ZESZYTY NAUKOWE POLITECHNIKI ŚLĄSKIEJ 2000
Seria: ELEKTRONIKA z. 12 Nr kol. 1492

Piotr KŁOSOWSKI
Politechnika Śląska, Instytut Elektroniki

USPRAWNIENIE PROCESU AUTOMATYCZNEGO ROZPOZNAWANIA MOWY OPARTEGO NA FONETYCE I FONOLOGII JĘZYKA POLSKIEGO

Streszczenie. Niniejszy artykuł ma na celu przedstawienie metody usprawnienia procesu rozpoznawania mowy opartego na artykulacyjnych cechach dystynktywnych fonemów języka polskiego. Metoda ta dedykowana jest dla systemów rozpoznających mowę języka polskiego. Dystynktywność, czyli zdolność do reprezentowania cech charakterystycznych obiektów z poszczególnych klas, uznano za najważniejszą cechę determinującą przydatność poszczególnych parametrów w procesie rozpoznawania mowy.

IMPROVED SPEECH RECOGNITION PROCESS BASED ON PHONETICS AND PHONOLOGY OF POLISH LANGUAGE

Summary. This article presents new improved method for speech recognition based on detection of distinctive acoustic parameters of phonemes in polish language. Distinctivity has been assumed as a most important selection of parameters, which have represented objects from recognized classes.

1. Wstęp

Mowa jest dźwiękową formą języka naturalnego, który w tradycyjnym ujęciu stanowi kod służący do porozumiewania się pomiędzy ludźmi. W dobie dynamicznie rozwijających się nauk technicznych oraz nowych dziedzin, takich jak multimedia, ta tradycyjna definicja uległa modyfikacji. Wyłaniają się nowe możliwości wykorzystania języka naturalnego, np. do komunikacji człowieka z maszyną oraz maszyny z człowiekiem. Powstają systemy porozumiewania się z komputerem za pomocą głosu oraz inteligentne przyrządy, które nie tylko komunikują się z człowiekiem za pomocą mowy, ale także rozumieją ludzką mowę.

Nikt nie jest w stanie przewidzieć, dokąd zawiedzie nas powszechna komputeryzacja, lecz na horyzoncie rysują się pewne interesujące perspektywy. Istnieją już obiecujące osiągnięcia w pracach nad maszynami rozpoznającymi głos i nad komunikacją z komputerem nie za pomocą klawiatury, lecz w ten najbardziej naturalny dla człowieka sposób - za pomocą mowy. Wprowadzanie informacji do maszyny za pośrednictwem głosu okazało się, w porównaniu z innymi sposobami, najbardziej efektywną i najwygodniejszą formą komunikacji człowieka z maszyną. Mowa jest najbardziej naturalną, elastyczną, wygodną i ekonomiczną formą komunikacji.

Po wielu latach badania nad rozpoznawaniem mowy zaczynają przekraczać próg zastosowań. Ostatnia dekada jest świadkiem znacznego postępu technologii rozpoznawania mowy. Systemy rozpoznawania mowy stają się dostępne w postaci oprogramowania oraz aplikacji sprzętowych. Jest już możliwe komunikowanie się za pomocą głosu z telefonem komórkowym oraz niektórymi urządzeniami powszechnego użytku. Rozpoznawanie mowy staje się obecnie najbardziej obiecującą technologią przyszłości. Sprzedaż produktów wykorzystujących tę technologię w roku 1997 wyniosła 500 milionów dolarów. W roku 2003 wskaźnik ten ma przekroczyć wartość 38 miliardów dolarów.

W Instytucie Elektroniki Politechniki Śląskiej w Gliwicach od kilku lat prowadzone są badania nad automatycznym rozpoznawaniem mowy polskiej [9,11,12,14,15] oraz wytwarzaniem mowy syntetycznej [13,16,17,18,19]. Początkowo prace były prowadzone z myślą o niewidomych i niedowidzących użytkownikach komputerów, czego dowodem był udział w realizacji grantu zamawianego, mającego na celu stworzenie stanowiska komputerowego przeznaczonego dla osób z upośledzeniem wzroku [10].

Skupiono się na próbie określenia, jakie cechy sygnału mowy odpowiedzialne są za przenoszenie informacji i jej subiektywny odbiór przez słuchacza. Posiadanie takiej wiedzy jest podstawą do konstruowania efektywnych systemów automatycznego rozpoznawania mowy polskiej. Spodziewane korzyści płynące z podjęcia tej tematyki nie ograniczają się tylko do określenia podstaw efektywnych systemów automatycznego rozpoznawania mowy polskiej, ale także pozwolą na generowanie syntetycznej mowy wysokiej jakości.

Mimo ogromnego zapotrzebowania, nie istnieją systemy dedykowane dla języka polskiego rozpoznające mowę ciągłą, tak więc porównanie proponowanych rozwiązań z innymi systemami jest utrudnione. Próby dostosowania aplikacji rozpoznających mowę obcojęzyczną do rozpoznawania mowy polskiej nie przynoszą zadowalających efektów. Zatem każde skuteczne rozwiązanie lub usprawnienie procesu automatycznego rozpoznawania mowy polskiej jest niezwykle cenne.

Przedmiotem prowadzonych badań jest usprawnienie procesu automatycznego rozpoznawania mowy języka polskiego. Podczas prowadzenia badań zaistniała wyraźna potrzeba stworzenia pomostu między wiedzą dotyczącą cyfrowego przetwarzania sygnału mowy a wiedzą lingwistyczną dotyczącą języka polskiego oraz mowy polskiej. Podjęto próbę wykorzystania dorobku dyscyplin językoznawczych, takich jak fonetyka i fonologia języka polskiego, do usprawnienia procesu automatycznego rozpoznawania mowy polskiej.

Istota procesu automatycznego rozpoznawania mowy polega na tym, aby je oprzeć na cechach pierwszorzędnych z punktu widzenia tego procesu, czyli tych, które są nośnikami przekazywanej za pomocą mowy informacji. Wykorzystanie cech drugo- lub trzeciorzędnych obniża wydajność rozpoznawania i sprawia, że jest ona uzależniona w wysokim stopniu od mówcy, a nawet jego stanu emocjonalnego.

2. Automatyczne rozpoznawanie mowy polskiej na poziomie fonemów

Z punktu widzenia fizyki istotą mowy są wywołane drganiami akustycznymi zmiany ciśnienia, rozchodzące się w środowisku sprężystym ruchem falowym. W takim ujęciu mowa nie różni się niczym szczególnym od innych fal akustycznych, wytwarzanych w przyrodzie lub emitowanych za pomocą różnego rodzaju urządzeń skonstruowanych przez człowieka. Fale takie, nazywane też głosowymi albo niezbyt ściśle dźwiękowymi, można wyczerpująco opisać opierając się na analizie niewielkiej liczby ich fizycznych właściwości.

Mowa to ukształtowany w toku biologicznej i społecznej ewolucji człowieka system elementów głosowych, spełniających w określonym czasie i w określonej wspólnocie ludzkiej funkcję komunikacyjną.

Istota automatycznego rozpoznawania mowy polega najogólniej na tym, że określony układ techniczny odbiera sygnał mowy i mając do wyboru pewną skończoną liczbę elementów językowych klasyfikuje każdy odebrany element, jako jeden z tego skończonego zbioru. Automatycznemu rozpoznawaniu mogą podlegać elementy językowe różnych rzędów, najczęściej jednak rozpoznaje się **fonemy**, najmniejsze jednostki dźwiękowe, z których jest zbudowane słowo, lub bardziej złożone struktury, np. wyrazy.

Rozpoznawanie mowy jest trudnym problemem. W dużym stopniu wynika to z istnienia wielu źródeł zmienności sygnału mowy, są nimi:

- zmienność fonetyczna realizacji fonemów, z których słowo jest zbudowane, zależna od kontekstu, w którym występują, przy czym na granicach słów zmiany kontekstowe są zasadnicze,
- zmienność akustyczna, która może wynikać ze zmian w otoczeniu mówiącego, jak również z umiejscowienia i charakterystyk przetwornika mowy,
- wewnętrzna zmienność mogąca wynikać ze zmian stanu fizycznego i emocjonalnego mówcy, tempa mówienia oraz jakości głosu,
- różnice socjolingwistyczne w dialekcie i rozmiarze traktu głosowego, które mogą być poważnym utrudnieniem przy zmianie mówcy.

Fonem natomiast to byt abstrakcyjny, będący zbiorem cech dystynktywnych, co jest równoznaczne z definicją, według której fonem jest najkrótszym odcinkiem ciągu mowy, zdolnym do różnicowania znaczeń wyrazów [5]. Fonem uznaje się za najmniejszy, dający się liniowo wydzielić funkcjonalny segment formy językowej [22]. Wariantem fonemu jest **allofon.** Reprezentuje on dźwiękową realizację fonemu [8,7]. Fonem jako podstawowa jednostka znaczeniowa ma tę właściwość, że jeśli jeden fonem zastąpimy innym, to znaczenie słowa może ulec zmianie. Ponieważ fonemy są składnikami mowy, to dowolne wypowiedzi mogą być wyrażane w postaci łańcucha fonemów.

Zespół fonemów może być traktowany jako alfabet bazowy dla danego języka. Posługując się alfabetem fonetycznym, jesteśmy w stanie określić zbiór fonemów języka polskiego, jako alfabet bazowy w rozpoznawaniu mowy polskiej. Zbiór taki zamieszczono w tabeli 1.

Ciąg fonemów tworzy bardziej złożone jednostki językowe, takie jak sylaby, wyrazy i zdania oraz całe wypowiedzi. Istniejące związki między poszczególnymi jednostkami można rozpatrywać na wielu płaszczyznach: fizycznej, fonetycznej i lingwistycznej. Ma to znaczenie nie tylko w procesie modelowania percepcji i rozumienia mowy przez człowieka, ale również w tworzeniu systemów automatycznego rozpoznawania mowy. Rozpoznawanie elementarnych jednostek językowych, jakimi są fonemy języka polskiego, pozwala automatycznie rozpoznawać dowolne wypowiedzi w tym języku. Istnieje zatem możliwość usprawnienia procesu rozpoznawania mowy polskiej przez oparcie go na rozpoznawaniu poszczególnych fonemów języka.

