Praca habilitacyjna

Henryk Kubzdela

METODA GLOBALNEGO ROZPOZNAWANIA WYRAZÓW NA PODSTAWIE SPEKTROGRAMÓW BINARNYCH

28/1986

WARSZAWA 1986

ISSN 0208-5658

Praca habilitacyjna

Praca wpłynęła do Redakcji dnia 2 czerwca 1986 r.

Na prawach rękopisu

Instytut Podstawowych Problemów Techniki PAN

Nakład 180 egz. Ark.wyd. 8,8. Ark.druk. 11 .
Oddano do drukarni w czerwcu 1986 r.
Nr zamówienia 349/86 Nakład 180+23.

Warszawska Drukarnia Naukowa, Warszawa,
ul.Śniadeckich 8

Henryk Kubzdela

Pracownia Fonetyki Akustycznej
IPPTPAN

METODA GLOBALNEGO ROZPOZNAWANIA WYRAZÓW NA PODSTAWIE SPEKTROGRAMÓW BINARNYCH

Streszczenie

Rozprawa dotyczy pewnej metody automatycznego rozpoznawania wyrazów wymawianych w izolacji i należących do pewnego ograniczonego zbioru haseł. Część zasadniczą rozprawy, prezentującą dorobek autora, poprzedza wstęp oraz przegląd problematyki rozpoznawania wyrazów, głównie w sposób globalny. Przegląd ten zawiera omówienie stosowanych przez różnych autorów metod rozwiązania podstawowych zagadnień występujących w globalnym rozpoznawaniu wyrazów, takich jak tworzenie obrazów akustycznych wypowiedzi wyrazów i obrazów wzorcowych poszczególnych wyrazów oraz wzajemne porównywanie ich. Przegląd zawiera też dane o wynikach automatycznego rozpoznawania wyrazów uzyskiwanych przez różnych badaczy, a także opisy zastosowań tego rozpoznawania.

Prezentację swojej metody automatycznego rozpoznawania wyrazów rozpoczął autor od przedstawienia układu, który stanowił nieodzowne narzędzie w jego badaniach, i którego przygotowanie stanowiło ważny etap przedsięwzięcia uwieńczonego niniejszą rozprawą. Poza minikomputerem i jego standardowymi urządzeniami peryferyjnymi, podstawowe elementy tego układu skonstruował autor sam. Odwołując się do wiedzy o strukturach akustycznych dźwięków mowy przyjęto w pracy, że w reprezentacji wypowiedzi wyrazu dla jego automatycznego rozpoznawania powinna być zawarta informacja o podstawowych cechach widmowych. Stwierdzono, iż temu wymaganiu sprostać może określonego typu obraz o strukturze binarnej zwany spektrogramem binarnym i charakteryzujący się korzystnie małą objętością informacyjną. Przedstawiona została opracowana przez autora metoda przetworzenia sygnału mowy w tego rodzaju obraz. Na metodę tę składa się m.in. sposób wygładzenia widma sygnału mowy za pomocą specjalnie do te-

go wyznaczonej funkcji wygładzającej oraz sposób wyznaczenia widma binarnego w oparciu o dwustopniową ocenę wypukłości obwiedni widma wygładzonego.

W kolejnym rozdziale przedstawiono sposób, w jaki autor rozwiązał w swojej metodzie zagadnienie porównywania spektrogramów binarnych. Charakteryzują go następujące cechy: 1/ porównywanie dwóch spektrogramów odbywa się fragmentami. Fragmentem jest ciąg kilku kolejnych widm binarnych; 2/ podobieństwo fragmentów wyraża względna liczebność identycznych, niezerowych parametrów widmowych w porównywanych fragmentach; 3/ podobne fragmenty we wzajemnie porównywanych spektrogramach binarnych poszukiwane są w otoczeniu trajektorii quasiliniowej normalizacji ich długości.

Autor zaproponował dwa sposoby wyznaczenia wzorcowego spektrogramu wyrazu oparte o zasadę, iż wzorzec powinien wyrażać te cechy widmowe, które wystąpiły w większości spektrogramów binarnych wypowiedzi adaptacyjnych.

Będąca przedmiotem niniejszej rozprawy metoda globalnego rozpoznawania wyrazów w oparciu o spektrogramy binarne zrealizowana została w opracowanym przez autora modelu rozpoznającym ROWBIR 1. Ze względu na niskie parametry jedynego minikomputera dostępnego przy wykonywaniu tej pracy, w modelu ROWBIR 1 użyto tylko jeden ze wspomnianych dwóch sposobów tworzenia wzorców. Wskazano na możliwość uproszczenia rozpoznawania wyrazów poprzez odpowiednią konstrukcję inwentarza wzorców polegającą na połączeniu częściowo podobnych wzorców w grupy i wyłonieniu dla każdej grupy osobnego reprezentanta. Przedstawiono szczegółowo tryb rozpoznawania wyrazu drogą porównywania obrazu binarnego danej wypowiedzi z poszczególnymi wzorcami. Wyłoniony wzorzec w porównaniu z innymi wykazuje lepsze lub analogiczne podobieństwo do obiektu w największej liczbie fragmentów.

Opracowana przez autora metoda rozpoznawania wyrazów osiągnęła swój końcowy kształt poprzez szereg modyfikacji modelu rozpoznającego ROWBIR 1. Poszczególne wersje modelu różniły się między sobą pod względem zakresu analizy widmowej, liczby parametrów widma binarnego oraz zakresu porównań lokalnych

dwóch spektrogramów. Próby testowe poszczególnych wersji modelu dotyczyły różnych słowników i różnych głosów. Zmniejszano liczbę parametrów widma binarnego rownocześnie obejmując rozpoznawaniem coraz więcej wyrazów. Wyniki testów każdej wcześniejszej wersji modelu rozpoznającego inspirowały do jego modyfikacji, w wyniku której powstawała wersja nowa, bardziej uproszczona. We wszystkich testach rozpoznawania w oparciu o indywidualne wzorce uzyskiwano bez względu na liczbę parametrów widma binarnego oraz wielkość słownika podobną poprawność rozpoznawania, zawartą w granicach 96-97%.

W wersji modelu rozpoznającego uznanej za dotychczas najkorzystniejszą operuje się widmami binarnymi 16-parametrycznymi oraz fragmentami dwuwidmowymi. Dla tej wersji szerokość otoczenia quasiliniowej normalizacji długości, w którym poszukiwany jest podobny fragment, równa się rozciągłości siedmiu kolejnych fragmentów. Wersję tę przetestowano w oparciu o słownik 100-wyrazowy.

Przeprowadzono też próby rozpoznawania wyrazów w oparciu o wzorce wspólne dla pary głosów uzyskując wyniki w granicach od 82.7 do 98.8% zależnie od głosu i partnera dzielącego z danym głosem wspólne wzorce. Zagadnienie rozpoznawania wyrazów niezależnego od cech indywidualnych głosów, wymaga dalszych wnikliwych studiów.

W osobnym rozdziale scharakteryzowano model ROWBIR 1, wyszczególniając zasadnicze cechy zarówno samej metody, jak i sposobu jej technicznej realizacji. Zaakcentowano atrakcyjność zastosowanej formy wyrażenia wyrazu, dzięki której przedstawiona w rozprawie metoda automatycznego rozpoznawania wyrazów stanowi korzystną propozycję dla konstrukcji przyszłych modeli użytkowych.

W ostatnim rozdziale przedstawiono hipotetyczną metodę rozpoznawania połączeń dwóch fonemów w mowie w oparciu o ich obrazy binarne. Przeprowadzono teoretyczną analizę porównawczą rozpoznawania wyrazów w sposób globalny oraz przy użyciu tej metody. Wynik tego porównania wskazuje, że dla słowników liczących mniej niż 230 słów zdecydowanie bardziej opłacalna jest metoda rozpoznawania w sposób globalny. W zakresie od 230 do

do 1000 słów obie metody są konkurencyjne, i to zarówno ze względu na rozmiary pamięci koniecznej do przechowywania wzorców, jak i na ilość operacji, z jakich składają się procedury rozpoznawania. Wyniki te świadczą korzystnie o globalnej metodzie rozpoznawania wyrazów. Dopiero powyżej 1000 słów wyłącznie opłacalną z rozpatrywanego punktu widzenia staje się metoda rozpoznawania w sposób segmentalny.

1. Wstęp

Rozprawa niniejsza stanowi podsumowanie kilkuletniej samodzielnej pracy autora nad metodą automatycznego rozpoznawania wyrazów wymawianych w izolacji i należących do pewnego zamkniętego zbioru haseł. Zadanie autora dotyczyło więc pewnego fragmentu bardzo szerokiego i złożonego problemu automatycznego rozpoznawania mowy w pełnym tego słowa znaczeniu. Zagadnienie automatycznego rozpoznawania jedynie izolowanych wyrazów wzbudzało zawsze szerokie zainteresowanie badaczy, którzy w perspektywie traktowali jego rozwiązanie jako ważny przyczynek do osiągnięcia przez automat pełnej symulacji zdolności człowieka w zakresie percepcji mowy. Wysoką rangę nadawały też zawsze temu zagadnieniu przewidywania natychmiastowych korzyści praktycznych wynikających z jego rozwiązania.

Wśród istniejących metod automatycznego rozpoznawania wyrazów wymawianych w izolacji, liczną grupę stanowią takie, w których wypowiedź rozpatruje się jako pojedynczy element z pominięciem faktu, iż jest on ciągiem określonych segmentów akustyczno-fonetycznych. Rozpoznawanie wyrazów według takich metod określić można mianem globalnego. Popularność rozpoznawania wyrazów w sposób globalny pochodzi prawdopodobnie z trzech przyczyn. Pierwszą z nich jest brak dotychczas niezawodnej metody segmentacji sygnału mowy na podstawowe elementy fonetyczno-akustyczne, mimo podejmowania wielu prób jej stworzenia . Przyczynę drugą stanowi trudność adaptacji systemu przewidzianego do rozpoznawania elementów segmentalnych w mowie, a także okoliczność, iż poszczególne języki różnią się między sobą pod względem rodzaju i liczby tych elementów. Przyczyną trzecią wydaje się być fakt, iż badacze pracujący nad metodami automatycznego rozpoznawania wyrazów są na ogół specjalistami w dzie-

dzinach technicznych lub matematyczno-fizycznych i jako tacy nie opierają swoich koncepcji na przesłankach, jakich dostarcza wiedza fonetyczno-akustyczna. Przedmiotem ich zainteresowania jest przede wszystkim dość ogólnie pojmowany sygnał mowy. Z pierwszego z wymienionych powodów autor skłonił się także ku koncepcji rozpoznawania wyrazów w sposób globalny, postanawiając jednocześnie uwzględnić w założeniach swojej metody możliwość późniejszego zaadaptowania jej do rozpoznawania określonego typu elementów segmentalnych, bez potrzeby uprzedniego wyznaczania granic pomiędzy nimi.

Głównym celem, jaki przyświecał autorowi rozprawy było opracowanie metody rozpoznawania wyrazów opartej o prostą, lecz jednocześnie reprezentatywną formę przedstawienia wypowiedzi wyrazu. Miernikiem prostoty tej formy jest wielkość jej objętości informacyjnej, zaś jej reprezentatywność wyraża się zawartością w niej informacji o dystynktywnych cechach dźwięków mowy.

Wiedza fonetyczno-akustyczna wskazuje, iż najbardziej dystynktywny opis dźwięków mowy przedstawiają niektóre szczegóły ich obrazów widmowych. One zatem powinny znaleźć się w uproszczonej reprezentacji wypowiedzi wyrazu.

Forma, w jakiej wyrażone zostają wypowiedzi wyrazów decyduje o złożoności procedur składających się na proces rozpoznawania. Innymi słowy, każdą metodę automatycznego rozpoznawania wyrazów cechują pewne rozwiązania, na które pozwala przyjęta w niej forma reprezentacji wyrazu.

Jedną z prostych form wyrażenia wypowiedzi wyrazów w zastosowaniu do ich rozpoznawania może być tak zwany „spektrogram binarny". Pod pojęciem spektrogramu binarnego spotyka się różne binarne reprezentacje wybranych cech widmowych. Autor decydując się na użycie tej formy postanowił obrać taką zasadę wyznaczania widma binarnego, która gwarantowałaby, że wyrażone zostaną w tym widmie ważne cechy dystynktywne dźwięków mowy, istotne dla roli, jaką spektrogram binarny ma odgrywać w rozpoznawaniu wyrazów. W tym zadaniu zawierał się też postulat, aby widmo binarne składało się z niewielkiej liczby parametrów gwarantujących dobrą efektywność rozpoznawania potwierdzoną odpowiednimi testami.

Innym ważnym zagadnieniem przy opracowywaniu metody globalnego rozpoznawania wyrazów jest zagadnienie porównywania obrazów będących konkretnymi realizacjami przyjętej formy reprezentacji wyrazu. Jego rozwiązanie polega na znalezieniu odpowiednich miar dla wyrażenia podobieństw lokalnych oraz podobieństwa globalnego porównywanych obrazów. Ponieważ porównywanie obrazów jest działaniem wielokrotnym w procesie rozpoznawania wyrazów w sposób globalny, dążyć należy do użycia takiej miary podobieństwa, której stosowanie nie wymaga wykonywania złożonych obliczeń. Miarę podobieństwa dostosowuje się do wybranej formy reprezentacji wyrazu. Obraz wypowiedzi wyrazu w formie spektrogramu binarnego pozwala na użycie miary podobieństwa, na którą składają się proste działania matematyczno-logiczne.

Posługując się spektrogramami binarnymi można łatwo skonstruowac obraz wzorcowy wyrazu. Posiada on także formę spektrogramu binarnego charakteryzującego się względnie małą objętością informacyjną.

W automatycznym rozpoznawaniu wyrazów występuje zagadnienie wzorców indywidualnych lub wspólnych dla pewnego grona głosów. W testach przedstawionej tutej metody rozpoznawania w oparciu o spektrogramy binarne zostało to zagadnienie w pewnym zakresie uwzględnione.

Decydując się na badania mające służyć opracowaniu metody globalnego rozpoznawania wyrazów na podstawie spektrogramów binarnych autor stanął przed koniecznością przygotowania sobie odpowiednich narzędzi badawczych. Najważniejsze z nich skonstruował autor osobiście, co zajęło mu kilka lat pracy. Urządzenia własnej konstrukcji zostały włączone w system minikomputerowy MERA 303, z którym razem utworzyły bazę techniczną pozwalającą na przeprowadzenie zamierzonych badań.

