wyjścia jest realizowany w syntezatorze formantowym, sterowanym sygnałami cyfrowymi wysyłanymi z komputera zgodnie ze zgromadzonymi w jego pamięci regułami syntezy elementarnych segmentów fonetycznych sygnału mowy o rozciągłości morfemów i diad.

Omówiono szczegółowo organizację i sposób realizacji doświadczalnego systemu dwustronnej komunikacji słownej człowiek - komputer, opracowywanego w ZAC i operującego słownikiem 60 wyrazów. Podano ogólną charakterystykę systemu w oparciu o jego wstępną weryfikację doświadczalną. W zakończeniu zwrócono uwagę na ukierunkowanie i zakres dalszych prac badawczych, teoretycznych i eksperymentalnych, jaki należy podjąć przy opracowywaniu użytkowej wersji docelowego systemu dwustronnej komunikacji słownej «operator - maszyna robocza (pojazd)», przystosowanego do pracy w rzeczywistych warunkach przemysłowych, różniących się od warunków laboratoryjnych m.in. niekorzystnym stosunkiem S/N sygnału do szumu (hałasu otoczenia) oraz ograniczeniem szerokości pasma częstotliwości sygnału, przesyłanego za pomocą elektronicznych urządzeń transmisyjnych i przetwórczych typu eksploatacyjnego.