Tabela 1

Zbiór fonemów języka polskiego [8]

lp.	Fonem	Przykłady występowania	lp.	Fonem	Przykłady występowania
1	i	wici	21	ś	siano
2	y	syty	22	ź	ziarno
3	e	serce	23	χ	higiena
4	a	baba	24	p	praca
5	o	oko	25	b	baba
6	u	buk	26	t	trawa
7	ɨ	jajo	27	d	dudek
8	ʉ	łysy	28	k	kot
9	r	rok	29	g	mogą
10	l	lato	30	ḱ	kino
11	m	mama	31	ǵ	magiczny
12	n	noc	32	c	cacko
13	ń	koń	33	ʒ	nadzy
14	ŋ	ręka	34	č	czarny
15	f	fala	35	ǯ	drożdże
16	v	wada	36	ć	ciasto
17	s	sok	37	ʒ́	dziedzic
18	z	koza			
19	š	szyszka			
20	ž	każdy			

Głównym problemem w automatycznym rozpoznawaniu mowy jest wybór adekwatnych parametrów, za pomocą których można przypisać fragmentom mowy przynależność do określonych klas. Zespoły parametrów służące do automatycznego rozpoznawania mowy powinny być [2,31]:

- skuteczne w sensie reprezentacji cech charakterystycznych dla danej klasy,
- łatwe do pomiaru,
- stabilne w czasie,
- odporne na zakłócenia, czyli możliwie mało zmieniające się wraz ze zmianami warunków pomiaru.

Najważniejszą cechą determinującą przydatność poszczególnych parametrów (lub ich zespołów) powinna być ich dystynktywność, czyli zdolność do reprezentowania cech charakterystycznych obiektów z poszczególnych klas. Rozpoznawanie tych cech, ma kluczowe znaczenie w procesie automatycznego rozpoznawania mowy polskiej.

W czasie prowadzenia badań własnych powstał pomysł usprawnienia procesu rozpoznawania mowy polskiej poprzez taki dobór cech dystynktywnych, aby statystyczna liczba cech niezbędnych do rozpoznania pojedynczego fonemu była mniejsza od dotychczas uzyskanych.

Zaproponowano więc nowy zestaw cech dystynktywnych oparty na elementarnych cechach artykulacyjnych sygnału mowy.

3. Propozycja usprawnienia procesu automatycznego rozpoznawania mowy polskiej oparta na nowym zestawie artykulacyjnych cech dystynktywnych

Propozycja usprawnienia procesu automatycznego rozpoznawania mowy sprowadza się do wykorzystania elementarnych artykulacyjnych cech dystynktywnych do rozpoznawania fonemów [11]. Rozpoznając fonemy mowy polskiej można rozpoznawać dowolną wypowiedź języka polskiego. Rozpoznawanie tych elementarnych składowych języka ma fundamentalne znaczenie w procesie rozpoznawania mowy. Dobór cech dystynktywnych jest nieco inny niż we wcześniej znanych rozwiązaniach [8].

Zaproponowany nowy zestaw artykulacyjnych cech dystynktywnych wykorzystywany jest przez fonologów do rozwiązywania zagadnień językowych (fonologicznych), a między innymi do klasyfikacji i badania elementarnych jednostek językowych [22].

Po wstępnych badaniach mających na celu ustalenie przydatności poszczególnych cech artykulacyjnych do rozpoznawania mowy, autor zdecydował się na wykorzystanie nowego zestawu 19 artykulacyjnych cech dystynktywnych do rozpoznawania fonemów języka polskiego. Analizowanych cech było bardzo wiele, ale tylko 19 z nich pełniło funkcję dystynktywną, a więc brało czynny udział w rozpoznawaniu fonemów. Wybrane cechy dystynktywne miały charakter elementarny, każda z cech odzwierciedlała konkretne zjawisko artykulacyjne. Dzięki temu cechy te stały się łatwiejsze do wyekstrahowania z sygnału mowy.

Liczba zaproponowanych cech dystynktywnych (dobranych według zaproponowanej w pracy metody) jest większa niż w rozwiązaniach znanych z literatury [8]. Niewątpliwą zaletą zwiększenia zbioru cech jest zmniejszenie się liczby cech niezbędnych do rozpoznania pojedynczego fonemu. Do rozpoznania większości fonemów języka polskiego potrzeba określenia tylko 3 cech, a w niektórych przypadkach tylko 2, a dla fonemów [r] i [l] nawet 1 cechy.

Elementarny charakter zaproponowanych cech dystynktywnych oraz bardzo mała liczba cech niezbędnych do rozpoznawania pojedynczego fonemu ma fundamentalne znaczenie dla usprawnienia automatycznego rozpoznawania mowy polskiej [11].

Zestaw 19 elementarnych cech artykulacyjnych można podzielić na 4 grupy:

- **grupa SA** - cechy określające sposób artykulacji (zwartość, szczelinowość, zwarto-szczelinowość, otwartość, nosowość, drżącość, boczność, półotwartość),
- **grupa MA** - cechy określające miejsce artykulacji (dwuwargowość, wargowo-zębowość, dziąslowość, przedniojęzykowość, środkowojęzykowość, tylno-językowość),
- **grupa CD** - cechy dodatkowe (język wysoko, język nisko),
- **grupa SW** - cechy określające stan wiązadeł głosowych (dźwięczność, bezdźwięczność).

Na uwagę zasługuje także fakt, iż podczas artykulacji fonemów w każdej z czterech grup cech może wystąpić jednocześnie tylko jedna z cech. Ma to wielkie znaczenie w usprawnieniu procesu automatycznego rozpoznawania mowy. W większości przypadków gwałtowne przechodzenie z jednej cechy do drugiej jest także ograniczone bezwładnością mechaniczną narządów artykulacji.

Tabela 2 przedstawia, które z cech dystynktywnych odpowiedzialne są za rozpoznawanie poszczególnych fonemów. Kolumny tabeli odpowiadają cechom określającym sposób artykulacji. Wiersze natomiast odpowiadają cechom określającym miejsce artykulacji. W miejscu przecięcia się danej kolumny i wiersza, odpowiadającym konkretnej parze cech, zapisano fonemy, dla których dana para cech odznacza się dystynktywnością. Kolor fonemu informuje, która z pozostałych cech odpowiedzialna jest za rozpoznawanie.

Większość fonemów opisana jest za pomocą dwóch cech dystynktywnych określających sposób oraz miejsce artykulacji. Wyjątkiem są fonemy [r] i [l], do rozpoznawania których wystarczy tylko określenie jednej cechy dotyczącej sposobu artykulacji. Jednak aby rozróżnić pozostałe fonemy, niezbędna jest jeszcze cecha określająca stan wiązadeł głosowych lub tzw. cecha dodatkowa wykorzystywana do rozpoznawania samogłosek.

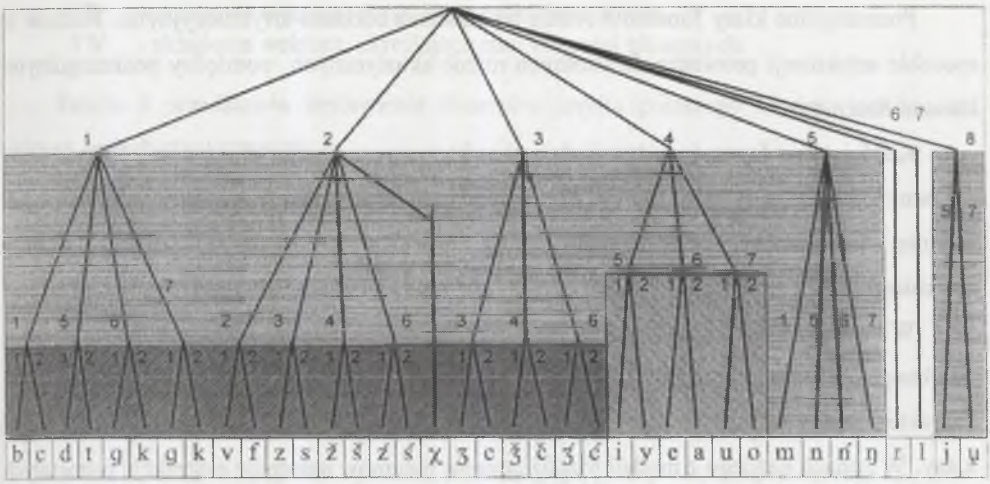
Tabela 2

Zbiór fonemów i odpowiadających im cech dystynktywnych odpowiedzialnych za ich rozpoznawanie

	SA1 zwartość	SA2 szczelin.	SA3 zw-szcz.	SA4 otwartość	SA5 nosowość	SA6 drżącość	SA7 boczność	SA8 półotw.
MA1 dwuwarg.	[b][1] [p][2]				[m]			
MA2 warg-zęb.		[v][1] [f][2]						
MA3 zębowość		[z][1] [s][2]	[ʒ][1] [c][2]					
MA4 dziąsłow.		[ž][1] [š][2]	[ǯ][1] [č][2]			[r][5]	[l][5]	
MA5 przed.-jęz.	[d][1] [t][2]			[i][3] [y][4]	[n]			[j]
MA6 środ.-jęz.	[ɡ́][1] [ḱ][2]	[ź][1] [ś][2]	[ʑ][1] [ć][2]	[e][3] [a][4]	[ń]			
MA7 tylno-jęz.	[g][1] [k][2]	[χ]		[u][3] [o][4]	[ŋ]			[u̯]

Odnośniki oznaczają kolejno:
1) fonemy dźwięczne (SW1),
2) fonemy bezdźwięczne (SW2),
3) fonemy artykuowane przy wysokim położeniu języka (CD1),
4) fonemy artykuowane przy niskim położeniu języka (CD2),
5) fonemy rozpoznawane za pomocą tylko jednej cechy dystynktywnej (SA).

Na rysunku 1 przedstawiono diagram cech dystynktywnych, ilustrujący proces rozpoznawania mowy polskiej na ich podstawie. Gałęzie diagramu pokazują, na podstawie jakich cech dystynktywnych rozpoznawane są kolejne fonemy języka polskiego.



Rys.1. Diagram rozpoznawania fonemów języka polskiego opartego na zaproponowanym zestawie artykulacyjnych cech dystynktywnych

Fig. 1. Recognition diagram of polish phonems based on suggested distinctive acoustic parameters of phonems

Odrębnymi odcieniami szarości zaznaczono części diagramu odpowiadające poszczególnym grupom cech dystynktywnych, a cyframi konkretne cechy z danej grupy, biorące udział w procesie rozpoznawania:

1 – SA1 zwartość, 2 – SA2 szczelinowość, 3 – SA3 zwarto-szczelinowość, 4 – SA4 otwartość, 5 – SA5 nosowość, 6 – SA6 drżącość, 7 – SA7 boczność, 8 – SA8 półotwartość,	1 – MA1 dwuwargowość, 2 – MA2 wargowo-zębowość, 3 – MA3 zębowość, 4 – MA4 dziąsłowość, 5 – MA5 przedniojęzykowość, 6 – MA6 środkowojęzykowość, 7 – MA7 tylnojęzykowość,
1 – SW1 dźwięczność, 2 – SW2 bezdźwięczność.	1 – CD1 język wysoko, 2 – CD2 język nisko.

Cechy dystynktywne określające sposób artykulacji pozwalają podzielić zbiór fonemów języka polskiego na poszczególne klasy fonemów, odpowiadające różnym sposobom ich artykulacji, tj.:

- fonemy zwarte,
- fonemy szczelinowe,
- fonemy zwartoszczelinowe,
- fonemy otwarte,
- fonemy nosowe,
- fonem drżący,

Poszczególne klasy fonemów różnią się znacznie cechami artykulacyjnymi. Różnice w sposobie artykulacji prowadzą do istotnych różnic akustycznych pomiędzy poszczególnymi klasami fonemów.