Metoda opracowana przez autora przedstawiona została w niniejszej rozprawie na tle przeglądu problematyki globalnego rozpoznawania wyrazów, któremu poświęcony został w całości rozdział drugi.

2. Przegląd problematyki

2.1. Rys historyczny

W społeczności ludzkiej mowa jest podstawowym środkiem komunikowania się. Za pomocą mowy ludzie przekazują sobie ustawicznie ogromne ilości informacji o najprzeróżniejszej treści, wadze czy wartości. Już we wczesnych dziełach literackich dawał człowiek wyraz tęsknotom za światem, w którym mowa ludzka trafiałaby także do zwierząt i martwych przedmiotów. Mowa jest bowiem najwygodniejszą dla człowieka formą wydawania rozkazów i poleceń, stawiania zapytań i podawania informacji. W rzeczywistym świecie próbuje człowiek posługiwać się mową w kontaktowaniu się z niektórymi zwierzętami. Zakres języka jaki w tym przypadku wchodzi w grę jest jednak bardzo ograniczony, a skuteczność oddziaływania niewielka. W miarę rozwoju cywilizacji technicznej zaczęła rodzić się możliwość użycia mowy do oddziaływania na maszyny. Powstała nowa dziedzina nauki zwana automatycznym rozpoznawaniem mowy - ARM .

Automatyczne rozpoznawanie mowy jest procesem technicznym prowadzącym do zdekodowania informacji zawartej w wypowiedzi człowieka na użytek automatu. Ze względu na ogromną złożoność problemu ARM zakres jego rozwiązywania ograniczono najpierw do izolowanych wyrazów, a następnie stopniowo go poszerzano. Badania prowadzone w laboratoriach akustycznych i fonetyczno-akustycznych instytutów naukowych w wielu krajach świata zaowocowały powstaniem różnych metod i modeli automatycznego rozpoznawania mowy. Jak słusznie stwierdza LEA (1980), rozpoznawanie mowy jest problemem interdyscyplinarnym i ma swoje korzenie w sięgających odległej przeszłości studiach nad językiem i dźwiękiem, w fizjologii, psychologii i automatyce. Prawdopodobnie pierwszą próbę rozpoznawania mowy przedstawił DREYFUS - GRAF w roku 1950. W jego urządzeniu zwanym „Stenosonografem" sygnał mowy przepuszczany był przez sześć filtrów środkowo-przepustowych. Ich wyjścia połączone były z cewkami odchylającymi rozmieszczonymi wokół lampy oscyloskopowej. Dla poszczególnych sekwencji dźwięków mowy pojawiały się na ekranie różne trajektorie. Brakowało w tym modelu układu automatycznie identyfiku-

jącego obrazy na ekranie. Pierwszy kompletny układ rozpoznający przedstawili w roku 1952 DAVIS, BIDDULPH i BALASHEK z Bell Telephone Laboratories. W ich modelu sygnał mowy wyrażono dwoma parametrami będącymi częstościami przejść przez zero w pasmach powyżej i poniżej 900 Hz. Obraz sygnału mowy utworzony z tych parametrów był więc jedynie dwuwymiarowy. Identyfikacja następowała poprzez określenie najwyższej korelacji skrośnej identyfikowanego obrazu z obrazami uprzednio wyznaczonymi dla cyfr od 0 do 9. Poprawność rozpoznawania wynosiła dla jednego mówcy 97 %. Uważa się, że model ten był pierwszym zależnym od głosu układem rozpoznawania mówionych cyfr, w którym sygnał mowy potraktowano z akustycznego punktu widzenia i w którym posłużono się ideą porównywania obrazów. W roku 1958 DUDLEY i BALASHEK skonstruowali układ rozpoznający, który operował cechami widmowymi ekstrahowanymi z sygnału mowy przy użyciu 10-kanałowego analizatora widma. W modelu tym, jak również w opublikowanej w tym samym czasie pracy FRY'a i DENES'a, wprowadzono segmentację wyrazu na jednostki fonetyczne, które identyfikowano na podstawie ich obrazów widmowych. Dobre rezultaty rozpoznawania uzyskiwano jedynie w zakresie jednego głosu. W roku 1960 DENES i MATHEWS wprowadzili pojęcie normalizacji czasowej. Pierwsze próby rozpoznawania mowy przy użyciu techniki komputerowej miały miejsce już w latach 1959 i 1960. W roku 1959 J.W. i C.D. FORGIE z Laboratorium Lincolna rozpoznawali programowo samogłoski angielskie w wyrazach typu /bVt/ na podstawie położenia dwóch pierwszych formantów uzyskując poprawność 93 %. W roku 1962 ci sami autorzy opracowali program rozpoznawania spółgłosek trących na początku i końcu izolowanych wyrazów angielskich. Wśród pierwszych, którzy użyli maszyn cyfrowych do rozpoznawania mówionych wyrazów angielskich wymienia się także HUGHES'a (1961), MARTINA i innych (1964) i REDDY'ego (1967). Rozpoznawanie komputerowe było początkowo bardzo kosztowne. Opierało się ono na dawnych koncepcjach analizy widmowej i identyfikacji i nie imponowało szybkością działania. W latach 60-tych zaczęły pojawiać się też pierwsze hardware'owe wersje urządzeń rozpoznających wyrazy przeznaczone do specjalnych celów (DERESCH, 1961, TEACHER i inni, 1967, ROS, 1967, KELLY i inni, 1968, HILL, 1969 i MARTIN, 1969). Rozpoznawaniem mowy

zajmowano się w tej dekadzie głównie w Stanach Zjednoczonych i w niewielkim stopniu także w Japonii, Związku Radzieckim i w Niemczech. Pod koniec lat sześćdziesiątych problematyka automatycznego rozpoznawania mowy zaczyna szerzej upowszechniać się w świecie. W jej nurt włącza się coraz więcej ośrodków badawczych w różnych krajach. Wpłynęło na to wiele przyczyn. Pierwszą i chyba najistotniejszą z nich był dynamiczny rozwój techniki komputerowej. Laboratoriom naukowo-badawczym przybył komputer, nowe, atrakcyjne narzędzie pozwalające na modelowanie różnych koncepcji automatycznego rozpoznawania mowy bez potrzeby konstruowania w tym celu specjalnych układów, jak miało to miejsce dotychczas. Jednocześnie w miarę upowszechniania się techniki komputerowej ożywiać zaczęła się idea stworzenia tak zwanego wejścia fonicznego (WEF-u) umożliwiającego kontaktowanie się człowieka z komputerem za pomocą głosu i ewentualne sterowanie głosem poprzez komputer różnymi procesami. Dążenia tego rodzaju wynikały z postępujących tendencji do pełnej automatyzacji wszelkich procesów technologicznych. Z powodu specyficznych cech fonetyczno-akustycznych każdego języka prace nad automatycznym rozpoznawaniem mowy zaczęły rozwijać się równolegle w wielu krajach. Podejmowanie własnych badań w tym zakresie przez różne ośrodki naukowe wynikało też z faktu, iż nie istniała wówczas jeszcze żadna wypróbowana metoda automatycznego rozpoznawania mowy i problem pozostawał szeroko otwarty. Główne osiągnięcia z tej fali badań, nieprzerwanie zresztą rozwijających się aż po dzień dzisiejszy, zostaną podane w następnych częściach niniejszej pracy.

2.2. Ograniczenia zakresu rozpoznawania mowy

Rozwiązywanie problemu automatycznego rozpoznawania mowy przebiega stopniowo i na każdym etapie podlega określonym ograniczeniom. Problem ten jest bowiem bardzo złożony i trudny. Osiągnięcie przez automat takiej zdolności rozpoznawania mowy, jaką posiada człowiek, wydaje się być celem jeszcze wciąż bardzo odległym. Znakomita większość prac poświęconych automatycznemu rozpoznawaniu mowy dotyczy jedynie rozpoznawania izolowanych wyrazów. Opublikowano też już szereg prac poświęconych rozpoznawaniu fraz będących ciągiem kilku połączonych wyrazów

(VINCJUK, 1971, BRIDLE i BROWN, 1979, SAKOE, 1979, FLANAGAN i inni, 1980) . Zastosowane w nich metody rozpoznawania odwołują się na ogół do technik rozpoznawania wyrazów izolowanych. Pod pojęciem wyraz izolowany rozumie się wypowiedź pojedynczego wyrazu, w której otoczeniu panuje cisza.

Kolejne ograniczenie dotyczy rozmiarów słownika, którego wyrazy podlegają automatycznemu rozpoznawaniu. Słowniki takie bywają różnej wielkości i zawierają od dziesięciu do kilkuset wyrazów. O wielkości słownika decyduje w dużym stopniu to, czy układ rozpoznający jest zależny, czy niezależny od głosu, innymi słowy czy jest rzeczą obojętną, czyje wypowiedzi mają być automatycznie rozpoznawane.

Z ograniczeniem rozpoznawania pod względem liczby wyrazów wiąże się zatem ograniczenie co do ilości głosów, dla których oczekiwać można poprawnych wyników rozpoznawania. Dla małych słowników udaje się na ogół rozpoznawanie niezależne od głosu. Niekiedy za cenę tej niezależności przyjmuje się celowo słownik niewielkich rozmiarów. Przykład takiego podejścia zawarty jest w pracy FLANAGANA i innych (1980) . Problem wrażliwości układu rozpoznawania mowy na cechy osobnicze głosu jest dość złożony i trudny. Mimo, iż niektórzy badacze (JASCHUL, 1979, 1981, 1983) podejmują próby normalizacji indywidualnych cech akustycznych głosu, to jednak ogólnie rzecz biorąc w zagadnieniach automatycznego rozpoznawania mowy problem ten stawiany jest na dalszym planie lub traktowany bywa w sposób uboczny. Model rozpoznawania mowy weryfikuje się zwykle wpierw dla jednego głosu uzyskując na ogół dobre rezultaty, a dopiero potem podejmuje się próby jego działania dla głosów różnych. Próby te zwykle nie przynoszą zadowalających wyników. Metody rozpoznawania wyrazów gwarantujące niezależność wyników od głosu stosowane w przypadkach małych i prostych słowników nie dają się zastosować dla dowolnych słowników.

O wielkości słownika wyrazów automatycznie rozpoznawanych decyduje też forma, w jakiej wyrazy te są reprezentowane w pamięci komputera. Spotyka się dwa podstawowe rodzaje tej formy. Pierwsza dotyczy takich metod rozpoznawania, w których wyraz wymówiony rozpatruje się jako ciąg elementów należących do określonych wcześniej klas. Liczba klas zależy od zastosowanej

definicji elementu i wynikających z niej zasad segmentacji. Wielu badaczy kierowało się dążeniem do znalezienia takiego podziału sygnału mowy na elementy, przy którym uzyska się najmniejszą liczbę klas. Wiedza fonetyczno-akustyczna sugeruje opłacalność zastosowania podziału mowy na fonemy, bowiem ich liczba jest stosunkowo nieduża. Ta sugestia wydawała się początkowo najwłaściwszą. Słownik rozpoznawanych wyrazów umieszczony w pamięci komputera byłby wówczas analogiem słownika napisanego w transkrypcji fonetycznej. Realizacja takiej koncepcji okazała się jednak z kilku względów wcale niełatwa. Pierwszą istotną trudność przedstawia wyznaczanie granic międzyfonemowych. Następujące po sobie fonemy nie zawsze dzieli ostra granica. Przejścia między niektórymi fonemami są bardzo łagodne, to znaczy, że szybkości zmian parametrów sygnału mowy są wówczas nie tylko skończone, ale zbyt małe, by zdyskryminować występującą granicę. FANT (1973) wskazał na rozbieżność pomiędzy fonemem a jego akustyczną lub artykulacyjną reprezentacją dowodząc, że w każdej wypowiedzi wyróżnić można więcej segmentów akustycznych niż fonemów. Trudność w realizacji klasyfikacji fonematycznej powodują też zróżnicowania osobnicze i kontekstowe obrazów akustycznych poszczególnych fonemów. Wartości parametrów danego fonemu są zależne od jego otoczenia oraz od głosu mówiącego.

Fonem w mowie ciągłej należy zatem traktować jako zjawisko niestacjonarne z rozmytymi granicami. W takim też rozumieniu należałoby go klasyfikować.

RUSKE i SCHOTOLA (1980) twierdzą, że łatwiej jest określać położenie środka fonemu niż jego granice, oraz że klasyfikacji należy poddawać nie fonemy, a tzw. półsylaby (demi-syllables), za które uważa się segmenty mowy ciągłej pomiędzy środkami następujących po sobie samogłosek i spółgłosek lub vice versa. Jeśli jako element podziału mowy przyjąć zamiast fonemu półsylabę, liczba klas staje się dużo większa. Klasy tworzą wówczas poszczególne połączenia międzyfonemowe (diphones) występujące w danym języku. Elementy tego rodzaju klasy różnić się mogą długością stanów ustalonych. Segmentacja półsylabiczna mowy jest faktycznie prostsza od segmentacji fonematycznej. W półsylabach zawarte są cechy kontekstowe fonemów komplikujące klasyfikację fonematyczną w zastosowaniu do rozpoznawania mowy.

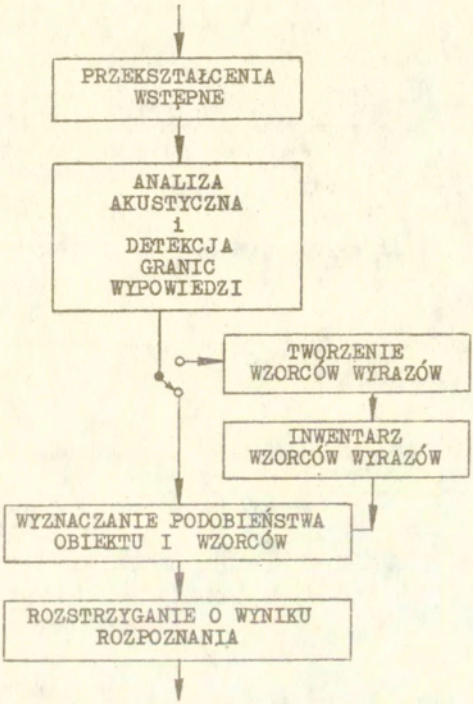
J. SHOUP (1980) przedstawiła korzyści i niedogodności związane z posługiwaniem się w automatycznym rozpoznawaniu mowy różnego rodzaju jednostkami fonologicznymi, do których zaliczyła allofony, fonemy, difony, sylaby i wyrazy. Z jej zestawienia wynika, że operowanie w automatycznym rozpoznawaniu mowy którąkolwiek z tych jednostek ma swoje wady i zalety. SHOUP stwierdziła też, że fonem identyfikuje się najtrudniej. Podobne oceny, lecz z nieco innymi proporcjami wad i zalet przedstawił G. MERCIER (1981).