Każdej z klas fonemów odpowiadają charakterystyczne cechy akustyczne, opisane parametrami fizycznymi sygnału mowy. Aby ustalić sposób artykulacji fonemów, niezbędne jest badanie akustycznych cech sygnału mowy charakterystycznych dla danego sposobu artykulacji.

W konsekwencji każda klasa fonemów języka polskiego rozpoznawana jest za pomocą odrębnego modelu rozpoznawania. Każdy z modeli umożliwia rozpoznawanie fonemów charakteryzujących się tym samym sposobem artykulacji, a więc należących do tej samej klasy. W ramach każdego z modeli rozpoznawania fonemów następuje ekstrakcja pozostałych artykulacyjnych cech dystynktywnych niezbędnych do rozpoznawania fonemów danej klasy. Zestaw parametrów fizycznych sygnału mowy, używany do ekstrakcji cech dystynktywnych, jest różny dla każdego z modeli. Spowodowane jest to faktem, iż fonemy należące do poszczególnych klas istotnie różnią się akustycznie.

Do parametrów fizycznych sygnału mowy wykorzystywanych w ekstrakcji cech dystynktywnych należą m.in.:

- częstotliwości formantowe, tzn. częstotliwości, w których przypadają maksima obwiedni widma zdjętego w małym przedziale czasowym. Najczęściej wyróżnia się pierwsze 4 formanty,
- względne amplitudy poszczególnych formantów,
- częstotliwości antyformantów, czyli minimów obwiedni widma, które występują na skutek pobudzenia zamkniętych bocznikujących układów rezonansowych kanału głosowego,
- częstotliwość podstawowa tonu krtaniowego F_0,
- przebieg czasowy obwiedni amplitudy lub energii,
- przebiegi czasowe wyekstrahowanych w dziedzinie częstotliwości parametrów takich jak: częstotliwość podstawowa tonu krtaniowego, trajektorie formantów itp.,
- spektrogramy oraz sonogramy sygnału mowy.

Wektor X cech dystynktywnych pozwalający rozpoznawać fonemy języka polskiego będzie miał postać:

$$X = [SA, MA, CD, SW] \tag{1}$$

gdzie: SA - składowa wektora, określająca sposób artykulacji,
MA - składowa wektora, określająca miejsce artykulacji,

CD - składowa wektora, określająca cechę dodatkową,
SW - składowa wektora, określająca stan wiązadeł głosowych.

Tabela 3 przedstawia zestawienie fonemów języka polskiego i odpowiadające im wektory cech dystynktywnych.

Tabela 3

Wektory cech dystynktywnych fonemów języka polskiego

Lp.	Fonem	Wektor cech dystynktywnych [SA,MA,CD,SW]	Liczba cech niezbędnych do rozpoznawania fonemu
1	b	[1,1,0,1]	3
2	p	[1,1,0,2]	3
3	d	[1,5,0,1]	3
4	t	[1,5,0,2]	3
5	ǵ	[1,6,0,1]	3
6	ḱ	[1,6,0,2]	3
7	g	[1,7,0,1]	3
8	k	[1,7,0,2]	3
9	v	[2,2,0,1]	3
10	f	[2,2,0,2]	3
11	z	[2,3,0,1]	3
12	s	[2,3,0,2]	3
13	ž	[2,4,0,1]	3
14	š	[2,4,0,2]	3
15	ź	[2,6,0,1]	3
16	ś	[2,6,0,2]	3
17	χ	[2,7,0,0]	2
18	ʒ	[3,3,0,1]	3
19	c	[3,3,0,2]	3
20	ǯ	[3,4,0,1]	3
21	č	[3,4,0,2]	3
22	ʒ́	[3,6,0,1]	3
23	ć	[3,6,0,2]	3
24	i	[4,5,1,0]	3
25	y	[4,5,2,0]	3
26	e	[4,6,1,0]	3
27	a	[4,6,2,0]	3
28	u	[4,7,1,0]	3
29	o	[4,7,2,0]	3
30	m	[5,1,0,0]	2
31	n	[5,5,0,0]	2
32	ń	[5,6,0,0]	2
33	ŋ	[5,7,0,0]	2
34	r	[6,0,0,0]	1
35	l	[7,0,0,0]	1
36	j	[8,5,0,0]	2
37	ɥ	[8,7,0,0]	2

Poszczególne pola (składowe) wektora cech odpowiadają cechom z kolejnych grup cech, tzn. grupy cech określających sposób artykulacji (SA), miejsce artykulacji (MA), cechę dodatkową (CD) oraz stan wiązadeł głosowych (SW). Wartość 0 w danym polu wektora cech dystynktywnych oznacza, że dana cecha nie pełni funkcji dystynktywnej dla danego fonemu, nie bierze więc udziału w procesie jego rozpoznawania. Zbędne więc jest jej określanie. Informacja o tym przekazywana jest do warstwy niższej (artykulacyjnej) w celu usprawnienia procesu rozpoznawania.

Uwzględniając rozkład prawdopodobieństwa występowania fonemów w języku polskim [29], jesteśmy w stanie wyliczyć średnią liczbę cech dystynktywnych *Ns* niezbędną do rozpoznawania pojedynczego fonemu zaproponowaną metodą. Jest ona równa [1]:

$$Ns = \sum_{k=1}^{M} p_k \cdot N_k = \sum_{k=1}^{37} p_k \cdot N_k = 2.71 \qquad (2)$$

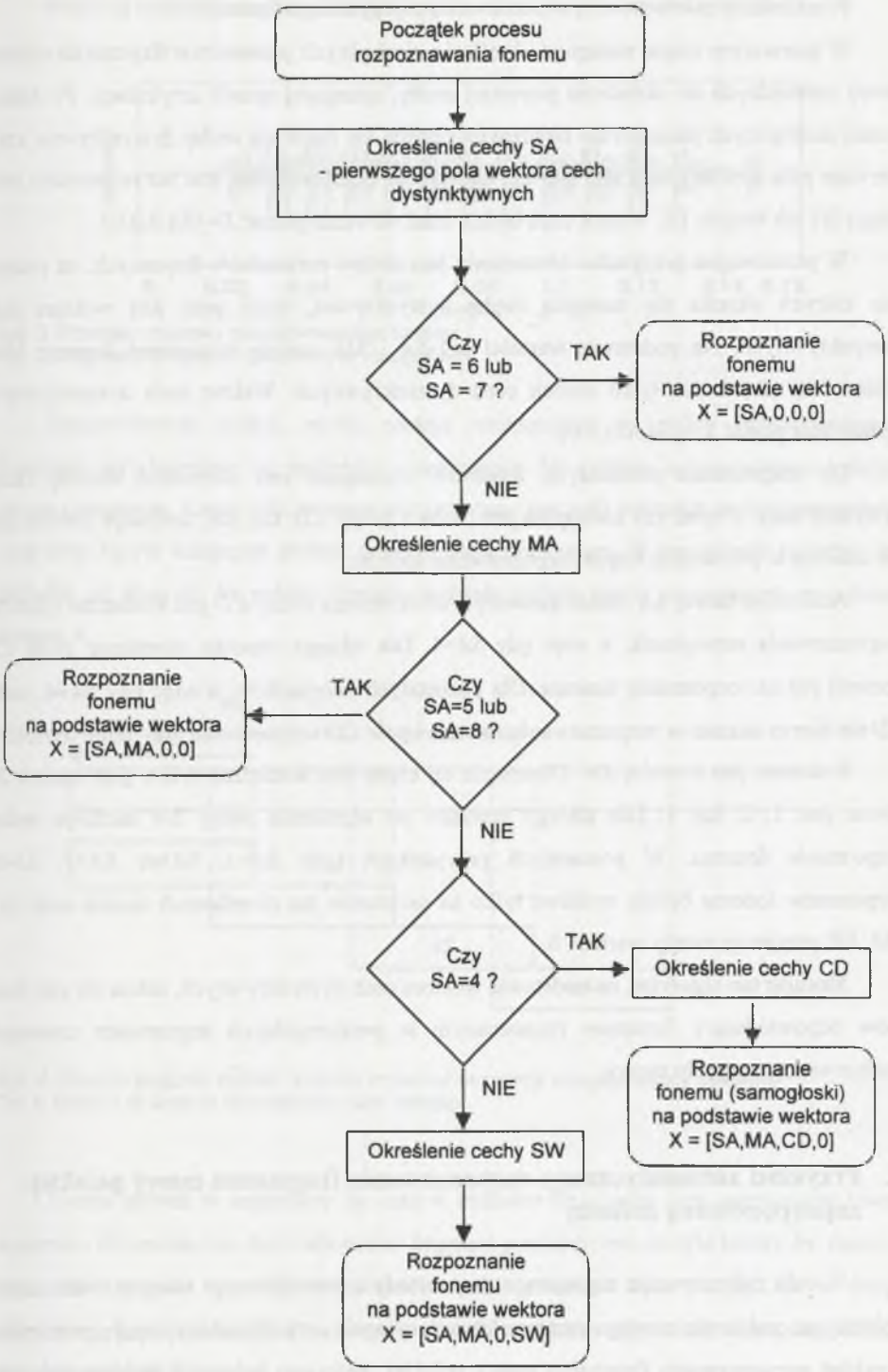
gdzie:

Ns - średnia liczba cech określająca pojedynczy fonem,
p_k - prawdopodobieństwo występowania k-tego fonemu,
N_k - liczba cech określająca k-ty fonem,
M - całkowita liczba fonemów języka polskiego.

Uzyskany wynik zasługuje na uwagę. Niewielka liczba cech niezbędnych do rozpoznania statystycznego fonemu jest ważnym czynnikiem usprawniającym proces automatycznego rozpoznawania mowy polskiej. Ta stosunkowo niska wartość *Ns* jest znacznie mniejsza, niż w przypadku rozpoznawania mowy na podstawie zestawu cech dystynktywnych zaproponowanych przez Jassema, dla którego *Ns*=5.1 [8].

4. Zaproponowany algorytm rozpoznawania pojedynczego fonemu

Algorytm rozpoznawania pojedynczego fonemu można prześledzić na zaprezentowanym na rysunku 2 schemacie blokowym.



Rys. 2. Schemat blokowy ilustrujący algorytm rozpoznawania pojedynczego fonemu
Fig. 2. Algorithm of recognition single pfonem

Prześledźmy zatem proces rozpoznawania pojedynczego fonemu:

W pierwszym etapie następuje określenie niezbędnych parametrów fizycznych sygnału mowy niezbędnych do określenia pierwszej cechy, opisującej sposób artykulacji. Po dostarczeniu niezbędnych parametrów fizycznych określa się pierwszą cechę dystynktywną, czyli pierwsze pole wektora cech *SA*. Gdy *SA=6* lub *SA=7*, dany fonem jest już rozpoznany jako drżący [r] lub boczny [l]. Wektor cech będzie miał wówczas postać *X=[SA,0,0,0]*.

W przeciwnym przypadku dostarczany jest zestaw parametrów fizycznych, na podstawie których określa się następną cechę dystynktywną, czyli pole *MA* wektora cech dystynktywnych. Na podstawie wartości pól *SA* i *MA* zostają rozpoznane fonemy, które opisane są za pomocą tylko dwóch cech dystynktywnych. Wektor cech dystynktywnych będzie miał postać *X=[SA,MA,0,0]*.

Do rozpoznania pozostałych fonemów wymagane jest określenie trzeciej cechy dystynktywnej. O tym, czy niezbędna jest cecha z grupy *CD*, czy *SW*, decyduje wartość pola *SA* ustalona w pierwszym etapie rozpoznawania fonemu.

Analizując tabelę 3.1 można zauważyć, że określenie cechy *CD* jest konieczne tylko dla rozpoznawania samogłosek, a więc gdy *SA*=4. Dla takiego warunku określenie cechy *CD* pozwoli już na rozpoznanie fonemu. Dla pozostałych przypadków, a więc gdy $SA \neq 4$, cecha *CD* nie bierze udziału w rozpoznawaniu fonemu i pole *CD* wektora cech przyjmuje wartość 0.

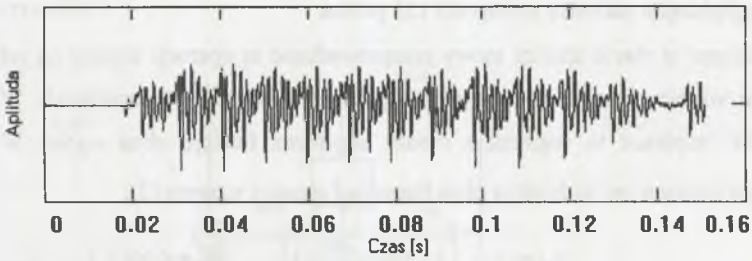
Podobnie jest z cechą *SW*. Określenie tej cechy jest konieczne tylko, gdy wartość *SA* równe jest 1, 2 lub 3. Dla takiego warunku po określeniu cechy *SW* możliwe będzie rozpoznanie fonemu. W pozostałych przypadkach (gdy *SA=5*, *SA=6*, *SA=7*, *SA=8*) rozpoznanie fonemu będzie możliwe tylko na podstawie już określonych dwóch cech *SA* i *MA*, *SW* przyjmuje wtedy wartość 0.

Stosując ten algorytm, na podstawie wektora cech dystynktywnych, ustala się ciąg fonemów odpowiadający fonemom rozpoznanym w poszczególnych segmentach czasowych analizowanego sygnału mowy.

5. Przykład automatycznego rozpoznawania fragmentu mowy polskiej zaproponowaną metodą

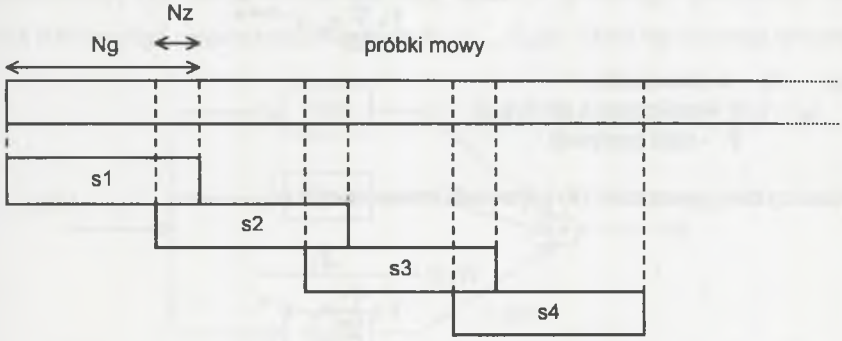
W celu zobrazowania zaproponowanej metody automatycznego rozpoznawania mowy polskiej na podstawie nowego zestawu artykulacyjnych cech dystynktywnych, prześledźmy przykład rozpoznawania fragmentu mowy polskiej, będącego jednym z izolowanych fonemów.

Przebieg czasowy rozpoznawanego fragmentu mowy przedstawia rysunek 3.



Rys. 3. Przebieg czasowy rozpoznawanego fonemu
Fig. 3. Fragment of speech (phonem) to recognition

Zarejestrowane próbki mowy zostają umieszczone w tablicy jednowymiarowej. Następnie wyodrębniane są segmenty zawierające *Ng* próbek, odpowiadające kolejnym oknom czasowym. Każdy taki segment otrzymywany jest z *Nz* ostatnich próbek poprzedniego okna oraz *Ng-Nz* kolejnych próbek nowego okna czasowego. W ten sposób powstaje tzw. zakładka od długości *Nz* próbek. Sposób podziału próbek mowy na segmenty przedstawia rysunek 4.



Rys. 4. Sposób podziału próbek sygnału mowy na segmenty uwzględniający zakładkę
Fig. 4. Method of speech segmentation with overlap

Liczba próbek w segmencie *Ng* oraz w zakładce *Nz*, a więc tym samym czas trwania segmentu, dobierana jest doświadczalnie. Segment powinien być na tyle krótki, by nawet dla segmentów sygnału o szybkich zmianach struktury można było przyjąć, że w obrębie segmentu sygnał jest lokalnie stacjonarny. Zwykle segment odpowiada oknu czasowemu o długości od 10 do 30 ms.

Liczba próbek zarejestrowanego sygnału mowy wynosi 1979, częstotliwość próbkowania 11 025 Hz. Segmentacja podzieliła sygnał na 14 segmentów po 256 próbek, o czasie trwania 23 ms, uwzględniając zakładkę o długości 128 próbek.

W następnym etapie analizy mowy przeprowadzane są operacje mające na celu estymację chwilowego widma mocy, a w rezultacie określenie trajektorii formantowych. W pierwszym kroku próbki zapisane w segmencie mnoży się przez funkcję okna $w_h(n)$. W przypadku omawianego systemu jest to funkcja okna Hanninga opisana wzorem (3):

$$w_h(n) = \frac{1}{2} \cdot \left[1 - \cos\left(\frac{2\pi n}{Ng - 1}\right) \right] \qquad 0 \le n \le Ng - 1 \tag{3}$$

gdzie Ng oznacza liczbę próbek sygnału mowy w danym segmencie.

W każdym segmencie wyznaczane są współczynniki predykcji. Wyznaczenie współczynników predykcji pozwala na wykorzystanie ich do określenia chwilowego widma mocy. Jeśli przyjmiemy, że wielomian predykcyjny reprezentuje mianownik funkcji transmitancji kanału głosowego, to charakterystykę częstotliwościową tego kanału dla określonego segmentu sygnału mowy można przedstawić jako:

$$H(e^{j2\pi fn}) = \frac{A}{1 + \sum_{p=1}^{P} a_p \cdot e^{-j2\pi fn \cdot p}} \tag{4}$$

gdzie: A - wzmocnienie,
a_p - p-ty współczynnik predykcji,
P - rząd predykcji.

Na płaszczyźnie z zależności (4) odpowiada transmitancja:

$$H(z) = \frac{A}{1 + \sum_{p=1}^{P} a_p \cdot z^{-p}} \tag{5}$$

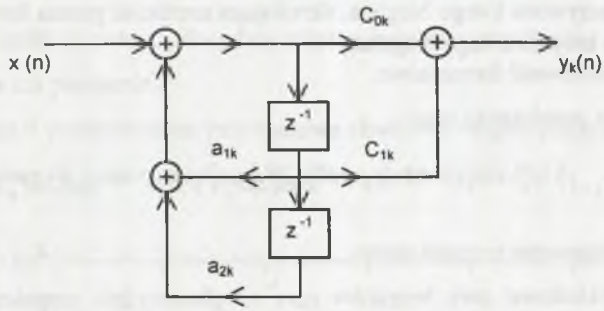
Transmitancję $H(z)$ z równania (5), będącą transmitancją filtru cyfrowego o nieskończonej odpowiedzi impulsowej IIR (*ang. Infinite Impulse Response*), można zapisać jako sumę M transmitancji składowych drugiego stopnia $H_k(z)$ w postaci:

$$H(z) = \sum_{k=1}^{M} H_k(z) \tag{6}$$

gdzie k-ta składowa $H_k(z)$ ma postać:

$$H_k(z) = \frac{c_{0k} + c_{1k} z^{-1}}{1 - a_{1k} z^{-1} - a_{2k} z^{-2}} \tag{7}$$

Powyższa składowa odpowiada transmitancji filtru cyfrowego IIR drugiego rzędu. Schemat składowego k-tego filtru cyfrowego o transmitancji $H_k(z)$ opisanej wzorem (7) przedstawia rysunek 5.

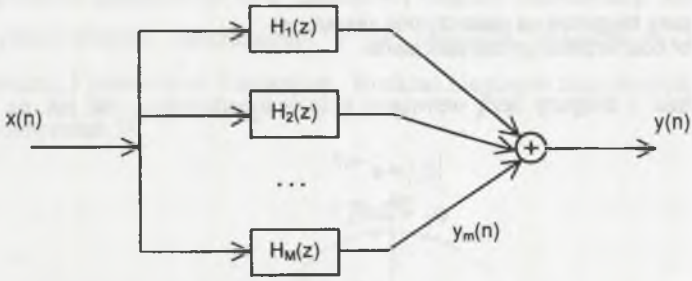


Rys. 5. Schemat pojedynczego k-tego filtru składowego o transmitancji Hk(z)
Fig. 5. Diagram of single k-th filter Hk(z)

Transmitancja zapisana wzorem (6) przyjmie więc postać:

$$H(z) = \sum_{k=1}^{M} \frac{c_{0k} + c_{1k} z^{-1}}{1 - a_{1k} z^{-1} - a_{2k} z^{-2}} \tag{8}$$

Transmitancję (8) można zrealizować jako układ M równolegle połączonych filtrów cyfrowych IIR drugiego rzędu o transmitancjach $H_1(z),\ldots,H_M(z)$. Układ taki ilustruje rysunek 6.



Rys. 6. Układ równolegle połączonych filtrów składowych
Fig. 6. Curcuit of parallely connected filter component

Widmo odpowiedzi impulsowej takiego układu filtrów będzie przybliżeniem widma sygnału mowy w danym segmencie (oknie czasowym). W ten sposób można łatwo określić częstotliwości formantowe, bowiem każdy z równolegle połączonych filtrów formuje pojedynczy formant. Częstotliwościom formantowym odpowiadają wartości pierwiastków wielomianu predykcyjnego, czyli bieguny transmitancji (5).

Najczęściej dwa sprzężone bieguny odpowiadają pojedynczemu formatowi:

$$s_k, s_k^* = -\sigma_k \pm j2\pi F_k \tag{9}$$

gdzie:

s_k, s_k^* - k-ta para biegunów sprzężonych transmitancji,
σ_k - część rzeczywista k-tego bieguna, określająca szerokość pasma formantu,
$j2\pi F_k$ - część urojona k-tego bieguna,
F_k - k-ta częstotliwość formantowa.