2.3. Uwagi ogólne o globalnym rozpoznawaniu mowy

Problemy, jakie stwarza segmentacja mowy oraz klasyfikacja i identyfikacja jednostek segmentalnych skłoniły wielu badaczy zajmujących się automatycznym rozpoznawaniem mowy do traktowania wypowiedzi wyrazu w sposób globalny, tzn. bez dokonywania w niej podziału na elementy należące do uprzednio określonych klas. Pożądana jest w takiej metodzie jedynie znajomość położenia początku i końca wypowiedzi, chociaż w niektórych wersjach globalnego rozpoznawania nie jest to bezwzględnie konieczne (BRIDLE i inni, 1981). Wyznaczenie granic wyrazu nie nastręcza żadnych trudności, jeżeli został on wymówiony w izolacji i przy braku hałasów. Stąd globalne rozpoznawanie stosuje się głównie do wyrazów izolowanych.

Wśród metod globalnego rozpoznawania na uwagę zasługują jedynie takie, w których nie istnieją żadne uzależnienia proceduralne od struktur i ilości rozpoznawanych wyrazów. Wszystkie spotykane metody globalnego rozpoznawania wyrazów łączy pewien wspólny schemat działania, który przedstawiono na rys. 1. Różnice pomiędzy indywidualnymi metodami dotyczą sposobów realizacji poszczególnych etapów procesu rozpoznawania oraz formy reprezentacji wyrazu. Warianty tej realizacji oraz różne formy reprezentacji wyrazu zostaną omówione poniżej.

W odróżnieniu od metod segmentalnych, metody globalne automatycznego rozpoznawania wyrazów nie wymagają adaptacji językowej, tzn. przystosowania do określonego języka. Żadna z metod globalnego rozpoznawania wyrazów nie jest ściśle przypisana do jakiegoś języka. Pod tym względem metody globalne mają charakter uniwersalny i przewyższają metody segmentalne.



Rys. 1. Ogólny schemat systemu globalnego rozpoznawania wyrazów

2.4. Analiza akustyczna - pierwszy etap globalnego rozpoznawania wyrazów

Część pierwszą ogólnego modelu globalnego rozpoznawania wyrazów stanowi analiza akustyczna. Zadaniem jej jest wyekstrahowanie z sygnału mowy pewnych istotnych parametrów i jednocześnie odrzucenie informacji mało znaczących i przez to zbytecznych. Stosowane są w tym celu różnego rodzaju analizatory akustyczne, przeważnie w wersjach cyfrowych, chociaż używane są także analizatory analogowe oraz analogowo-cyfrowe.

W większości systemów rozpoznających stosowana jest analiza widmowa. Wykonuje się ją za pomocą zespołu filtrów środkowo-przepustowych o jednakowych lub zróżnicowanych szerokościach

pasm analizy. Brak jest zgodności w opiniach w kwestii doboru właściwego rodzaju analizatora widmowego. Stosowane są np. analizatory z podziałem tercjowo-oktawowym, analizatory ze skalą mel, analizatory będące modelem funkcyjnym ucha środkowego i inne. Na uwagę zasługuje analizator widma przedstawiony przez ZWICKERA i innych (1979) i używany przez RUSKE w badaniach nad rozpoznawaniem mowy na Uniwersytecie Technicznym w Monachium. W analizatorze tym pełen zakres częstotliwości słyszalnych podzielony jest na 24 pasma o jednakowej szerokości subiektywnej równej 1 barkowi. Mierzoną w każdym paśmie wielkością jest głośność uważana za miarę psycho-akustycznych wrażeń intensywności dźwięku. Analizator ten uwzględnia także właściwości percepcyjne ucha w zakresie reagowania na wahania ciśnienia akustycznego.

W niektórych systemach dysponujących dużą mocą obliczeniową widmo obliczane jest metodą szybkiej transformacji Fourier'a. Szybkie wykonanie operacji składających się na wyznaczanie widma umożliwia moduł typu „butterfly" zawierający 3 niezależne układy trzech podstawowych działań przekształcenia Fourier'a: mnożenia, sumowania i odejmowania liczb zespolonych. Moduł „butterfly" produkowany jest obecnie w układzie zintegrowanym według technologii bipolarnej. Osiąganie dużych zdolności obliczeniowych w systemach analizy akustycznej sygnału mowy umożliwiają także szybkie akumulatory mnożąco-sumujące produkowane już obecnie masowo w formie modułowej (ALLEN, 1981).

Wśród wielu różnych metod analizy akustycznej służących parametryzacji sygnału mowy w systemach automatycznego rozpoznawania wyrazów szczególne uznanie zyskały sobie metoda predykcji liniowej oraz metoda cepstralna.

Jak podają M. i G. SORENSON (1970), a za nimi MARKEL i GRAY (1976) predykcji liniowej dała początek stworzona przez GAUSSA w roku 1975 metoda liniowej oceny za pomocą najmniejszych kwadratów. Jako pierwszy użył pojęcia „predykcja liniowa" w roku 1949 WIENER. SAITO i ITAKURĘ (1966) oraz ATALA i SCHRÖDERA (1967) uważa się za pierwszych, którzy zastosowali metodę predykcji liniowej do analizy i syntezy mowy. Obszerne przedstawienie teorii i zastosowań predykcji liniowej w analizie, syntezie i rozpoznawaniu mowy zawierają prace MARKELA i GRAY'A

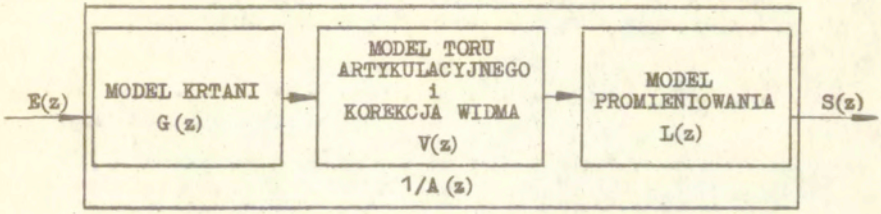
(1976) oraz MAKHOULA (1975 a i 1975 b).

Analiza metodą predykcji liniowej przyjmuje jako model wytwarzania mowy biegunowy filtr cyfrowy pobudzany periodycznie impulsami jednostkowymi dla dźwięcznych fragmentów mowy lub szumem przypadkowym dla fragmentów bezdźwięcznych. Filtr ten wywodzi się z modelu wytwarzania mowy podanego przez FANTA w roku 1960.

Określenie biegunowy oznacza, że funkcja przeniesienia tego filtru wyrażona wzorem:

$$H(z) = \frac{G}{1 + \sum_{i=1}^{M} a_i z^{-i}} \qquad (2.1)$$

posiada jedynie bieguny (nie ma zer).



Rys. 2. Model wytwarzania mowy przyjmowany w predykcji liniowej

Model wytwarzania mowy wyrażony wzorem (2.1) reprezentuje w domenie czasowej zależność:

$$x_n = -\sum_{i=1}^{M} a_i x_{n-1} + e_n \qquad (2.2)$$

Ponieważ e_n wyraża próbkę sygnału z niewiadomego źródła pobudzającego, zalicza się ten wyraz do błędu, z jakim liniowo ważona suma M-1 wcześniejszych próbek $x_{n-1}, \ldots, x_{n-M}$ analizowanego sygnału rokuje o wartości bieżącej próbki x_n. Analiza predykcyjna polega na wyliczeniu współczynników wagowych a_i

będących jednocześnie współczynnikami filtru biegunowego, który jest założonym modelem wytwarzania mowy. Przyjmując zasadę, że właściwa wartość każdego ze współczynników a_i minimalizuje ogólny błąd predykcji zdefiniowany jako suma kwadratów błędów chwilowych e_n na przestrzeni pewnego przedziału czasu, wyznaczenie współczynników filtru a_i redukuje się do rozwiązania układu M równań liniowych:

$$\sum_{i=1}^{M} a_i c_{ij} = - c_{oj} \qquad (2.3)$$

gdzie M jest rzędem filtru biegunowego, a

$$c_{ij} = \sum_{n=n_o}^{n_1} x_{n-i} \cdot x_{n-j} \qquad (2.4)$$

dla $j = 1,2,\ldots,M$. n_o i n_1 oznaczają granice przedziału, na przestrzeni którego dokonuje się minimalizacja błędu. Istnieją dwie metody obliczania współczynników predykcyjnych a_i - metoda autokorelacji oparta o założenie, że n_o i n_i przypadają odpowiednio w $-\infty$ i $+\infty$ oraz metoda kowariancji zakładająca $n_o = 0$ i $n_1 = N-1$.

Współczynniki filtru biegunowego a_i zwane też współczynnikami predykcji liniowej służą często jako parametry reprezentujące sygnał mowy w systemach rozpoznawania mowy. Szereg badaczy wyróżnia metodę predykcji liniowej spośród innych metod analizy akustycznej sygnału mowy. Np. ZUE (1980) wymienia dwie zalety korzystnie odróżniające metodę predykcji liniowej od klasycznej analizy widmowej sygnału mowy. Pierwszą jest to, że uwalnia ona widmo od efektów harmonicznego charakteru dźwięków mowy. Drugą jest duża zgodność położenia wierzchołków funkcji przeniesienia filtru liniowo predykcyjnego z położeniem formantów, jeśli wystarczająco wysoki jest rząd predyktora. Dzięki tej drugiej zalecie można za pomocą predykcji liniowej wyznaczać przebiegi częstotliwości formantów w mowie.

DAVIS i MERMELSTEIN (1980) natomiast udowadniają, że do rozpoznawania mowy korzystniejszą zarówno od analizy predykcyjnej, jak i od zwykłej analizy widmowej, jest analiza cepstralna.

Zespolone cepstrum ciągu próbek x(n) jest definiowane jako ciąg $\hat{x}(n)$ będący odwrotną transformatą Fouriera zlogarytmowanej transformaty Fouriera ciągu x(n). Zapis matematyczny tej definicji jest następujący:

$$X(k) = \sum_{n=0}^{N-1} x(n) \exp\left(-j \frac{2\pi}{N} \cdot kn\right) \qquad (2.5)$$

$$\hat{X}(k) = \mathrm{Log}\left(X(k)\right) \qquad (2.6)$$

$$\hat{x}(n) = \frac{1}{N} \sum_{k=1}^{N-1} \hat{X}(k) \exp\left(j \frac{2\pi}{N} \cdot kn\right) \qquad (2.7)$$

Jeśli przyjąć, że ciągi x(n) są minimalno-fazowe, czyli że zera i bieguny położone są wewnątrz koła jednostkowego, cepstrum zespolone sprowadza się do cepstrum rzeczywistego (OPPENHEIM, SCHAFER, 1968), które jest wówczas odwrotną transformatą Fouriera logarytmu modułu transformaty ciągu x(n). DAVIS i MERMELSTEIN (1980) zaproponowali przekształcenie cepstralne z widma w skali MEL, jako bardziej odpowiadającej kryteriom percepcyjnym. GAGNOULET i inni (1983) potwierdzili słuszność propozycji DAVISA i MERMELSTEINA przytaczając korzystne wyniki globalnego rozpoznawania trudnych wyrazów francuskich z zastosowaniem reprezentacji cepstralnej i MEL-owej skali częstotliwości. Te same wyrazy były gorzej rozpoznawane przy użyciu dwóch innych rodzajów analizy akustycznej, a mianowicie analizy za pomocą 12-kanałowego wokodera oraz analizy cepstralnej poprzez predykcję liniową i z zastosowaniem liniowej skali częstotliwości. Stosując do przekształcenia cepstralnego MEL-ową skalę częstotliwości można wystarczająco reprezentatywnie dla potrzeb rozpoznawania przedstawić sygnał mowy za pomocą zaledwie 6-8 parametrów cepstralnych.

Jeśli etap analizy akustycznej realizowany jest w całości przez komputer, wówczas poprzedza go filtracja dolno-przepustowa i konwersja analogowo-cyfrowa sygnału mowy, a następnie wpis próbek do pamięci komputera. Kod cyfrowy próbki w spotykanych analizatorach akustycznych używanych do rozpoznawania mowy ma wymiar od 8 do 12 bitów. Częstość próbkowania mieści się zwykle w granicach od 10 do 15 kHz, zaś o połowę niższą od niej przyjmuje się z wiadomych względów granicę filtracji dolno-przepustowej. Jeżeli natomiast analizę akustyczną wykonują układy analogowe, wówczas kodowanie cyfrowe i wpis do pamięci komputera dotyczy dopiero wyników tej analizy, a są nimi przeważnie określonego rodzaju parametry widmowe stanowiące reprezentację sygnału mowy w pewnym przedziale czasu. Ten przedział czasu wynoszący od 10 do kilkunastu milisekund decyduje o częstości operacji wpisu do pamięci komputera kolejnych prób kilkunastu lub kilkudziesięciu parametrów będących wynikiem analizy akustycznej sygnału mowy. Wpis jednej próby reprezentującej pewne stadium chwilowe sygnału przebiega w miarę technicznych możliwości jak najszybciej. Próba taka ma w języku angielskim nazwę „frame". Zwykle podczas analizy akustycznej następuje równocześnie identyfikacja początku i końca wyrazu i tym samym określenie jego rozciągłości czasowej.

Wśród badaczy panuje niemal zgodne przekonanie, że najwłaściwiej reprezentują sygnał mowy parametry częstotliwościowe i nimi też przeważnie operują spotykane metody globalnego rozpoznawania izolowanych wyrazów. Mimo to jednak nie brak przykładów tworzenia systemów rozpoznawania opartych o kodowanie czasowe sygnału mowy. Zacytować tu można prace NIEDERJOHNA (1975) oraz BAUDRY'ego i DUPEYRAT'a (1982). Np. w modelu rozpoznawania wyrazów przedstawionym przez tych ostatnich dwóch autorów parametrami reprezentującymi sygnał mowy są liczebności odstępów pomiędzy kolejnymi zerami pochodnej funkcji sygnału mowy w 16 klasach długości czasowej i w zakresie pewnego okna czasowego, powiększone odpowiednio o składnik proporcjonalny do długości czasowej określającej każdą klasę. Taki rodzaj reprezentacji sygnału mowy jest bardzo prosty w realizacji, co niewątpliwie stanowi jego dużą zaletę. Mimo to jednak parametryzacja sygnału mowy w dziedzinie czasu stosowana jest w rozpoznawaniu mowy bar-

dzo rzadko.