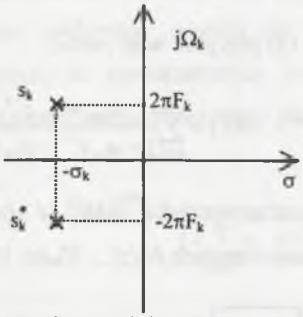
Bieguny można także przedstawić jako:

$$z_k, z_k^* = e^{-\sigma_k T} e^{\pm j2\pi F_k T} = e^{-\sigma_k T}\cos(2\pi F_k T) \pm j e^{-\sigma_k T}\sin(2\pi F_k T) \tag{10}$$

gdzie:

T - okres próbkowania sygnału mowy.

Położenie przykładowej pary biegunów s_k, s_k^* na płaszczyźnie zespolonej przedstawia rysunek 7.

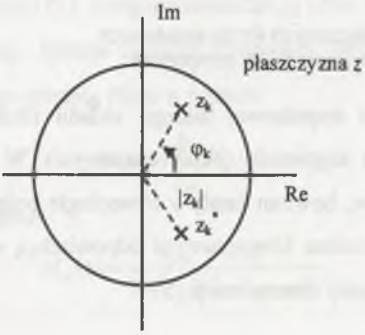


Rys. 7. Położenie pary biegunów na płaszczyźnie zespolonej
Fig. 7. Placement of couple poles on complex plane

Na płaszczyźnie z bieguny leżą wewnątrz koła jednostkowego, tak jak na rysunku 8, gdzie:

$$|z_k| = e^{-\sigma_k T} \tag{11}$$

$$\varphi_k = 2\pi F_k T \tag{12}$$



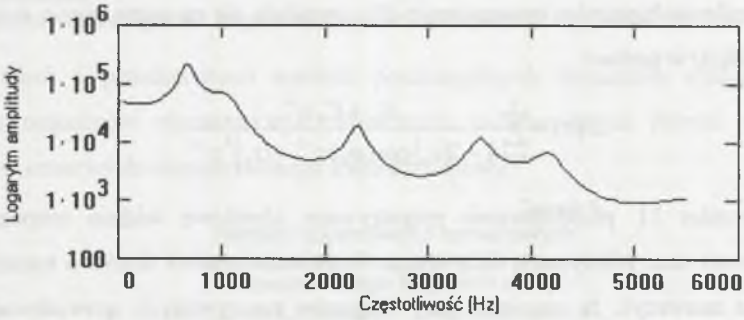
Rys.8. Położenie pary biegunów na płaszczyźnie z
Fig. 8. Placement of couple poles on 'z' plane

Transmitancja opisana równaniem (8) będzie miała teraz następującą postać:

$$H(z)=\sum_{k=1}^{M}\frac{B_k+C_k z^{-1}}{1-2|z_k|\cos(\varphi_k)z^{-1}+|z_k|^2 z^{-2}} \qquad (13)$$

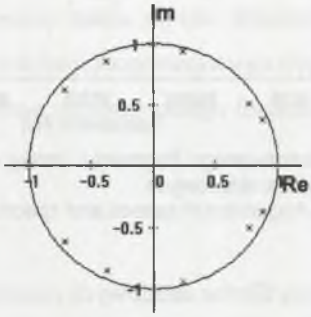
Współczynniki B_k oraz C_k decydują między innymi o amplitudzie formantów, lecz nie mają wpływu na ich położenie.

Na rysunku 9 przedstawiono przykładowe chwilowe widmo rozpoznawanego fragmentu mowy otrzymane za pomocą analizy LPC dla rzędu predykcji P=12.



Rys. 9. Widmo chwilowe rozpoznawanego fragmentu mowy
Fig. 9. Spetrum of recognized fragment of speech

Po określeniu parametrów LPC znaleziono bieguny transmitancji *H(z)*. Po usunięciu biegunów rzeczywistych, otrzymujemy 5 par biegunów zespolonych odpowiadających częstotliwościom 5 pierwszych formantów. Rozkład biegunów zespolonych na płaszczyźnie z przedstawia rysunek 10.



Rys. 10. Rozkład biegunów transmitancji H(z) na płaszczyźnie z po usunięciu biegunów rzeczywistych
Fig. 10. Placement of poles H(z) on 'z' plane after remove of real poles

Wartości wyliczonych na podstawie rozkładu biegunów przedstawiono w tabeli 4.

Tabela 4

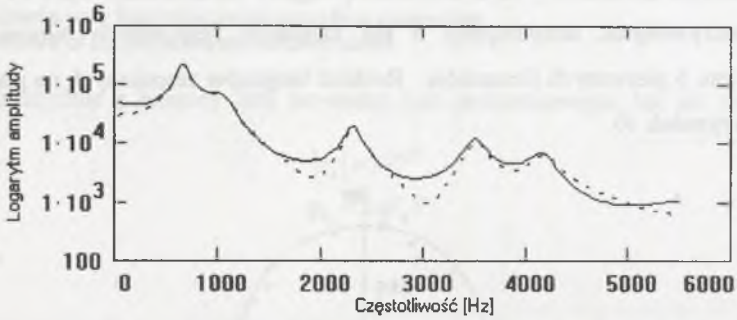
Wyliczone sprzężone pary biegunów zespolonych

Pary biegunów sprzężonych	Wartości
s1,s1*	0.882 ± 0.378i
s2,s2*	0.771 ± 0.507i
s3,s3*	0.234 ± 0.939i
s4,s4*	-0.379 ± 0.857i
s5,s5*	-0.713 ± 0.618i

Po określeniu biegunów transmitancję *H(z)* rozkłada się na sumę pięciu transmitancji składowych $H_k(z)$ w postaci:

$$H(z)=\sum_{k=1}^{5}\frac{B_k+C_k z^{-1}}{1-2|z_k|\cos(\varphi_k)z^{-1}+|z_k|^2 z^{-2}} \qquad (14)$$

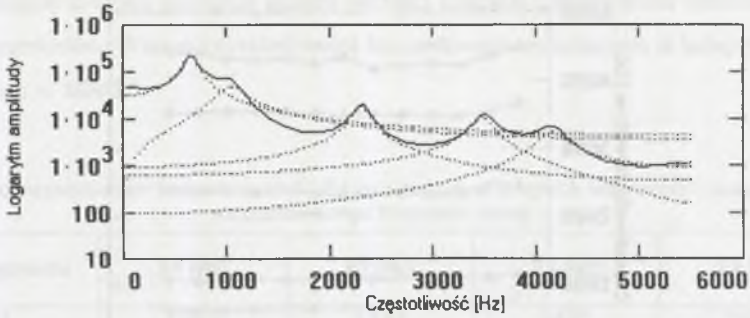
Na rysunku 11 przedstawiono rozpatrywane chwilowe widmo rozpoznawanego fragmentu mowy oraz przerywaną linią widmo formowane przez 5 filtrów o transmitancjach $H_k(z)$. Można zauważyć, że usunięcie pary biegunów rzeczywistych spowodowało zmianę charakterystyki dla niektórych częstotliwości. Nie ma to jednak wpływu na określanie położenia formantów.



Rys. 11. Widmo chwilowe rozpoznawanego fragmentu mowy oraz widmo estymowane za pomocą sumy pięciu charakterystyk składowych

Fig. 11. Spectrum of recognized fragement of speech and spectrum estimeted througt adding spectral-response of 5-th filters

Charakterystyki widmowe filtrów składowych przedstawiono na rysunku 12 i zaznaczono je linią przerywaną. Maksima poszczególnych charakterystyk składowych odpowiadają kolejnym częstotliwościom formantowym.



Rys. 12. Widmo chwilowe rozpoznawanego fragmentu mowy oraz poszczególne charakterystyki składowe

Fig. 12. Spectrum of recognized fragement of speech and spectral-response of 5-th filters

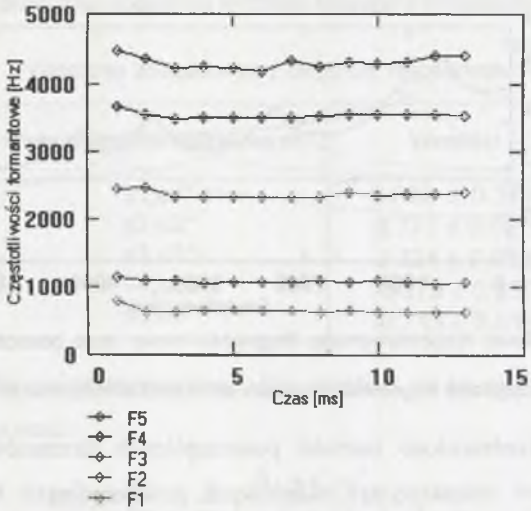
W tabeli 5 przedstawiono wartości poszczególnych formantów obliczone poprzez określenie maksimów charakterystyk składowych poszczególnych filtrów, w jednym z segmentów czasowych rozpoznawanego fragmentu mowy.

Tabela 5

Wartości częstotliwości formantowych wyliczonych w jednym z segmentów czasowych rozpoznawanego fragmentu mowy

Formant	Częstotliwość
F1	646 Hz
F2	1055 Hz
F3	2304 Hz
F4	3488 Hz
F5	4245 Hz

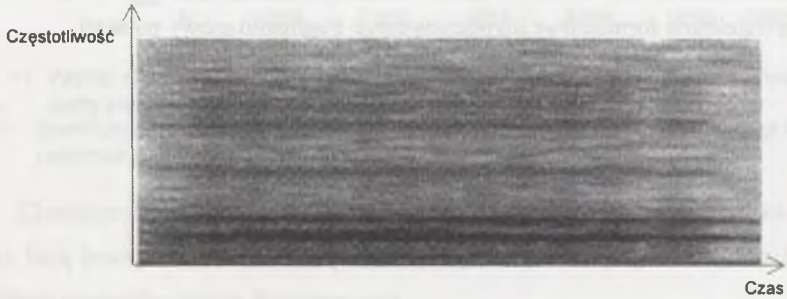
Po wyliczeniu wartości częstotliwości formantowych w każdym z segmentów czasowych rozpoznawanego sygnału mowy ustala się tzw. trajektorie formantowe, czyli zmiany położenia formantów w czasie artykulacji rozpoznawanego fragmentu mowy. Na rysunku 13 przedstawiono trajektorie formantowe rozpoznawanego fragmentu mowy polskiej.



Rys. 13. Trajektorie formantowe rozpoznawanego fragmentu mowy
Fig. 13. Formants trajektory of recognized speech

Powstałe trajektorie formantowe poddaje są dalszej analizie w celu ustalenia poszczególnych cech dystynktywnych. Minimalne różnice rozmieszczenia formantów w kolejnych segmentach czasowych świadczą o regularnym, prążkowym widmie analizowanego sygnału mowy. Regularny charakter widma może świadczyć, że rozpoznawany fragment sygnału mowy jest fonemem otwartym.