Analiza akustyczna w systemie rozpoznawania mowy przebiegać może w sposób ciągły, co ma miejsce w przypadku stosowania analizatorów analogowych, lub dyskretnie, jeśli analizator zmodelowany jest w maszynie cyfrowej lub wykonany w wersji cyfrowej. Pierwszy wariant spotykany jest już coraz rzadziej. Z potoku danych ciągle napływających z analizatora analogowego, do operacji rozpoznawania wystarczają jedynie próby reprezentujące sygnał mowy w kolejnych momentach czasu oddalonych od siebie o skończoną odległość czasową. Próbę tworzą chwilowe wartości parametrów, których liczba, rodzaj i zakres wynikają z typu analizatora analogowego. W przypadku, gdy analizę akustyczną wykonuje wyspecjalizowany układ cyfrowy lub standardowy komputer, reprezentacja sygnału mowy w formie ciągu prób charakteryzujących tenże sygnał jedynie w kolejnych momentach odległych od siebie o skończony przedział czasu wynika niejako naturalnie z faktu, że obliczenie wartości parametrów składających się na jedną próbę wymaga pewnego czasu. Stosowane długości odstępu czasowego dzielącego kolejne próby są dość zróżnicowane i wynoszą od kilku do około 20 ms.

2.5. Pojęcie obrazu akustycznego

Wynikiem analizy akustycznej wypowiedzi w systemach globalnego rozpoznawania wyrazów jest zatem macierz wartości parametrów reprezentujących sygnał mowy w kolejnych przedziałach czasu. Określa się tę macierz w terminologii angielskiej przeważnie terminem PATTERN. Brak jest dotychczas w specjalistycznym słownictwie polskim usankcjonowanego odpowiednika tego określenia angielskiego. Unika się raczej stosowania w tym znaczeniu polskiego wyrazu „wzór", będącego leksykalnym odpowiednikiem angielskiego „pattern". Próbuje się używać wyrazu „obraz" w znaczeniu, jakie w rozpoznawaniu mowy ma angielskie „pattern". Dla ścisłości należałoby dodawać „akustyczny" dla odróżnienia od podstawowego znaczenia, jakie ma ten wyraz w języku polskim. Trafne i wygodne w użyciu mogłyby też być określenia „akustobraz" lub „mowobraz". W dalszej części niniejszej pracy stosowane będzie określenie „obraz akustyczny" lub krótko „obraz". W niektórych pracach angielskojęzycznych, np. RABI-

NERA (1978), spotyka się określenie TEMPLATE zamiast PATTERN. Słowo TEMPLATE nie ma w ogóle polskiego odpowiednika leksykalnego. W słownictwie polskim brakuje także trafnego odpowiednika wyrazu FRAME oznaczającego między innymi to, co powyżej określono przez PRÓBA. Słowo PRÓBA użyte w znaczeniu jak wyżej wydaje się być określeniem mało precyzyjnym, gdyż posiada zbyt szerokie pole semantyczne. Brak niestety dotychczas polskiego terminu trafniej określającego zbiór wartości parametrów charakteryzujących bardzo wąski segment sygnału mowy.

2.6. Porównywanie obrazów akustycznych

Po parametryzacji sygnału mowy i wyznaczeniu początku oraz końca wypowiedzi, kolejnym etapem w procesie globalnego rozpoznawania wyrazów izolowanych jest z reguły porównywanie obrazów akustycznych. To działanie wykonuje się także na etapie UCZENIA lub ADAPTACJI, podczas którego następuje przygotowanie systemu rozpoznającego do późniejszego rozpoznawania wyrazów wchodzących w skład założonego słownika. Problemowi globalnego porównywania dwóch obrazów akustycznych poświęcono bardzo wiele prac, w których przedstawiono różne szczegółowe rozwiązania. Każde rozwiązanie odnosi się zasadniczo do dwóch podstawowych zagadnień, a mianowicie wyboru właściwej miary odległości oraz uwzględnienia różnic w rozkładzie czasowym odpowiadających sobie fragmentów porównywanych obrazów. Na globalne podobieństwo lub odległość składają się podobieństwa i odległości odpowiadających sobie segmentów porównywanych wyrazów. Reprezentację segmentu stanowi to, co wyżej nazwano próbą, czyli zbiór wartości parametrów charakteryzujących sygnał mowy w wąskim przedziale czasu, nazywany też nieraz wektorem cech. Wobec tego w grę wchodzą dwie miary odległości. Pierwsza wyraża podobieństwo lokalne porównywanych obrazów akustycznych czyli podobieństwo ich segmentów. Druga miara odległości wyraża podobieństwo całych obrazów akustycznych.

2.6.1. Wyznaczanie podobieństwa lokalnego

Wyznaczanie podobieństw lokalnych porównywanych obrazów A i B polega na obliczeniu odległości pomiędzy poszczególnymi segmentami A_x i B_y obu tych obrazów. W systemach globalne-

go rozpoznawania wyrazów, w których analizę akustyczną wykonują wielopasmowe analizatory widma, segment taki reprezentują wartości energii sygnału mowy w pasmach analizy. Taką reprezentację segmentu stosowali w swoich pracach nad rozpoznawaniem izolowanych wyrazów między innymi SAKOE i CHIBA (1978), DAS (1982) oraz LAMEL i inni (1982). Autorzy ci stosowali różne miary lokalnej odległości, co świadczyć może o dopuszczalnej dowolności w wyborze takiej miary dla rozpoznawania mowy. Np. SAKOE i CHIBA (1978) użyli jako miarę odległości porównywanych segmentów moduł różnicy wektorów cech A_x i B_y reprezentujących te segmenty:

$$d(x,y) = ||A_x - B_y|| \tag{2.8}$$

DAS (1982) przyjął jako miarę odległości sumę bezwzględnych różnic współrzędnych wektorów cech

$$d(x,y) = |A_x - B_y| = \sum_{i=1}^{n} |a_{xi} - b_{yi}| \tag{2.9}$$

Indeksy x i y są odpowiednio numerami kolejnymi wektorów cech lub fragmentów obrazów A i B. Indeks i odnosi się do numeru cechy (współrzędnej wektora). Np. a_{xi} oznacza cechę i wektora A_x. Inną miarę odległości pomiędzy wektorami cech użyli LAMEL i ZUE (1982). Wyraża się ona wzorem:

$$d(x,y) = \log \frac{\sum_{i=1}^{n} (a_{xi} b_{yi})}{\left(\sum_{i=1}^{n} a_{xi}^2\right)^{\frac{1}{2}} \left(\sum_{i=1}^{n} b_{yi}^2\right)^{\frac{1}{2}}} \tag{2.10}$$

We wszystkich przytoczonych wzorach użyto jednakowego oznaczenia cech, wektorów cech i odległości w celu ułatwienia dostrzeżenia różnic między wyrażonymi przez te wzory miarami odległości. W pracach, z których te wzory pochodzą, użyto innych oznaczeń.

W ostatnich latach bardzo często używa się współczynników predykcji liniowej jako parametrów reprezentujących sygnał mo-

wy w procesach automatycznego rozpoznawania mowy. Metoda predykcji liniowej odegrała doniosłą rolę w rozwoju automatycznego rozpoznawania wyrazów. Dla oceny lokalnego podobieństwa dwóch obrazów akustycznych wyrażanych przez współczynniki predykcji liniowej ITAKURA (1975) zaproponował bardzo korzystną miarę odległości opartą o pewne założone właściwości statystyczne zbiorów parametrów predykcji liniowej. Miara Itakury określa, czy segment sygnału mowy X wyrażony N kolejnymi próbkami wartości chwilowych i zbiorem â współczynników predykcji liniowej uważać można za podobny do sygnału ukształtowanego przez model filtru biegunowego reprezentowany zbiorem a współczynników. Kanoniczną postacią miary odległości Itakury jest wzór:

$$d(x/a) = \log\left(\frac{aVa^t}{\hat{a}V\hat{a}^t}\right) \qquad (2.11)$$

w którym symbolem d(x/a) oznaczono odległość Itakury, a^t i $\hat{a}^t$ oznaczają odpowiednio macierze transponowane wektorów a i â, a V jest macierzą autokorelacji segmentu X sygnału mowy wyrażanego też zbiorem â współczynników predykcyjnych. Itakura proponuje jako wygodniejszy do obliczeń wzór na odległość w następującej postaci:

$$d(x/a) = c + \log\left[(br)/(\hat{a}r)\right] \qquad (2.12)$$

gdzie c = log(aa), wszystkie iloczyny typu (XY) są iloczynami wewnętrznymi (INNER PRODUCT), b jest wektorem [1, b(1), b(2),... .., b(p)], którego współrzędne wylicza się ze wzoru:

$$b(i) = 2\sum_{j=0}^{p-1} a(j)a(j+1)/(aa) \qquad (2.13)$$

r jest znormalizowanym wektorem korelacji $r = \left(v(i)/\, v(o)\right)$ $(i = 0, \ldots, p)$ przy czym $v(i) = \frac{1}{N}\sum_{n=1}^{N-1} x(n)x(n+1)$, p jest rzędem predykcji liniowej.

Miarę odległości Itakury uważać można za najczęściej stosowaną w metodach rozpoznawania mowy opartych na predykcji liniowej. Istnieją bowiem też inne miary odległości odnoszące się do rozpoznawania z zastosowaniem predykcji liniowej, np. zaproponowane przez GRAY'a i MARKEL'a (1976) lub SAMBURA i RABINERA (1976). Metody rozpoznawania mowy posługujące się predykcją liniową są niewątpliwie bardzo racjonalne z teoretycznego punktu widzenia. W realizacji są natomiast bardzo czasochłonne i drogie. Wymagają bardzo wydajnych maszyn liczących. Stąd na efektywne posługiwanie się nimi pozwolić sobie mogą jedynie zamożne i dobrze wyposażone laboratoria badawcze.

2.6.2. Wyznaczanie odległości globalnej porównywanych obrazów akustycznych

Oddzielny problem w porównywaniu obrazów akustycznych stanowi wyznaczanie tzw. odległości globalnej. Składają się na to następujące dwie przyczyny: Na ogół każdy z dwóch wzajemnie porównywanych obrazów akustycznych posiada inną liczbę prób, gdyż każda wypowiedź tego samego wyrazu ma inną rozciągłość czasową. W każdej wypowiedzi tego samego wyrazu występuje niepowtarzalny rozkład w czasie segmentów fonetyczno-akustycznych. We wczesnych próbach rozpoznawania mowy starano się tego rodzaju różnice czasowe uwzględnić stosując liniową normalizację czasową. WINCJUKA (1969), WJELICZKĘ i ZAGORUJKĘ (1970) oraz SAKOE i CHIBĘ (1971) uważać można za pierwszych, którzy zastosowali technikę dynamicznego programowania dla optymalnego uwzględnienia tych różnic podczas wyznaczania odległości globalnej porównywanych wypowiedzi.

Programowanie dynamiczne jest metodą nieliniowej normalizacji czasowej. Różnice w pomiarach czasowych porównywanych obrazów akustycznych zostają uwzględnione przez kształtowanie skali czasu jednego z nich tak, aby uzyskać optymalne skojarzenie odpowiadających sobie segmentów w obu wypowiedziach. Dysponując lokalnymi odległościami $d(c(k))$ odpowiednich segmentów porównywanych wyrazów akustycznych A i B wyznacza się odległość globalną między nimi stosując następujący wzór:

$$D(A,B) = \min_{F} \left[\frac{\sum_{k=1}^{K} d\left(c(k) \cdot w(k)\right)}{\sum_{k=1}^{K} w(k)} \right] \qquad (2.14)$$

F oznacza funkcję, według której następuje kojarzenie segmentów jednego i drugiego obrazu akustycznego. Funkcja F nazywa się w terminologii angielskiej Time Warping Function. Brak niestety dotychczas trafnego odpowiednika tej nazwy w terminologii polskiej. W niniejszej pracy stosowane będzie określenie funkcja normalizacji czasowej lub w skrócie NC. c(k) symbolizuje skojarzone fragmenty obu obrazów, a argument k wyraża numery kolejne poszczególnych skojarzeń fragmentu i(k) obrazu A i fragmentu j(k) obrazu B. c(k) wyrażane jest następującym zapisem: $c(k) = (i(k), j(k))$. w(k) we wzorze (2.14) jest dodatnim współczynnikiem wagowym, a suma jego wartości w obszarze funkcji F występująca w mianowniku tego wzoru ma na celu uniezależnienie obliczanej odległości globalnej od całkowitej liczby skojarzeń wynikającej z przebiegu funkcji normalizacji czasowej NC . Przytoczone wyżej zależności ilustruje rys. 3.

Funkcja normalizacji czasowej podlega kilku ograniczeniom wynikającym z podstawowych cech mowy. Z tych powodów spełniać ona musi następujące warunki podane między innymi przez SAKOE i CHIBĘ (1976):

1. Warunek monotoniczności, który wymaga, aby:

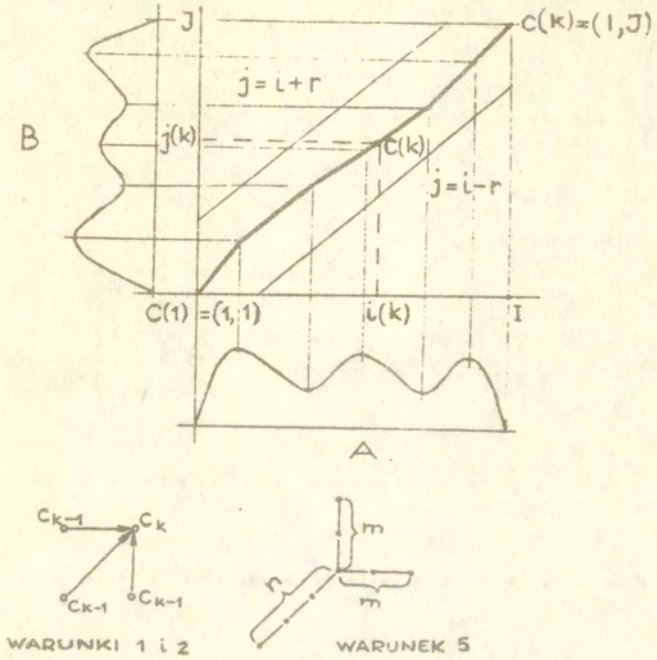
$$i(k-1) \leq i(k) \quad \text{oraz} \quad j(k-1) \leq j(k).$$

2. Warunek ciągłości wyrażony nierównościami:

$$i(k) - i(k-1) \leq 1 \quad \text{oraz} \quad j(k) - j(k-1) \leq 1.$$

Z warunków 1 i 2 wynika, iż punkt c(k-1) bezpośrednio poprzedzający punkt c(k) odnosić się może do jednej z trzech par fragmentów porównywanych wyrazów, co zapisać można następująco wyrażając te fragmenty ich indeksami:

$$c(k-1) = \left[(i(k), j(k)-1) \vee (i(k)-1, j(k)-1) \vee (i(k)-1, j(k)) \right].$$



Rys. 3. Ilustracja zasady dynamicznej nieliniowej normalizacji czasowej

3. Trzeci warunek tzw. brzegowy wymaga, aby pary pierwszych i ostatnich fragmentów obu obrazów stanowiły odpowiednio początek i koniec funkcji NC, czyli aby

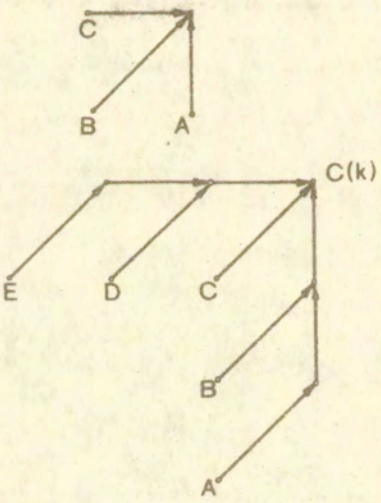
$$i(1) = 1, \quad j(1) = 1 \quad \text{oraz} \quad i(K) = I, \quad j(K) = J.$$

4. Różnice rozciągłości czasowych porównywanych obrazów mieszczą się w pewnym ograniczonym zakresie, z czego wynika korzystny warunek, iż

$$|i(k) - j(k)| \leq r \quad ,$$

gdzie r jest liczbą całkowitą dodatnią określającą strefę przebiegu funkcji NC.

5. Ostatni piąty warunek określa dopuszczalne nachylenie funkcji NC wyrażone stosunkiem liczb n i m. n jest liczbą kolejnych ruchów punktu c(k) w kierunku przekątnej, po których dopuszczalnych jest m kolejnych ruchów punktu c(k) w kierunku i lub j. Rozpatrywane były przez różnych badaczy możliwe trajektorie prowadzące do punktu c(k) z dozwolonych punktów wyjściowych (ITAKURA, 1976, SAKOE i CHIBA, 1978, MYERS i inni, 1981, OKOCHI i SAKAI, 1982). Każda taka trajektoria posiada przebieg zgodny z dopuszczalnym nachyleniem funkcji NC określonym przez stosunek n/m. Na rys. 4 podano przykłady różnych trajektorii i ruchów prowadzących do punktu c(k). Zaczerpnięto je z pracy SAKOE i CHIBY (1978).



Rys. 4. Przykłady trajektorii ruchów w kierunku punktu c(k).

Algorytm wyznaczania odległości globalnej oparty jest o zasadę programowania dynamicznego, która polega na poszukiwaniu takiego porządku doboru ważonych odległości lokalnych, aby osiągnąć minimum odległości globalnej. Do optymalnej odległości globalnej dochodzi się rozwiązując wielokrotnie równanie programowania dynamicznego:

$$g_k(c(k)) = \min_{c(k-1)} \left[g_{k-1}(c(k-1)) + d(c(k) \cdot w(k)) \right], \qquad (2.15)$$

w którym $g_k(c(k))$ oznacza odległość porównywanych obrazów w zakresie od początku, czyli od $c(1)$ do punktu $c(k)$. Składają się na tę odległość cząstkową odległość w zakresie do punktu $c(k-1)$ oraz ważona odległość lokalna w punkcie $c(k)$. Współczynnik wagowy przyjmuje się według jednej z dwóch następujących definicji:

1. $w(k) = (i(k) - i(k-1)) + (j(k) - j(k-1))$ (2.16)

dla tzw. formy symetrycznej,

2. $w(k) = (i(k) - i(k-1))$ (2.17)

dla formy niesymetrycznej.

Mianownik we wzorze (2.14) staje się równy 1+J w przypadku przyjęcia definicji pierwszej, a jest równy I dla definicji drugiej. Dla pierwszego zbioru trajektorii dozwolonych dojść do punktu $c(k)$ pokazanych na rys. 4 oraz przy uwzględnieniu pierwszej definicji dla współczynnika wagowego równanie programowania dynamicznego ma 3 następujące warianty:

$$g(i,j) = \min \begin{bmatrix} g(i,\ j-1) + d(i,j) \\ g(i-1,\ j-1) + 2d(i,j) \\ g(i-1,\ j) + d(i,j) \end{bmatrix} \qquad (2.18)$$

Poszczególne wiersze odnoszą się do trajektorii prowadzących z punktów A, B i C do punktu $c(k)$.

Dla drugiego zbioru trajektorii analogiczne równanie ma 5 wariantów:

$$g(i,j) = \min \begin{bmatrix} g(i-1,j-3) + 2d(i,j-2) + d(i,j-1) + d(i,j) \\ g(i-1,j-2) + 2d(i,j-1) + d(i,j) \\ g(i-1,j-1) + 2d(i,j) \\ g(i-2,j-1) + 2d(i-1,j) + d(i,j) \\ g(i-3,j-1) + 2d(i-2,j) + d(i-1,j) + d(i,j) \end{bmatrix} \quad (2.19)$$

Podobnie jak poprzednio, poszczególne wiersze dotyczą trajektorii wychodzących z punktów A, B, C, D, E i zmierzających do c(k). Najmniejsza z wartości g(I,J) stanowi rozwiązanie zadania znalezienia odległości globalnej pomiędzy porównywanymi obrazami A i B.

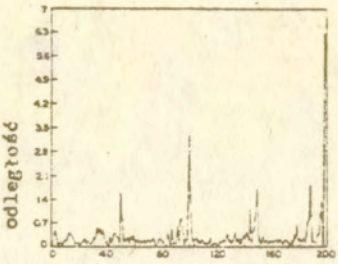
Technika programowania dynamicznego jest obecnie często stosowana w systemach automatycznego rozpoznawania wyrazów. Próbuje się ją ustawicznie udoskonalać modyfikując ją w celu zredukowania ilości koniecznych operacji bez uszczerbku dla wyników rozpoznawania.

2.7. Metody uczenia (adaptacji)

Rozpoznanie wypowiedzianego wyrazu następuje w wyniku porównania jego obrazu akustycznego z wzorcowymi obrazami akustycznymi wyrazów, do rozpoznania których układ rozpoznający został zaadaptowany. W terminologii angielskiej obraz akustyczny rozpoznawanego wyrazu nazywa się test pattern, a wzorcowy obraz akustyczny wyrazu - reference pattern lub reference template. W polskich pracach na temat automatycznego rozpoznawania wyrazów stosowane są już od kilku lat uproszczone nazwy obiekt i wzorzec zamiast obraz akustyczny rozpoznawanego wyrazu i wzorcowy obraz akustyczny wyrazu. Określenia obiekt i wzorzec będą też odtąd stosowane w niniejszej pracy. Wyznaczanie wzorców jest procesem bardzo ważnym, bowiem od jego wyniku zależy jakość rozpoznawania wyrazów. RABINER i WILPON (1980) uważają, że spotykane techniki adaptacji lub treningu można ogólnie podzielić na 3 klasy. Do pierwszej zaliczają się metody treningu przypadkowego, który przebiega następująco: Osoba, dla której ma być zaadaptowany system rozpoznający wypowiada każdy wyraz słownika pojedynczo lub kilkakrotnie i każda z tych wypowiedzi jest uważana za wzorzec. Ten rodzaj uczenia stosowany był przez

ITAKURĘ (1975) oraz ROSENBERGA i ITAKURĘ (1976). Druga klasa obejmuje metody z uśrednianiem, stosowane do tworzenia wzorców indywidualnych, jak i grupowych. Osoba lub grupa osób, gdy w grę wchodzi system rozpoznawania niezależny od głosu, wymawia każdy wyraz słownika wielokrotnie. Wzorzec powstaje w wyniku uśrednienia obrazów akustycznych poszczególnych wypowiedzi danego wyrazu. Rozciągłość czasową przyszłego wzorca przyjmuje się jako średnią z długości czasowych poszczególnych wypowiedzi tego wyrazu. Dla każdego wyrazu utworzony zostaje tylko jeden wzorzec. Metody z uśrednianiem osłabiają prawdopodobieństwo wpływu ewentualnych zakłóceń wypowiedzi na tworzony wzorzec. Nie gwarantują natomiast właściwego wzorca, gdy wypowiedzi składające się na wzorzec są bardzo różnymi realizacjami artykulacyjnymi tego samego wyrazu. Ten rodzaj treningu stosowali HERSHER i COX (1972), MARTIN (1976) oraz SAMBUR i RABINER (1976). W ostatnich latach użyto do tworzenia wzorców metod statystycznej klasteryzacji (statistical clustering methods). Wiodącą rolę miały w tym prace RABINERA (1978), LEVINSONA i innych (1979) oraz dwie prace RABINERA i WILPONA (1979). Obrazy akustyczne wypowiedzi łączone są w skupienia (clusters) na podstawie ich wzajemnego podobieństwa. Wzorcami wyrazów są centralne lub średnie obrazy akustyczne poszczególnych skupień. Dla pojedynczego wyrazu utworzonych może być kilka wzorców na podstawie wypowiedzi wielu głosów. LEVINSON i inni (1979) podają 4 procedury klasteryzacji. Pierwsza, zastosowana przez PATRICKA (1972) nazywana metodą łańcuchową (chainmap) polega na uszeregowaniu obrazów akustycznych różnych wypowiedzi tego samego wyrazu w takiej kolejności, aby każdy obraz był bardziej podobny do bezpośrednio poprzedzającego go niż do wszystkich następujących po nim. Na rys. 5 zamieszczono przykładowy wykres odległości $d_k = \delta(x_{k-1}, x_k)$ dla $1 \leq k \leq N-1$, kolejnych obrazów x_k, uszeregowanych w porządku łańcuchowym, od ich bezpośrednich poprzedników x_{k-1}. Wykres ten pochodzi z pracy LEVINSONA i innych (1979). Widać na nim, że dla pewnych obrazów odległości d_k są wydatnie większe niż dla pozostałych.

Każdy obraz znacznie oddalony od swego poprzednika zapoczątkowuje nowe skupienie obrazów. Obrazy podzielone zostają na tyle skupień, ile wydatnych maksimów posiada wykres.



Rys. 5. Wykres odległości sąsiednich obrazów w szeregu łańcuchowym

W innej metodzie skupia się obrazy akustyczne na zasadzie najbliższego sąsiedztwa (Shared Nearest Neighbors). Podstawę tej metody zastosowanej między innymi przez JARVISA i PATRICKA (1973) stanowi zasada, że dwie wypowiedzi mające przynajmniej pewną ilość k_s wspólnych sąsiadów należą do tego samego skupienia. Precyzyjniej zasadę tę można przedstawić następująco: Jeśli wypowiedź x_i ma k_i bliskich sąsiadów tworzących uporządkowany zbiór wypowiedzi R_i oraz wypowiedź x_j ma k_j bliskich sąsiadów składających się na zbiór R_j, oraz jeśli równocześnie $x_i \in R_j$ a $x_j \in R_i$ i zbiory R_i oraz R_j mają co najmniej k_s wspólnych elementów, czyli $|R_i \cap R_j| \geq k_s$, wówczas wypowiedzi x_i i x_j (a ściślej ich obrazy akustyczne) mają też co najmniej k_s wspólnych sąsiadów włączając siebie i tym samym należą do tego samego skupienia.

Do tworzenia wzorców wyrazu poprzez skupianie obrazów akustycznych różnych jego wypowiedzi stosowana jest też procedura k-krotnej iteracji. Składa się ona z trzech etapów. Pierwszym jest klasyfikacja obrazów według reguły najbliższego sąsiedztwa wyrażającej się zapisem:

$$x_j \in \omega_i, \text{ jeżeli } \delta(x_j, x_p^{(i)}) \leq \delta(x_j, x_p^{(k)}), \qquad 1 \leq k \leq M \qquad (2.20)$$

Obraz x_i zaliczony zostaje do klasy ω_i, jeżeli jego odległość do najbliższego obrazu $x_p^{(i)}$ tej klasy jest najmniejszą z wszystkich odległości dzielących go od innych bliskich jemu obrazów $x_p^{(k)}$ z poszczególnych innych klas. Na początku klasyfikacji przyjmuje się liczbę skupień M oraz typuje się w sposób dowolny M wypowiedzi jako tymczasowe środki przyszłych skupień. W drugim etapie następuje weryfikacja środków skupień według kryterium Minimax. W obrębie każdego skupienia poszukuje się takiego obrazu, który dzieli najmniejsza odległość od obrazu najbardziej oddalonego. Innymi słowy środkiem $x_p^{(i)}$ i-tego skupienia zostaje obraz $x_j^{(i)}$, od którego najbardziej oddalony element $x_k^{(i)}$ tej samej klasy dzieli najmniejsza odległość, $\min\left(\delta\left(x_j^{(i)}, x_k^{(i)}\right)\right)$. Kolejnym etapem metody k-krotnej iteracji jest test zbieżności, który polega na sprawdzeniu, czy lub nie środkami skupień są te same wypowiedzi, co w poprzednim kroku iteracji. Jeśli nie, iteracja jest kontynuowana.