Jednym ze sposobów prezentacji zmian kształtu widma w czasie są tzw. **sonogramy**. Sonogram prezentuje się na płaszczyźnie w postaci prostokąta. Wymiar poziomy sonogramu odpowiada jednostce czasu, natomiast wymiar pionowy - częstotliwości. Amplitudę oznacza się poprzez proporcjonalne zaczernienie odpowiednich miejsc prostokąta. Rysunek 14 prezentuje sonogram rozpoznawanego fragmentu mowy.



Rys. 14. Sonogram rozpoznawanego fragmentu mowy
Fig. 14. Sonogram of recognized speech

W następnym kroku następuje analiza wartości czterech częstotliwości formantowych w każdym z segmentów. Wartości częstotliwości formantowych wyliczonych w kolejnych oknach przedstawiono w tabeli 6.

Tabela 6

Wartości częstotliwości formantowych F1-F4 wyliczonych w kolejnych segmentach czasowych rozpoznawanego fragmentu mowy

Nr segmentu	F1 [Hz]	F2 [Hz]	F3 [Hz]	F4 [Hz]
1	796.7	1141	2433	3639
2	646.0	1098	2433	3531
3	624.5	1077	2326	3488
4	646.0	1055	2304	3488
5	667.5	1055	2326	3488
6	667.5	1055	2326	3488
7	646.0	1055	2326	3488
8	624.5	1077	2326	3510
9	646.0	1077	2390	3531
10	646.0	1077	2347	3553
12	624.5	1055	2369	3531
13	646.0	1055	2369	3531
14	624.5	1055	2369	3510

Analizując wartości częstotliwości formantowych F1÷F4 można zauważyć, że są one charakterystyczne dla fonemów otwartych. Pierwsza cecha dystynktywna zostaje rozpoznana jako **otwartość.** W tabeli 7 przedstawiono zakresy częstotliwości formantowych charakterystyczne dla fonemów otwartych.

Tabela 7

Zakresy częstotliwości formantowych fonemów otwartych

Formant	Częstotliwość
F1	210 ÷ 900 Hz
F2	700 ÷ 2700 Hz
F3	2150 ÷ 3500 Hz
F4	3200 ÷ 4200 Hz

Rozróżnianie fonemów otwartych jest możliwe poprzez analizę wartości dwóch najniższych częstotliwości formantowych F1,F2. Ustalenie drugiej cechy dystynktywnej (MA) jest możliwe dzięki analizie wartości częstotliwości formantowej F2. Tabela 8 przedstawia wartości formantu F2 w zależności od drugiej cechy dystynktywnej, określającej miejsce artykulacji.

Tabela 8
Zakres częstotliwości formantowej F2 w zależności od cechy dystynktywnej określającej miejsce artykulacji dla fonemów otwartych

Cecha dystynktywna określająca miejsce artykulacji	F2 [Hz]
MA5 - przedniojęzykowość	1700 ÷ 2700 Hz
MA6 – środkowojezykowość	1000 ÷ 1700 Hz
MA7 – tylnojęzykowość	600 ÷ 1000 Hz

Wartości częstotliwości formantowej F2 rozpoznawanego fonemu wskazują na to, iż badany fonem jest środkowojęzykowy. Analiza wartości pierwszej częstotliwości formantowej F1 pozwala na ustalenie trzeciej cechy dystynktywnej określającej pionową pozycję języka podczas artykulacji. W tabeli 9 przedstawiono zakresy częstotliwości formantu F1 w zależności od drugiej i trzeciej cechy dystynktywnej.

Tabela 9
Zakres częstotliwości formantowej F1 w zależności od drugiej i trzeciej cechy dystynktywnej dla fonemów otwartych

Cecha dystynktywna określająca miejsce artykulacji	Wektor cech dystynktywnych	F1 [Hz]
MA5 – przedniojęzykowość CD1 – język wysoko	[4,5,1,0]	160 ÷ 300 Hz
MA5 – przedniojęzykowość CD2 – język nisko	[4,5,2,0]	300 ÷ 500 Hz
MA6 – środkowojęzykowość CD1 – język wysoko	[4,6,1,0]	300 ÷ 600 Hz
MA6 – środkowojezykowość CD2 – jezyk nisko	[4.6.2.0]	600 ÷ 800 Hz
MA7 – tylnojęzykowość CD1 – język wysoko	[4,7,1,0]	150 ÷ 400 Hz
MA7 – tylnojęzykowość CD2 – język nisko	[4,7,2,0]	400 ÷ 600 Hz

Na podstawie tabeli 9, analizując wartości pierwszego formantu F1 wyliczone w kolejnych segmentach czasowych analizowanego fragmentu sygnału mowy, można ustalić wektor cech dystynktywnych niezbędny do rozpoznania badanego fonemu otwartego. Tabela 10 przedstawia wektory cech dystynktywnych fonemów otwartych.

Tabela 10
Wektory cech dystynktywnych odpowiadające fonemom otwartym j. polskiego

Fonem	Wektor cech dystynktywnych
i	[4,5,1,0]
y	[4,5,2,0]
e	[4,6,1,0]
a	[4,6,2,0]
u	[4,7,1,0]
o	[4,7,2,0]

Określenie wektora cech dystynktywnych kończy proces automatycznego rozpoznawania mowy. Na podstawie otrzymanego w prezentowanym przykładzie wektora cech dystynktywnych $X=[4,6,2,0]$ można stwierdzić, że rozpoznawanym fonemem jest fonem otwarty, środkowojęzykowy, o niskim położeniu języka podczas artykulacji. Takie cechy dystynktywne odpowiadają fonemowi [a] języka polskiego.

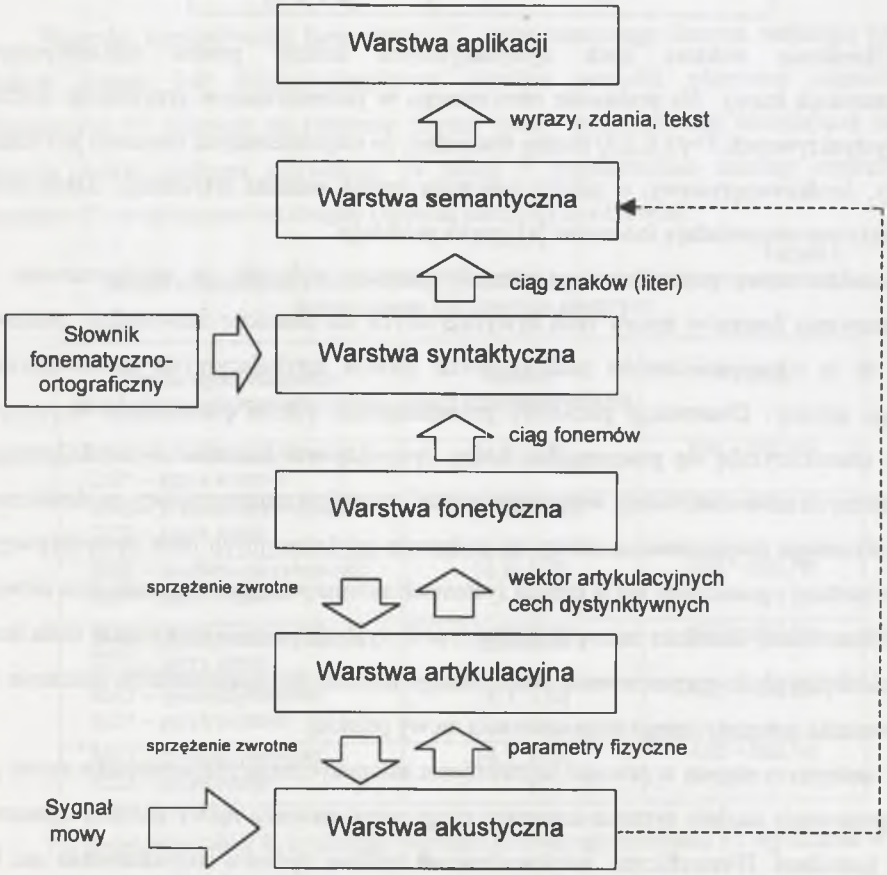
Analiza mowy pochodzącej od różnych mówców wykazała, że zaproponowany do rozpoznawania fonemów zestaw cech dystynktywnych ma charakter uniwersalny, ponieważ cechy te są odzwierciedleniem podstawowych zjawisk artykulacyjnych zachodzących u każdego mówcy. Obserwacje pozwoliły przeanalizować, jakimi parametrami fizycznymi mowy charakteryzują się poszczególne cechy dystynktywne fonemów wyartykułowanych przez różnych mówców. Należy więc przypuszczać, że wpływ zmiany mówcy na skuteczność automatycznego rozpoznawania mowy na podstawie artykulacyjnych cech dystynktywnych będzie bardziej ograniczony niż w innych systemach automatycznego rozpoznawania mowy.

Elementarny charakter zaproponowanych cech dystynktywnych oraz bardzo mała liczba cech niezbędnych do rozpoznawania pojedynczego fonemu ma fundamentalne znaczenie dla usprawnienia automatycznego rozpoznawania mowy polskiej.

Następnym etapem w procesie usprawnienia automatycznego rozpoznawania mowy jest zaproponowanie modelu systemu automatycznego rozpoznawania mowy polskiej opracowanymi metodami. Hierarchiczna, wielowarstwowa budowa systemu rozpoznawania ma być dodatkowym czynnikiem do usprawnienia całego procesu.

6. Propozycja wielopoziomowej, hierarchicznej struktury systemu rozpoznawania mowy polskiej

W wyniku przeprowadzonych badań nad usprawnieniem procesu rozpoznawania mowy polskiej opracowano model systemu automatycznego rozpoznawania mowy na podstawie opracowanego zestawu cech. Schemat blokowy hierarchicznego, wielowarstwowego systemu automatycznego rozpoznawania mowy polskiej przedstawiono na rysunku 15.



Rys. 15. Schemat blokowy wielowarstowego hierarchicznego systemu automatycznego rozpoznawania mowy poskiej
Fig. 15. Block diagram of multilevel, hierarchical system of speech recognition polish language

System ma budowę hierarchiczną, wielowarstwową. Każda z warstw systemu automatycznego rozpoznawania mowy realizuje odrębną część procesu rozpoznawania mowy. Dane wyjściowe z warstwy niższej są danymi wejściowymi dla warstwy nadrzędnej.

Dodatkowe sprzężenia zwrotne umożliwiają warstwom wyższym przesyłanie informacji do warstw niższych. Dzięki odpowiednim sprzężeniom zwrotnym (15) unika się określania cech nie biorących udziału w rozpoznawaniu danego fonemu oraz wyliczania parametrów fizycznych niezbędnych do ich ekstrakcji. Jest to czynnik usprawniający proces automatycznego rozpoznawania mowy polskiej.