Odmianą procedury k-krotnej iteracji jest metoda ISODATA (Iterative Self Organizing Data Analysis Technique A) BALL'a i HALL'a (1965), zastosowana przez LEVINSONA i innych (1979). Służy ona ogólnie rzecz ujmując do weryfikacji klas ze względu na ich liczbę oraz tzw. jakość klasyfikacji wyrażoną stosunkiem średniej odległości międzyklasowej i średniej odległości wewnątrzklasowej. Zasadniczym elementem procedury ISODATA jest klasteryzacja metodą k-krotnej iteracji, lecz po każdym kroku iteracji liczba skupień podlega weryfikacji. Jeśli aktualna liczba klas M przekracza dopuszczalną M_{max}, albo jeżeli liczebność $|\omega_i|$ i-tej klasy jest mniejsza od dopuszczalnego minimum m_{min}, lub też gdy odległość $\delta(x_p^{(i)}, x_p^{(j)})$ pomiędzy środkami i-tej i j-tej klasy jest mniejsza niż pewien próg θ_m, wówczas następuje łączenie klas. Natomiast, gdy aktualna liczba klas M jest mniejsza niż M_{min}, albo gdy liczebność $|\omega_i|$ którejś z klas jest większa od dopuszczalnej m_{max} lub jeśli jedna z klas jest znacznie rzadsza od pozostałych, wówczas procedura ISODATA umożliwia rozdzielanie klas.

Adaptacja w oparciu o techniki statystycznej klasteryzacji jest bardzo uciążliwa, gdy w treningu bierze udział tylko jeden głos np. przyszłego operatora systemu rozpoznawania wyrazów. Techniki te wymagają od 50 do 100-krotnej wypowiedzi każ-

dego wyrazu słownika. RABINER i WILPON (1980) zaproponowali metodę treningu, która zachowuje szereg zalet metody opartej o techniki statystycznej klasteryzacji, a jest jednocześnie dla mówiącego mniej forsowna. W metodzie tej każdy wyraz reprezentowany jest przez jeden wzorzec utworzony z dwóch najbardziej podobnych wypowiedzi. Podczas treningu mówiący wypowiada po kolei każdy wyraz słownika jednokrotnie. Gdy czyni to samo po raz drugi, wyznaczona zostaje dla każdego wyrazu odległość obrazu z poprzedniej i aktualnej wypowiedzi. Jeśli odległość ta dla danego wyrazu jest mniejsza od pewnego założonego progu, wówczas utworzony zostaje dla tego wyrazu wzorzec jako średni obraz z obu wypowiedzi. Jeśli nie, oczekiwana jest kolejna runda wypowiedzi już tylko tych wyrazów, dla których nie utworzono dotychczas wzorca. Nowa wypowiedź danego wyrazu może okazać się bardzo podobna do którejś z poprzednich. Jeśli mimo wielu replikacji brak jest pary podobnych wypowiedzi, wzorzec utworzony zostaje z dwóch najbardziej zbliżonych wypowiedzi.

RABINER i WILPON (1981) zaproponowali metodę rozpoznawania uwzględniającą znaczne podobieństwo niektórych wyrazów. Elementem tej metody jest procedura dyskryminacyjna. W procesie adaptacji dla każdej pary podobnych wyrazów wyznaczony zostaje oprócz wzorców szereg współczynników wagowych, które w procesie rozpoznawania ważą wpływ lokalnych podobieństw na podobieństwo globalne porównywanych obrazów. Współczynniki te zostają wyliczone dla poszczególnych segmentów dwóch podobnych wyrazów na podstawie różnic średnich odległości odpowiadających sobie segmentów pewnej liczby wypowiedzi tego samego wyrazu oraz dwóch wyrazów podobnych i z uwzględnieniem globalnej wariancji odległości odpowiadających sobie segmentów we wszystkich możliwych parach tych wypowiedzi.

Wynikiem adaptacji systemu jest zbiór wzorców wybranych wyrazów. Utworzone zostają wzorce indywidualne dla poszczególnych głosów, jak również wzorce wspólne dla wielu głosów. W pierwszym przypadku każdy wyraz reprezentowany jest zwykle przez jeden wzorzec, natomiast w drugim przez kilka. Wyrazy objęte adaptacją powinny być przez system poprawnie rozpoznawane.

2.8. Uzyskiwane wyniki automatycznego rozpoznawania wyrazów

Metodę rozpoznawania wyrazów oceniać należy biorąc pod uwagę takie względy, jak: koszt jej realizacji, rodzaj reprezentacji wyrazu, przystosowalność do ewentualnych zmian w zakresie słownika wyrazów oraz do nowych głosów. O wartości metody rozpoznawania wyrazów świadczą jednak przede wszystkim uzyskiwane wyniki rozpoznawania, które oceniać należy biorąc pod uwagę rozmiary słownika i stopień zależności od głosu mówiącego.

Poniżej przytoczono dane o wynikach globalnego rozpoznawania wyrazów zaczerpnięte z wybranych prac różnych autorów. Wśród osiągnięć w dziedzinie globalnego rozpoznawania wyrazów poczesne miejsce zajmuje model ITAKURY (1975). Test swojej metody globalnego rozpoznawania izolowanych wyrazów przeprowadził Itakura w oparciu o słownik złożony z 200 wyrazów, którymi były japońskie nazwy geograficzne wymawiane przez głos męski. Na 2000 wypowiedzi zebranych w okresie 3 tygodni poprawnie rozpoznanych zostało 97,3%. Itakura przyznaje jednak, że na wynik taki miał wpływ szczególny wybór słownika. Bowiem dla słownika złożonego z angielskich nazw alfa-numerycznych, na 720 wypowiedzi testowych tym samym głosem co w poprzednim teście, poprawność rozpoznania wyniosła 88,6%. Itakura zacytował prace WJELICZKI i ZAGORUJKI (1970) oraz REDDY'ego (1969), w których innymi metodami dla podobnych rozmiarów słownika uzyskano porównywalne wyniki mimo, iż w jego eksperymencie mówiący przebywał w normalnych warunkach akustycznych, a sygnał mowy przesyłany był do systemu rozpoznającego poprzez konwencjonalne urządzenia telefoniczne.

System rozpoznawania wyrazów izolowanych LEVINSONA, ROSENBERGA i FLANAGANA (1977) przystosowany do rozpoznawania wypowiedzi jednego mówcy w obrębie 127-wyrazowego słownika popełniał błędy w 11,7% przypadków. Test rozpoznawania wykonany przy pomocy tego systemu z udziałem kilku głosów męskich i żeńskich przyniósł 34,9% błędów.

RABINER (1978) przedstawił metodę rozpoznawania wyrazów izolowanych dla kilku głosów. W przeprowadzonych przez niego

badaniach brały udział 2 grupy głosów, jedna licząca 4 głosy, druga 8. Każdą grupę tworzyły po połowie głosy męskie i żeńskie. Test metody przeprowadzono na podstawie słownika złożonego z 54 wyrazów należących do terminologii komputerowej, uzyskując poprawność rozpoznawania w granicach 85%. Poszczególne wyrazy słownika miały po jednym lub po dwa wzorce. Uzyskiwano je dla każdego wyrazu grupując obrazy wypowiedzi różnych głosów i wyłaniając dla każdej grupy obraz uśredniony.

Dane o uzyskiwanych wynikach rozpoznawania izolowanych wyrazów podawane są w literaturze zazwyczaj w kontekście prezentacji jakiejś innowacji wprowadzonej do znanych już metod rozpoznawania lub też przy okazji przedstawienia całkiem nowego sposobu rozpoznawania. Ten drugi przypadek występuje znacznie rzadziej.

SAKOE i CHIBA (1978) dążąc do optymalizacji logarytmu dynamicznego programowania stanowiącego zasadniczy element dużej części metod globalnego rozpoznawania wyrazów uzyskali wyniki na poziomie 0,2 i 0,8% błędu. Pierwsza z tych wielkości odnosi się do słownika złożonego z japońskich nazw dziesięciu cyfr, a druga do słownika 50 japońskich nazw geograficznych. W teście rozpoznawania cyfr brało udział 10 głosów męskich. Każdy mówiący wypowiedział 6-krotnie każdą cyfrę. Przeprowadzono 6 serii testów. W każdej serii jedna wypowiedź każdej cyfry pełniła rolę wzorca, a pozostałe 5 rozpoznawano. W teście rozpoznawania wypowiedzi nazw geograficznych uczestniczyły 2 głosy męskie i 2 żeńskie. Z sześciu wypowiedzi każdej nazwy przez każdy z czterech głosów również pierwszą wypowiedź przyjmowano jako wzorzec, a pozostałe rozpoznawano.

RABINER i WILPON (1979) badając różne techniki klasteryzacji w zastosowaniu do globalnego rozpoznawania wyrazów wymawianych przez dowolny głos uzyskali poprawność rozpoznawania ok. 80% i 86%. 39 wyrazów stanowiących głównie angielskie nazwy liter alfabetu i dziesięciu cyfr zostało wypowiedzianych przez 50 głosów męskich i tyleż samo żeńskich. Tych 100 wypowiedzi poddano klasteryzacji różnymi metodami w celu wyłonienia reprezentatywnych wzorców poszczególnych wyrazów słownika. Wynik 80% pochodzi z testu rozpoznawania, w którym uczestniczyły 4 głosy męskie i 4 żeńskie wypowiadając jednokrotnie każ-

dy z 39 wyrazów słownika. Głosy te nie brały udziału w etapie uczenia. 80-procentową poprawność rozpoznawania uzyskano podczas testu, w którym 100 mówców biorących wcześniej udział w adaptacji wypowiadało 10-krotnie w porządku przypadkowym każdy z 39 wyrazów słownika.

Ci sami autorzy w późniejszej pracy (1980), w której zaproponowali prostą metodę treningu dla systemu rozpoznawania izolowanych wyrazów, posługując się tym samym co poprzednio słownikiem, uzyskali poprawność rozpoznawania w granicach od 78 do 87% zależnie od głosu. Porównywalne z powyższymi wyniki rozpoznawania wyrazów izolowanych uzyskiwano także w pracach: BROWNA i RABINERA (1982), MYERSA i innych (1980) oraz DAS'a (1982), które poświęcone były próbom zmodyfikowania algorytmów dynamicznej normalizacji czasowej porównywanych obrazów akustycznych.

NARA i inni (1982) przedstawili metodę globalnego rozpoznawania wyrazów z bardzo uproszczonym algorytmem dynamicznej normalizacji czasowej, dzięki któremu ilość obliczeń zmalała 10-krotnie, a wymagania co do rozmiarów pamięci 9-krotnie w porównaniu z tradycyjną normalizacją czasową. System funkcjonujący w oparciu o tę metodę przetestowano na bardzo dużym słowniku, bo liczącym aż ponad 1000 wyrazów, z udziałem 5 głosów męskich z indywidualnymi zbiorami wzorców uzyskując 95,8-procentową poprawność rozpoznawania. System rozpoznaje w czasie rzeczywistym, co zawdzięcza się zarówno bardzo uproszczonej algorytmizacji pewnych zwykle bardzo czasochłonnych procedur rozpoznawania, jak i nowoczesnej realizacji technicznej.

Metodę rozpoznawania wyrazów izolowanych w czasie rzeczywistym przedstawili także GREER, LOWERRE i WILCOX (1982) z Laboratorium Hewlett-Packarda. Metoda ich odbiega nieco od zasad stosowanych w globalnym rozpoznawaniu wyrazów. W jej skład wchodzi segmentacja wypowiedzi polegająca na lokalizacji wydatnych zmian cech sygnału mowy. Tradycyjna dynamiczna normalizacja czasowa zastąpiona została porównywaniem ciągów swoiście pojętych segmentów reprezentowanych przez średnie wektory cech. Test tej metody przeprowadzony dla słownika złożonego jedynie z angielskich nazw dziesięciu cyfr, z udziałem 3 głosów żeńskich i 3 męskich przyniósł wyniki bardzo zróżnicowane w za-

leżności od głosu, mimo iż rozpoznawanie tą metodą zaliczono do niezależnych od głosu. Obok 100-procentowej poprawności rozpoznawania uzyskanej dla dwóch głosów, dla innych dwóch głosów wyniki wyniosły 94 i 95%.

LAMEL i ZUE (1982) zaproponowali udoskonalenia w rozpoznawaniu angielskich nazw liter alfabetu oraz dziesięciu cyfr w systemie opartym o dynamiczne programowanie poprzez wykorzystanie niektórych informacji fonetycznych. Udoskonalenia te miały na celu zredukowanie ilości obliczeń wykonywanych podczas rozpoznawania. Pierwsze z nich polegało na podziale słownika na podzbiory. Każdy podzbiór charakteryzowała identyczna struktura sylabiczna poszczególnych elementów. Drugim udoskonaleniem było nadanie szczególnej wagi spółgłosce oraz segmentowi przejściowemu pomiędzy spółgłoską i samogłoską w rozpoznawanym wyrazie. Porównywanie obiektu z wzorcem ograniczono do zakresu spółgłoski oraz segmentu przejściowego. Podział słownika pozwolił na około 40-procentowe zredukowanie ilości obliczeń, a zastosowanie częściowego porównywania przyniosło ponad 30-procentowy zysk w ilości obliczeń i zapotrzebowaniu na pamięć. W teście rozpoznawania brało udział 5 głosów męskich i 5 żeńskich a dokładność rozpoznawania wykazywała znaczne uzależnienie od głosu wahając się w granicach od ok. 80 do 94%. Każdy wyraz słownika wypowiedziany był 7-krotnie.

Na zakończenie tego przeglądu wybranych przykładów globalnego rozpoznawania wyrazów wspomnieć warto o wynikach, jakie osiągnęli BAUDRY i DUPEYRAT (1982) stosując metodę opartą o parametry czasowe. Metodę tę scharakteryzowano już pokrótce w niniejszej pracy w rozdziale poświęconym omówieniu metod parametryzacji sygnału mowy w systemach rozpoznawania mowy. W próbie testowej uzyskano 98-procentową poprawność rozpoznawania dla 10-krotnej wypowiedzi każdego z 32 haseł słownika złożonego z nazw 10 cyfr oraz wybranych wyrazów z dziedziny radiotransmisji. Model Baudry'ego i Dupeyrata zalicza się do jednogłosowych, a więc rozpoznających tylko wypowiedzi osoby, dla której utworzony został zbiór wzorców.