Warstwa akustyczna systemu automatycznego rozpoznawania mowy, zwana także warstwą fizyczną, ma na celu wyodrębnienie parametrów fizycznych (akustycznych) sygnału mowy niezbędnych do określenia artykulacyjnych cech dystynktywnych, na podstawie których rozpoznane zostaną elementy językowe – fonemy.

Głównym zadaniem **warstwy artykulacyjnej** jest określenie artykulacyjnych cech dystynktywnych, na podstawie których rozpoznawane są kolejne fonemy. Określenie tych cech następuje na podstawie parametrów fizycznych otrzymanych w warstwie niższej (akustycznej). Sprzężenie zwrotne między warstwą artykulacyjną i akustyczną pozwala na przekazywanie informacji o tym, jakie parametry fizyczne sygnału mowy są aktualnie niezbędne do określenia wektora cech dystynktywnych.

Zadaniem **warstwy fonetycznej** jest rozpoznanie fonemu na podstawie wektora cech dystynktywnych określonych przez warstwę niższą (artykulacyjną). Wyjściową informacją z warstwy fonetycznej jest ciąg fonemów rozpoznanych w kolejnych segmentach czasowych analizowanego sygnału mowy. Dalsze jego przetwarzanie następuje w warstwie syntaktycznej systemu rozpoznawania mowy.

W **warstwie syntaktycznej** nieuporządkowany ciąg fonemów zamieniany jest na uporządkowany ciąg zgodny z fonematycznym zapisem rozpoznawanych słów. Etap ten jest bardzo złożony. Propozycja rozwiązania tego problemu sprowadza się do wykorzystania metod statystycznych [3]. Na tym etapie można także zastosować modele ukryte Markowa [25,6]. W następnym kroku następuje dekodowanie przenoszonego za pomocą mowy komunikatu, w efekcie czego otrzymuje się wypowiedź złożoną z wyrazów i zdań. Następuje przekształcenie uporządkowanego ciągu fonemów na ciąg liter odpowiadający wyrazom w języku polskim. Jest to możliwe dzięki zastosowaniu transkrypcji fonematycznej, która przekształca postać fonematyczną fragmentów mowy na postać ortograficzną [30].

Warstwa semantyczna pełni rolę uzupełniającą w systemie automatycznego rozpoznawania mowy i zajmuje się analizą treści rozpoznanej wypowiedzi. W warstwie tej przeprowadzana jest analiza zjawisk prozodycznych oraz gestów artykulacyjnych, które to niosą cenne informacje, bardzo istotne z punktu widzenia rozpoznawania [28].

Analiza zjawisk prozodycznych (takich jak głośność, wysokość, czas trwania, barwa i intonacja) pozwala wzbogacić rozpoznawane fragmenty mowy o dodatkowe informacje. Zjawiska te są istotne w sensie przekazywania informacji językowych i procesu porozumiewania się za pomocą mowy. Zastosowanie analizy prozodycznej języka polskiego pozwala usprawnić i udoskonalić proces automatycznego rozpoznawania mowy polskiej.

Informacja wyjściowa z warstwy semantycznej jest końcowym produktem procesu automatycznego rozpoznawania mowy. Jej wykorzystanie zależy od konkretnych zastosowań, które to znajdują swoje miejsce w najwyższej warstwie systemu automatycznego rozpoznawania mowy, zwanej **warstwą aplikacji**.

Zaproponowana wielowarstwowa, hierarchiczna struktura systemu automatycznego rozpoznawania mowy polskiej wykorzystuje szeroko rozumianą inżynierię wiedzy, w skład której wchodzi nie tylko wiedza dotycząca cyfrowego przetwarzania sygnału mowy, ale także wiedza wywodzącą się z fonetyki i fonologii języka polskiego. Starano się także wykorzystać wiedzę o mowie w strukturach bazujących na różnego rodzaju regułach, umożliwiających interpretację zgodną z semantyczną informacją zawartą w sygnale mowy.

Wiedza jest pozyskiwana z różnych źródeł. Przetwarzanie sygnału mowy polskiej, w poszczególnych warstwach systemu rozpoznawania opierało się na wiedzy:

- akustycznej, dotyczącej parametrów fizycznych sygnału mowy,
- fonetycznej, dotyczącej dźwięków języka polskiego,
- syntaktycznej, dotyczącej gramatyki,
- semantycznej, dotyczącej znaczenia,
- pragmatycznej, dotyczącej dziedziny zadań i aplikacji, czyli zastosowań automatycznego rozpoznawania mowy.

Tak rozbudowana struktura przekracza możliwości typowego systemu rozpoznawania mowy i posiada pewne cechy systemu automatycznego rozpoznawania i **rozumienia mowy** [20,24]. Złożone architektury systemów często są przyczyną spowolnienia ich pracy, dlatego też zdecydowano się na podział zadań na etapy realizowane przez poszczególne warstwy systemu automatycznego rozpoznawania mowy.

Dzięki odpowiednim sprzężeniom zwrotnym zdołano uniknąć zbędnej analizy mającej na celu ekstrakcję cech dystynktywnych nie biorących udziału w rozpoznawaniu danego fonemu oraz wyliczania parametrów fizycznych niezbędnych do ekstrakcji cech.

Taka organizacja systemu automatycznego rozpoznawania mowy polskiej w znaczącym stopniu przyczynia się do usprawnienia tego procesu.

Przeprowadzone próby rozpoznawania mowy oparte na zaproponowanym systemie wykazują wiele korzystnych własności polegających między innymi na skutecznym rozpoznawaniu słów o podobnym brzmieniu. Stało się to możliwe dzięki możliwości wychwycenia niewielkich różnic w artykulacji poszczególnych fonemów.

Tego typu właściwości systemu automatycznego rozpoznawania mowy polskiej świadczą o prawidłowym doborze artykulacyjnych cech dystynktywnych przeznaczonych do rozpoznawania fonemów. Potwierdza to także tezę, że rozpoznawanie mowy polskiej można usprawnić poprzez skuteczne rozpoznawanie fonemów na podstawie artykulacyjnych cech dystynktywnych.

Dodatkowym czynnikiem zwiększającym możliwości systemu rozpoznawania mowy polskiej jest zastosowanie analizy zjawisk prozodycznych i gestów artykulacyjnych. Analiza tych zjawisk pozwala wzbogacić rozpoznawane fragmenty mowy o dodatkowe informacje, pełniące istotną rolę w procesie porozumiewania się za pomocą mowy.

7. Podsumowanie

Główną ideą precentowanych koncepcji jest stwierdzenie, że informacja zawarta w mowie języka polskiego przenoszona jest przez cechy artykulacyjne elementów mowy. Znaczy to, że można opisać mowę małym zbiorem wolno zmieniających się artykulacyjnych cech dystynktywnych. Rozpoznawanie mowy można zatem oprzeć na rozróżnianiu poszczególnych cech artykulacyjnych jej elementów.

Zaproponowano metodę rozpoznawania fonemów języka polskiego. W tym celu dla każdego fonemu dobrano te cechy artykulacyjne, które pełnią funkcję dystynktywną, czyli takie, które odpowiedzialne są za ich rozróżnianie, a co za tym idzie – rozpoznawanie. W ten sposób powstał zestaw artykulacyjnych cech dystynktywnych, na podstawie których można rozpoznawać fonemy. Na szczególną uwagę zasługuje niewielka liczba cech dystynktywnych niezbędnych do rozpoznania pojedynczego fonemu. Do rozpoznawania większości fonemów wystarczające jest określenie trzech cech dystynktywnych. Ale są także fonemy rozpoznawane za pomocą dwóch lub nawet jednej cechy dystynktywnej. Fakt ten ma zasadnicze znaczenie w usprawnieniu całego procesu.

Aby skonstruować system automatycznego rozpoznawania mowy polskiej na podstawie zaproponowanego zestawu cech, niezbędne jest określenie związku między parametrami fizycznymi mowy polskiej a jej artykulacją. Podjęto próbę znalezienia odwzorowania między

parametrami fizycznymi sygnału mowy a układami artykulacji, które go wytwarzają. Było to możliwe dzięki skonstruowanym przez autora narzędziom służącym do analizy mowy, zaimplementowanym w środowisku Mathcad, o roboczej nazwie AM10.

Żmudne badania, obserwacje i pomiary prowadzone za pomocą systemu analizy mowy AM10, szczegółowo opisanego w jednym z rozdziałów rozprawy doktorskiej, pozwoliły na opracowanie metod określania poszczególnych cech artykulacyjnych na podstawie fizycznych parametrów sygnału mowy.

W pracy nad usprawnieniem procesu automatycznego rozpoznawania mowy polskiej skupiono się głównie na rozpoznawaniu fonemów języka polskiego. Przeprowadzono doświadczenia mające na celu rozpoznawanie fonemów języka polskiego na podstawie zaproponowanego zestawu artykulacyjnych cech dystynktywnych. W tym celu rozbudowano system analizy mowy AM10 o dodatkowy moduł służący do rozpoznawania fonemów. Rozpoznawanie fonemów języka polskiego pozwala na rozpoznawanie dowolnych fragmentów mowy polskiej. Próby rozpoznawania przeprowadzono dla izolowanych fonemów, wyrazów oraz dla bardziej złożonych struktur.

Na szczególną uwagę zasługuje skuteczne rozpoznawanie fonemów o podobnym brzmieniu. Stało się to możliwe dzięki wychwyceniu niewielkich różnic w artykulacji poszczególnych fonemów. Świadczy to o trafności wyboru zaproponowanego zestawu artykulacyjnych cech dystynktywnych do procesu rozpoznawania fonemów języka polskiego oraz potwierdza przyjętą tezę, iż automatyczne rozpoznawanie mowy polskiej można usprawnić skutecznie rozpoznając fonemy języka polskiego na podstawie zestawu elementarnych artykulacyjnych cech dystynktywnych.

Głównym celem badań jest wykazanie możliwości dalszego usprawnienia procesu rozpoznawania mowy polskiej przez zastosowanie odpowiedniej architektury systemu rozpoznawania.

Zaproponowano model kompletnego systemu rozpoznawania mowy polskiej opartego na architekturze hierarchicznej – wielowarstwowej.

Podstawową cechą pełnego systemu automatycznego rozpoznawania mowy polskiej jest umiejętność rozpoznawania i przetwarzania fonemów języka polskiego. Realizację tego celu powierza się poszczególnym warstwom systemu, które odpowiedzialne są za kolejne etapy procesu rozpoznawania. Kolejne warstwy, poczynając od najniższej, realizują proces rozpoznawania mowy na coraz wyższym poziomie abstrakcji. Dane wyjściowe warstwy niższej stają się informacją wejściową dla warstwy wyższej. Odpowiednie sprzężenia zwrotne pozwalają warstwom wyższym na przekazywanie informacji do warstw niższych, dzięki

czemu zwiększa się efektywność rozpoznawania mowy, bowiem unika się wyznaczania parametrów fizycznych i ekstrakcji cech dystynktywnych nie biorących udziału w rozpoznawaniu danego fonemu.