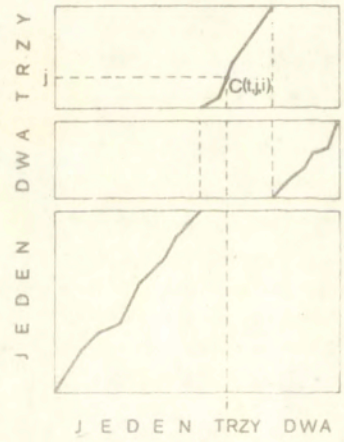
Rozpoznawanie wyrazów izolowanych stanowi jedynie fragment szerokiego problemu rozpoznawania mowy. Globalne rozpoznawanie wyrazów jest jednym ze sposobów rozpoznawania wyrazów, sposo-

bem nie bez powodów krytycznie dotychczas ocenianym. Jego główną wadą jest to, że operuje wyrazem jako jednostką podstawową, i że każdy wyraz przewidziany do automatycznego rozpoznawania posiadać musi co najmniej jeden wzorzec. Nakłada to zrozumiałe ograniczenie rozmiarów słownika rozpoznawanych wyrazów.

Aktualne badania w dziedzinie rozpoznawania mowy obejmują między innymi rozpoznawanie łączonych wyrazów (connected words). Jak wiadomo, z k wyrazów utworzyć można

$$N = k\,! + \sum_{n=1}^{k-1} \binom{k}{n}(k-n)! \qquad (2.21)$$

kombinacji. BRIDLE i inni (1982) dali przykład wykorzystania technik stosowanych w rozpoznawaniu wyrazów izolowanych do rozpoznawania wyrazów łączonych. Dysponując wzorcami wyrazów izolowanych można przy użyciu dynamicznej normalizacji czasowej rozpoznać wypowiedź wyrazów łączonych bez konieczności posiadania jej wzorca. Ilustruje to rys. 6, przedstawiający trajektorie optymalnych skojarzeń kolejnych segmentów wypowiedzi z odpowiednimi segmentami wzorców trzech izolowanych wyrazów.



Rys. 6. Trajektorie optymalnych skojarzeń kolejnych elementów wypowiedzi trzech połączonych wyrazów z elementami wzorców tych wyrazów

Podobieństwo C części wypowiedzi sięgającej miejsca „i" z wzorcami poszczególnych wyrazów izolowanych tworzą podobieństwa optymalnie skojarzonych pierwszych „i" segmentów tej wypowiedzi z segmentami t-1 wzorców oraz z pierwszymi „j" segmentami wzorca „t". Porządek kojarzenia segmentów wypowiedzi i wzorców ilustruje krzywa na rys. 6. Do wyznaczania jej przebiegu autorzy użyli zmodyfikowanego algorytmu WINCJUKA (1971). W teście rozpoznawania wypowiedzi różnych ciągów cyfr błąd wyniósł 2,5%. Rozpoznawaniu poddano łącznie 250 cyfr w różnych połączeniach. W tej metodzie podstawowym elementem jest wyraz a rozpoznawanie wypowiedzi następuje poprzez rozpoznanie kolejnych wyrazów składających się na wypowiedź. Nie zachodzi potrzeba uprzedniej segmentacji na wyrazy, aczkolwiek dokonuje się ona ubocznie w trakcie identyfikacji poszczególnych wyrazów.

2.9. Zastosowania automatycznego rozpoznawania wyrazów

Aspekty zastosowawcze automatycznego rozpoznawania mowy zostały wszechstronnie przedstawione w pracach LEA (1980) oraz MARTINA i WELCHA (1980). Istnieją różne sytuacje wywołujące potrzebę sięgnięcia po automatyczne rozpoznawanie mowy. Kontakt człowieka z maszyną następuje zwykle przy użyciu rąk, z udziałem wzroku i przy określonym pulpicie. W sytuacji, gdy operator musi wykonać rękoma inną czynność, odwrócić wzrok od pulpitu lub odejść od niego, wówczas jego kontakt z maszyną zostaje przerwany. Tymczasem w wielu przypadkach jest to niepożądane lub nie powinno mieć miejsca. Jeśli np. osoba czerpiąca informacje lub dysponująca takimi ma je przekazać komputerowi, lecz nie może użyć w tym celu rąk, wówczas potrzebny jej jest pomocnik. Jest niemal regułą, że dane wprowadzane przez operatora do maszyny odczytywane są przeważnie przez niego z jakiegoś zapisu na papierze. Wzrok operatora spoczywa na przemian na zapisie i na klawiaturze pulpitu. Jedynie wysoko kwalifikowani operatorzy potrafią obsługiwać klawiaturę nie odrywając wzroku od źródła danych. Osiągnięcie takiej perfekcji wymaga jednak szczególnych predyspozycji i długotrwałego szkolenia. Cykliczne czynności odczytu danych i przeniesienia ich na klawiaturę pulpitu odczuwane są przez przeciętnego operatora jako uciąż-

liwość. Dla operatora zmuszonego do mobilności podczas przekazywania maszynie danych dostępne są obecnie przenośne pulpity z klawiaturą umożliwiające zdalne wprowadzanie danych. Przydatność takiego pulpitu jest jednak ograniczona przez jego wymiary, ciężar i niepełny zestaw klawiszy. Automatyczne rozpoznawanie mowy pozwala spełnić potrzeby i usunąć uciążliwości występujące w powyższych przykładach. Wejście do maszyny dla sterowania głosem nazywa się wejściem fonicznym (voice imput). Wprowadzenie danych za pomocą głosu uwalnia ręce operatora, który użyć je może równocześnie do innych czynności. Dzięki temu czynności dotychczas wykonywane przez dwie osoby wykonywać może jedna osoba. WELCH (1977) wykazał, że z wejścia fonicznego pochodzi mniej błędów odczytu i interpretacji niż z wejścia poprzez klawiaturę i karty danych. Posługiwanie się wejściem fonicznym daje operatorowi całkowitą swobodę ruchu. Wyposażony on jest tylko w bezprzewodowy nadajnik rozmiarów małego pudełka od papierosów oraz w mikrofon umieszczony w stałej pozycji względem ust. Jak podaje MARTIN (1976) pierwsze systemy z wejściem fonicznym znalazły się w użyciu na przełomie lat 1972 i 1973. Wyniki posługiwania się tymi systemami przez robotników fabrycznych okazały się zadowalające. Poprawność sterowania głosem była identyczna lub lepsza od wcześniej uzyskiwanej poprawności sterowania z klawiatury przez ten sam personel. Okazało się też, że głos operatora był przez wiele miesięcy wystarczająco stabilny, dzięki czemu nie zachodziła konieczność częstych readaptacji.

MARTIN i WELCH (1980) oraz LEA (1980) podają szereg konkretnych przykładów zastosowania wejścia fonicznego (WEF-u). Oto kilka z nich:

Kontrola jakości ściany przedniej kineskopu telewizji kolorowej składa się z 54 niezależnych pomiarów. Kiedyś kontrola taka wymagała współdziałania dwóch osób. Dzięki zastosowaniu wejścia fonicznego czynności kontrolne wykonuje obecnie tylko jedna osoba manipulując obiema rękoma przy dużej i ciężkiej ścianie ekranu oraz przy skomplikowanych przyrządach pomiarowych i podając na bieżąco głosem wyniki kontroli. W podobny sposób usprawniono kontrolę jakości dekli do pojemników na płyny, dzięki czemu szybkość kontroli tego produktu wzrosła na niektórych stanowiskach aż o blisko 40%.

W ostatnich latach wzmógł się bardzo ruch wszelkiego rodzaju towarów. Wywołało to rozwój zautomatyzowanych urządzeń sortujących. Użycie automatycznego rozpoznawania mowy w procesie sortowania przesyłek przyczyniło się do zwiększenia prędkości sortowania przy zmniejszonej ilości personelu obsługi. Zmalała też liczba popełnianych pomyłek. Po raz pierwszy użyto WEF-u w automatycznych sortowniach już w 1973 r.

Inną dziedziną, w której z dużym pożytkiem stosuje się wejście głosowe jest kartografia. Np. dla uzyskania mapy ukształtowania terenu określonego obszaru dna morskiego należy wprowadzić do komputera dużą ilość danych o głębokości w poszczególnych miejscach geograficznych. W tym celu operator ustawia tzw. kursor w poszczególne miejsca geograficzne na mapie i podaje głosem odnośne dane o głębokości. W podobnej roli WEF znalazło zastosowanie w fabrykach obwodów scalonych przy tworzeniu maskownic obwodu scalonego. Ręce i oczy operatora zajęte są ustawianiem kursora na wielkiej tablicy świetlnej w różnych pozycjach, dla których podane być muszą odpowiednie parametry elementów obwodu scalonego. Dzięki zastosowaniu automatycznego rozpoznawania mowy dane te przekazane zostają głosem.

Powyższe przykłady świadczą, iż rozpoznawanie mowy wyszło już na ogół poza obręb laboratoriów naukowo-badawczych i z pożytkiem stosowane jest w różnych dziedzinach życia. Rozwinęła się już produkcja i sprzedaż układów rozpoznawania mowy. W rozwiniętych krajach świata nabyć można tanie przystawki akustyczne do komputera umożliwiające tworzenie prostych hobbystycznych układów rozpoznawania mowy. Sprzedawane są też podręczniki i instrukcje dla zainteresowanych zastosowaniem automatycznego rozpoznawania mowy. Potencjalnym klientom oferuje się szeroką gamę kompletnych systemów rozpoznawania mowy po cenach od kilku do kilkudziesięciu tysięcy dolarów. Np. firma Nippon Electric Company reklamuje 2-kanałowy system rozpoznający z poprawnością ponad 99,5% izolowane wyrazy, łączone cyfry lub ciągi wyrazów. Jego cena wynosi około 80 tys. dolarów. Przemysł, władze i instytucje badawcze pracują nad ustaleniem standardowych testów dla porównawczych ocen osiągalnych systemów, gdyż ich ilość w ostatnich kilku latach uległa zwielokrotnieniu, a poprawność działania i ceny są bardzo zróżnicowane.

3. Model globalnego rozpoznawania wyrazów na podstawie spektrogramów binarnych - ROWEIR 1

3.1. Uwagi wstępne

Nauka polska już od kilkunastu lat wykazuje wieloma pracami swoją obecność w nurcie badań nad automatycznym rozpoznawaniem mowy. Pierwsze polskie prace dotyczące tej problematyki [63], [64], [65], [101] ukazały się w latach sześćdziesiątych. W roku 1972 przedstawiono w formie publikacji [56] informację o podejmowanych w Polsce badaniach nad automatycznym rozpoznawaniem mowy. Do chwili obecnej ukazało się w kraju kilkadziesiąt prac poruszających różnorakie zagadnienia z tej dziedziny. Niektóre z nich [35], [47], [55], [66], [127] dotyczyły zagadnień ogólnych lub podstawowych, inne natomiast [101], [28], [29], [73], [34], [31] pewnych konkretnych metod rozpoznawania wybranych elementów mowy, najczęściej samogłosek. Szereg prac określić można jako przyczynkowe, gdyż poświęcono je pewnej wycinkowej problematyce z dziedziny rozpoznawania mowy. Nieliczne polskie publikacje [137], [31], [83], [87] donoszą o opracowaniu w kraju kompletnych modeli automatycznego rozpoznawania wyrazów nadających się po odpowiedniej adaptacji do wybranych zastosowań. Zakres i poziom prac polskich uwarunkowany był zawsze pozostającą w dyspozycji poszczególnych laboratoriów techniką badawczą. Mimo, iż z upływem czasu technika ta w Polsce ustawicznie rozwija się, to jednak swoim poziomem nigdy nie dorównywała technikom stosowanym przez przodujące ośrodki badawcze na świecie zajmujące się problemem automatycznego rozpoznawania mowy. Prace polskie z konieczności więc cechowała rezygnacja z metod rozpoznawania mowy opartych o najnowocześniejsze środki techniczne, gdyż takie okazywały się aktualnie i w najbliższym czasie w kraju niedostępne. Badania polskie ukierunkowywały się zatem na poszukiwanie metod własnych, dostosowanych do aktualnie istniejących możliwości realizacyjnych. Brak właściwych środków technicznych zmuszał do znajdowania metod automatycznego rozpoznawania mowy nacechowanych pro-

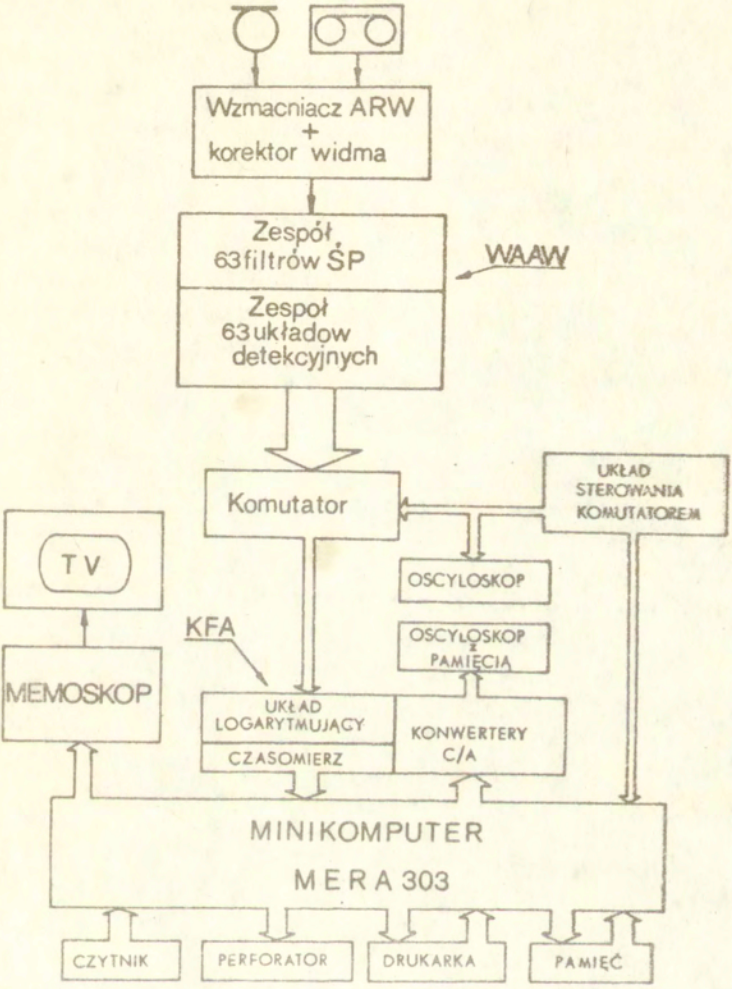
stotą i łatwością realizacji, lecz jednocześnie wystarczająco niezawodnych.