Istnieje możliwość jeszcze większego usprawnienia całego procesu dzięki współbieżnej, równoległej pracy poszczególnych warstw systemu rozpoznawania. Aby zbadać efektywność wielowarstwowej struktury systemu rozpoznawania mowy polskiej przeprowadzono szereg doświadczeń. Częściowe zaimplementowanie tej struktury w środowisku Mathcad pozwoliło zauważyć wzrost efektywności całego procesu rozpoznawania fonemów języka polskiego.

Zastosowanie wyższych warstw zaproponowanego systemu, szczególnie warstwy semantycznej realizującej elementy rozumienia mowy, wpłynęło na zwiększenie jego możliwości.

Dodatkowe możliwości usprawnienia procesu rozpoznawania mowy polskiej tkwią w zastosowaniu elementów rozumienia mowy, przez analizę zjawisk prozodycznych oraz gestów artykulacyjne zawierających dodatkowe informacje w rozpoznawanym sygnale mowy. Analiza prozodyczna wchodzi w skład najwyższej warstwy semantycznej zaproponowanego systemu. Mowa jest niejednoznaczna. Wypowiadane te same słowa i zdania często mają wiele różnych interpretacji, a różnice w wymowie różnych znaczeniowo kwestii mogą być bardzo małe. Mimo że warstwa semantyczna pełni rolę uzupełniającą, niepodważalne jest jej znaczenie w usprawnieniu całego procesu.

Zaproponowano wiele sposobów usprawnienia automatycznego rozpoznawania mowy polskiej w różnych etapach tego procesu. Na szczególną uwagę zasługują:

- opracowane i zaimplementowane w środowisku Mathcad algorytmy wyliczania parametrów fizycznych sygnału mowy niezbędnych do ekstrakcji artykulacyjnych cech dystynktywnych, a w szczególności trajektorii formantowych i zmian częstotliwości tonu krtaniowego,
- wybór artykulacyjnych cech dystynktywnych pozwalających efektywnie rozpoznawać fonemy języka polskiego,
- określenie zależności między parametrami fizycznymi sygnału mowy a cechami dystynktywnymi, określającymi artykulację,
- możliwość rozpoznawania fonemów o podobnym brzmieniu, dzięki wychwyceniu niewielkich różnic w ich artykulacji,
- mała liczba cech dystynktywnych niezbędnych do rozpoznawania pojedynczego fonemu,

- zaproponowanie wielowarstwowej, hierarchicznej struktury systemu automatycznego rozpoznawania mowy oraz usprawnienie jego działania poprzez odpowiednie sprzężenia zwrotne między poszczególnymi warstwami,
- wzbogacenie możliwości systemu automatycznego rozpoznawania mowy polskiej poprzez dodanie warstwy semantycznej, realizującej badanie znaczenia rozpoznawanych fragmentów mowy polskiej poprzez analizę prozodyczną.

Zaproponowane rozwiązania nie wyczerpują tematu automatycznego rozpoznawania mowy polskiej, ale są krokiem do usprawnienia tego procesu. Wiedza zdobyta w trakcie prowadzenia badań nad automatycznym rozpoznawaniem mowy może być także pomocna w konstruowaniu systemów generacji syntetycznej mowy wysokiej jakości.

Rozpoznawanie mowy jest problemem trudnym. Jest wiele zagadnień związanych z tym procesem, których realizację można podjąć w przyszłości. Po przeanalizowaniu wyników badań własnych nasuwają się pomysły dalszego usprawnienia rozpoznawania mowy. Jednym z rozwiązań może być zastosowanie innych elementarnych jednostek językowych, a mianowicie allofonów. W tym wypadku zestaw artykulacyjnych cech dystynktywnych uległby zmianie. Należałoby zbadać, czy korzyści płynące z możliwości rozpoznawania mniejszych jednostek językowych zrekompensowałyby problemy związane z określaniem większej liczby artykulacyjnych cech dystynktywnych.

Literatura

1. Abramson N.: Teoria informacji i kodowania, PWN, Warszawa 1969.
2. Basztura C.: Źródła, sygnały i obrazy akustyczne, WKiŁ, Warszawa 1988.
3. Becchetti C., Ricotti L.P.: Speech Recognition, Wiley & Sons London 1999.
4. Deller J.: Discrete-Time Processing of Speech Signals, McMillan Co., 1995.
5. Dukiewicz L., Sawicka I.: Fonetyka i fonologia, Gramatyka współczesnego języka polskiego, Instytut Języka Polskiego PAN, Kraków 1995.
6. Huang X. D., Ariki Y., Jack M. A.: Hidden Markov Models for Speech Recognition, Edinburgh University Press, 1990.
7. Jassem W.: Mowa a nauka o łączność, PWN, Warszawa 1974.
8. Jassem W.: Podstawy fonetyki akustycznej, PWN, Warszawa 1973.
9. Kłosowski P.: Automatyczne rozpoznawanie obrazów akustycznych sygnału mowy, Materiały konferencyjne Krajowego Sympozjum Telekomunikacji , Bydgoszcz 1998, wyd. Instytutu Telekomunikacji Politechniki Warszawskiej, t. B, ss. 245-252, Warszawa 1998.

10. Kłosowski P.: Oprogramowanie wspomagające niewidomych, Materiały konferencyjne Krajowego Sympozjum Telekomunikacji , Bydgoszcz 1997, wyd. Instytutu Telekomunikacji Politechniki Warszawskiej, t. D, ss. 61÷66, Warszawa 1997.
11. Kłosowski P.: Usprawnienie procesu rozpoznawania mowy w oparciu o cechy dystynktywne fonemów języka polskiego, Materiały konferencyjne Krajowego Sympozjum Telekomunikacji, Bydgoszcz 1999, wyd. Instytutu Telekomunikacji Politechniki Warszawskiej, t. B, ss.353÷359, Warszawa 1999.
12. Kłosowski P.: Wykorzystanie spektrogramów binarnych do rozpoznawania mowy języka polskiego, Materiały konferencyjne Krajowego Sympozjum Telekomunikacji , Bydgoszcz 1999, wyd. Instytutu Telekomunikacji Politechniki Warszawskiej, t. D, ss. 335÷341, Warszawa 1999.
13. [Kłosowski P.: Programowa synteza mowy polskiej, Materiały konferencyjne Krajowego Sympozjum Telekomunikacji, Bydgoszcz 1996, wyd. Instytutu Telekomunikacji Politechniki Warszawskiej, t. A, ss. 297÷305, Warszawa 1996.
14. Kłosowski P., Izydorczyk J.: Automatyczne rozpoznawanie mowy, Międzynarodowa Konferencja Multimedia w Nauczaniu Języka Rodzimego jako obcego. 7-8 XII 1998, Katowice 1998, Postscriptum – kwartalnik Szkoły Języka i Kultury Polskiej, ss. 81-88 , 1998.
15. Kłosowski P., Izydorczyk J.: Acoustic properties of Polish vowels, Bulletin of the Polish Academy of Science – Technical Sciences, vol. 47, no.1, pp. 29÷37, Warsaw 1999.
16. Kłosowski P., Izydorczyk J.: Allofoniczna synteza mowy polskiej, Materiały konferencyjne Krajowego Sympozjum Telekomunikacji, Bydgoszcz 1997, wyd. Instytutu Telekomunikacji Politechniki Warszawskiej, t. D, ss. 66÷72, Warszawa 1997.
17. Kłosowski P., Izydorczyk J.: Allophonic speech synthesis for Polish, Proceedings of he XX-th National Conference Circuit Theory and Electronics Circuits, pp. 535÷541, Kołobrzeg 1997.
18. Kłosowski P., Izydorczyk J.: Wytwarzanie mowy syntetycznej INFOTEL 3'98 – kwartalnik, 3 ss. 84÷87, Bydgoszcz 1998.
19. Kłosowski P., Konopacki J.: Text-to-Speech System for Polish, Proceedings of the XIX-th National Conference Circuit Theory and Electronics Circuits, vol. 2, p.445÷450, Krynica 1996.
20. Laface P., De Mori R.: Speech Recognition and Understanding, Recent Advances, Trends and Applicationcs, NATO ASI Series, Springer, Verlag Berlin 1992.

21. Oppenheim V., Schafer W.: Cyfrowe przetwarzanie sygnałów, WKiŁ, Warszawa 1979.
22. Ostraszewska D., Tambor J.: Podstawowe wiadomości z fonetyki i fonologii współczesnego języka polskiego, Skrypt Uniwersytetu Śląskiego nr 526, Wydawnictwo Uniwersytetu Śląskiego, Katowice 1997.
23. Picone J.: Fundamentals of Speech Recognition, Institute for Signal and Information Processing, Department of Electrical and Computer Engineering Mississippi State University, (http://www.isip.msstate.edu), 1998.
24. Picone J., Ebel W.J., Deshmukh N.: Automated Speech Understanding: The Next Generation, Digital Signal Processing Technology, vol. CR57, pp.101÷114, 1995.
25. Rabiner L. R. and Juang B. H.: An Introduction to Hidden Markov Models, IEEE Acoustics, Speech and Signal Processing Magazine, vol. 3, pp. 4÷16, 1986.
26. Rabiner L. R., Biing-Hwang Juang, Fundamentals of Speech Recognition, Prentice Hall, Englewood Cliffs, New Jerswey 1993.
27. Rabiner L. R., Schafer R. W.: Digital Processing of Speech Signals, Signal Processing Series, Prentice-Hall, Englewood Cliffs, New Jersey 1978.
28. Sagisaka Y., Campbell N., Higuchi N.: Computing Prosody, Computational Models for Processing Spontaneous Speech, Springer Verlag, Berlin 1996.
29. Steffen M.: Częstość występowania głosek polskich, Biuletyn PTJ XVI, Warszawa 1957.
30. Steffen-Batogowa M.: Automatyzacja transkrypcji fonematycznej tekstów polskich, PWN, Warszawa 1975.
31. Tadeusiewicz R.: Sygnał mowy, WKiŁ, Warszawa 1988.

Recenzent: Prof. Piotr Wach

Abstract

Speech recognition is, in its most general form, a conversion from an acoustics waveform to a written equivalent of the spoken message information. The nature of the speech recognition problem is heavy dependent upon the constraints placed on speaker, speaking situation and message context. The potential applications of speech recognition systems are many and varied; e.g. a voice operated typewriter and voice communication with computers. Also a speech recognizing system combined with a speech synthesizing system comprises the ultimate low bit rate communication systems. There are generally recognized to be three major areas within the

general area of man-machine communication by voice. These areas include voice response systems, speaker recognition systems and speech recognition systems. This paper presents new improved method for speech recognition based on detection of distinctive acoustic parameters of phonemes in polish language. Distinctivity has been assumed as a most important selection of parameters, which have represented objects from recognized classes. This method are sufficient grounds for development of speech recognition system for polish language.