Jedna z takich metod, opracowana przez autora, jest przedmiotem niniejszej rozprawy. Dotyczy ona na razie globalnego rozpoznawania izolowanych wyrazów, lecz da się prawdopodobnie przekształcić w przyszłości w metodę rozpoznawania elementów segmentalnych w mowie, a następnie wypowiedzi ciągów wyrazowych lub zdań. Przedstawiona poniżej koncepcja modelu globalnego rozpoznawania wyrazów podporządkowana została pewnym specyficznym warunkom technicznym, w jakich praca niniejsza powstawała.

3.2. Baza techniczna modelu ROWBIR 1

Podjęcie badań nad automatycznym rozpoznawaniem mowy wymagało zbudowania odpowiedniego zestawu badawczego, który zresztą w trakcie późniejszego trwania badań ulegał ustawicznym modyfikacjom i rozbudowie. Badania przypadły bowiem na okres dynamicznego rozwoju mikroelektroniki i informatyki. Nie wszystko jednak, co przynosił postęp techniki, było dostępne; bądź to ze względu na nieosiągalność na rynku, bądź też ze względu na wysokie ceny i brak środków na zakup. Wybór narzędzi badawczych odbywał się zatem w warunkach bardzo wielu ograniczeń, które wywarły znaczący wpływ na ostateczny kształt układu, w oparciu o który powstał model globalnego rozpoznawania wyrazów ROWBIR 1, będący przedmiotem niniejszej pracy. Schemat blokowy tego układu przedstawia rys. 7. Nazwa modelu ROWBIR jest mnemonicznym skrótem określenia: rozpoznawanie wyrazów na podstawie ich reprezentacji binarnej. Pojęcie binarnej reprezentacji wyrazu zostanie wyjaśnione w jednym z dalszych rozdziałów tej pracy.

Centralnym elementem wyżej wspomnianego układu do badań modelowych w zakresie rozpoznawania wyrazów jest minikomputer MERA 303 produkcji polskiej, zaliczany do klasy minikomputerów biurowych. Operuje on słowem ośmiobitowym, a objętość jego pamięci operacyjnej wynosi zaledwie 8 K bajtów. Standardowe wyposażenie minikomputera MERA 303 stanowią czytnik i perforator taśmy papierowej oraz drukarka znakowo-mozaikowa. Urządzenia te są także produkcji polskiej. W trakcie trwania badań nad auto-



Rys. 7. Schemat blokowy układu, w oparciu o który powstał model ROWBIR 1

matycznym rozpoznawaniem wyrazów poszerzono ten zestaw o jednostkę pamięci zewnętrznej na dyskach elastycznych /dyskietkach/. Pojemność pamięciowa jednej dyskietki wynosi 256 K bajtów. Tego rodzaju pamięć zewnętrzna nie spełnia jednak wymagań szybkiej transmisji danych do pamięci operacyjnej jednostki centralnej, stawianego przez model rozpoznawania wyrazów.

Istnieje wiele przekonujących dowodów na to, że dystynktywnie najostrzejszy opis elementów mowy możliwy jest jedynie w zakresie częstotliwości. Tak więc nieodzownym etapem procesu rozpoznawania mowy powinno być przekształcenie czasowo-częstotliwościowe sygnału mowy. Ten etap określa się zwykle mianem analizy akustycznej, a jego wynikiem jest zbiór parametrów wyrażających mniej lub bardziej dokładnie różne cechy widmowe. Skłaniając się definitywnie ku poglądowi, iż najbardziej dystynktywny jest opis sygnału mowy za pomocą parametrów widmowych, założono, że próby stworzenia własnego modelu rozpoznawania wyrazów oparte będą o tego rodzaju opis, aczkolwiek przy użyciu szczególnego rodzaju parametrów widmowych. Z takiego założenia wynikła konieczność wyposażenia zestawu w analizator widmowy. Programowa analiza widmowa sygnału mowy przy pomocy minikomputera MERA 303 była nie do przyjęcia ze względu na małą zdolność obliczeniową tego komputera, wynikającą zarówno z jego małej szybkości działania, jak i znikomej pamięci operacyjnej. Zrozumiały wymóg, aby analiza widmowa rozpoznawanej wypowiedzi przebiegała w czasie rzeczywistym, jest niemożliwy do spełnienia nie tylko przez minikomputer MERA 303, lecz także przez inne nieporównywalnie doskonalsze komputery. Etap analizy widmowej może być wykonany w czasie rzeczywistym jedynie przez wyspecjalizowane urządzenie. Może nim być wielokanałowy analizator widma w wersji analogowej lub cyfrowej, lub też układ szybkiej transformacji Fouriera zbudowany z bardzo szybko działających elementów mnożąco-sumujących oraz odpowiednich buforów pamięciowych o bezpośrednim dostępie. Gotowe urządzenia tego rodzaju są w Polsce nieosiągalne. Nieosiągalne są także w naszym kraju elementy do konstrukcji analizatora cyfrowego oraz układu szybkiej transformacji Fouriera. Nie rysują się żadne perspektywy poprawy tej sytuacji w najbliższej przyszłości. W naszych krajowych warunkach technika analogowa pozostaje od lat

jedynym realnym środkiem zapewniającym realizację analizy widmowej sygnału mowy w czasie rzeczywistym.

Z wyżej wymienionych powodów w przedstawionym tutaj zestawie urządzeń do badań nad rozpoznawaniem mowy znajduje się analogowy 63-kanałowy analizator widma WAAW koncepcji i konstrukcji autora. Na rys. 8 przedstawiono schemat blokowy jednego kanału analizatora. Tworzy go przede wszystkim aktywny filtr środkowo-przepustowy trzeciego rzędu, którego operatorowa funkcja transmitancji określona jest wzorem:

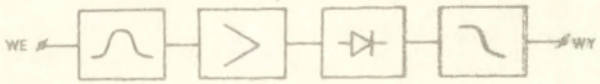
$$H(s) = \frac{1}{(s+1) \cdot (s^2 + s + 1)} \qquad (3.1)$$

Charakterystyka amplitudowo-częstotliwościowa tego filtru jest aproksymacją Butterworth'a do kształtu charakterystyki idealnego filtru.

Każdy kanał zawiera ponadto układ detekcyjny złożony ze wzmacniacza różnicowego, prostownika dwupołówkowego oraz czynnego filtru dolno-przepustowego trzeciego rzędu, także Butterworth'a, o szerokości pasma przepustowego wynoszącej 80 Hz.

Pasma analizy kanałów o numerach od 1 do 43 posiadają jednakową szerokość równą 80 Hz i pokrywają łącznie zakres częstotliwości od 120 Hz do 3560 Hz. Szerokości pasm analizy pozostałych dwudziestu kanałów analizatora są liniową funkcją numeru kanału zgodnie z następującą zależnością:

$$\Delta f_{pk} = 80 + (k - 43) \cdot 20 \qquad (3.2)$$

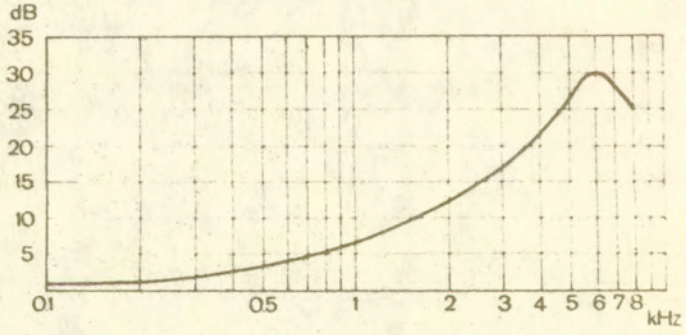


Rys. 8. Schemat blokowy jednego kanału analizatora widma WAAW

Szerokość pasma pierwszego z tych 20 kanałów wynosi 100 Hz, a ostatniego - 480 Hz. Łącznie pokrywają one zakres częstotliwości od 3560 Hz do 8310 Hz.

Przyjęty w analizatorze podział zakresu analizy na pasma jest wynikiem dążenia do uzyskania możliwie dobrej rezolucji częstotliwościowej analizatora przy utrzymaniu jego wymiarów w rozsądnych granicach. Przyjęcie stosunkowo dużej liczby pasm analizy wynikało z przeznaczenia analizatora do badań, które miały dopiero doprowadzić do ustalenia, ile określonego rodzaju parametrów widmowych należy brać pod uwagę w rozpoznawaniu mowy. Do połączonych razem wejść poszczególnych kanałów analizatora sygnał mowy zostaje doprowadzony poprzez układ wyposażony w automatyczną regulację wzmocnienia i preemfazę. Zastosowaną charakterystykę preemfazy przedstawiono na rys. 9. Dobrano ją tak, aby uzyskać w przybliżeniu jednakowy średni poziom znaczących grup składowych sygnału mowy.

Wyjścia poszczególnych kanałów łączone są komutacyjnie na wyjście analizatora. Wyjście każdego kanału zostaje połączone z wyjściem analizatora na przeciąg czasu $\triangle t_k$ wynoszący około



Rys. 9. Charakterystyka preemfazy zastosowanej w WAAW-ie

100 mikrosekund raz na każdy okres komutacji T_k równy około 23 ms. Wartość czasu $\triangle t_k$ oraz okresu komutacji T_k podlegają ograniczeniu ze względu na szybkość działania minikomputera MERA 303 oraz logarytmującego konwertera analogowo - cyfrowego, który przedstawiony zostanie poniżej. W dodatku znajdującym się na końcu obecnej pracy zamieszczono tabelę zawierającą dane o częstotliwościach środkowych i szerokościach pasm poszczególnych kanałów wyżej opisanego analizatora, a także ideowy schemat pojedynczego kanału.

Współpraca wielokanałowego analogowego analizatora widma WAAW z minikomputerem MERA 303 wymaga odpowiedniego urządzenia pośredniczącego. Rolę takiego urządzenia pełni odpowiednio zmodyfikowany kanał funkcji analogowych skonstruowany w Pracowni Fonetyki Akustycznej IPPT PAN przez K.MYTKOWSKIEGO [103] w ramach jego pracy magisterskiej, wykonanej pod opieką autora.

Zasadniczą częścią zmodyfikowanego kanału funkcji analogowych jest logarytmujący konwerter analogowo-cyfrowy, logarytmujący i przetwarzający w postać cyfrową dane odczytywane na wyjściu analizatora. Konwerter ten wykorzystuje powszechnie znaną logarytmiczną zależność

$$t = a \log U_c + b \qquad (3.3)$$

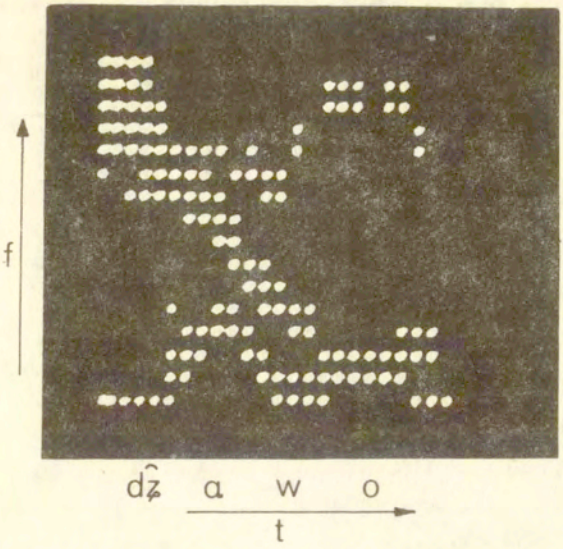
pomiędzy czasem t rozładowywania się kondensatora a napięciem U_c, do którego nastąpiło rozładowanie. Współczynniki a i b są określone przez stałą czasu T obwodu rozładowania kondensatora oraz napięcie początkowe U_o panujące na kondensatorze w chwili $t = 0$, następującymi zależnościami: $a = -T$, $b = T \log U_o$. Od momentu $t = 0$ inicjującego kolejny okres konwersji analogowo-cyfrowej napięcie elektryczne podlegające konwersji jest ustawicznie porównywane z napięciem na rozładowującym się kondensatorze. W czasie oczekiwania na zrównanie się obu tych napięć następuje zliczanie impulsów o częstotliwości około 1 MHz. Zrównanie napięć powoduje zatrzymanie zliczania. Stan licznika będący wynikiem konwersji wartości analogowej do postaci cyfrowej z równoczesnym jej zlogarytmowaniem zostaje przepisany programowo do komputera MERA 303. Dokładność konwersji wynosi około 0,5 dB.

W skład zestawu powstałego do badań nad rozpoznawaniem mowy wchodzą ponadto urządzenia do prezentacji informacji w formie obrazów. Zalicza się do nich między innymi standardowy oscyloskop monitorujący stan wyjść WAAW-u. Obserwacja widm sygnału mowy możliwa jest przy jego pomocy tylko w przypadku dźwięków stacjonarnych, a więc przeciągle wymawianych niektórych głosek izolowanych. Oscyloskop ten służy przede wszystkim do testowania analizatora.

Do prezentacji obrazów akustycznych dowolnego sygnału mowy zestaw wyposażono w oscyloskop dwukanałowy z długą poświatą, odbiornik telewizji czarno-białej oraz urządzenie zwane MEMOSKOPEM, skonstruowane w całości w Pracowni Fonetyki Akustycznej IPPT PAN specjalnie do badań nad rozpoznawaniem mowy.

Obraz akustyczny mowy na ekranie oscyloskopu z poświatą powstaje w wyniku cyklicznej generacji przez komputer sygnałów sterowania odchylaniem plamki świetlnej. Źródłem dla tej generacji są dane o konkretnym obrazie akustycznym zawarte w pamięci komputera. Sygnały wysyłane przez komputer nadają się do sterowania odchylaniem plamki świetlnej oscyloskopu dopiero po przekształceniu ich do postaci analogowej. Do tego przekształcenia służy tor konwersji cyfrowo-analogowej będący częścią składową wyżej wymienionego kanału funkcji analogowych. Kanał ten jest ważnym elementem prezentowanego tutaj zestawu. Na rys. 10 pokazano przykład spektrogramu binarnego wyświetlonego na ekranie oscyloskopu z długą poświatą.

Memoskop jest urządzeniem do generacji sygnału telewizyjnego przenoszącego obraz akustyczny sygnału mowy na ekran odbiornika telewizji czarno-białej. Obraz akustyczny mowy zaprogramowany w komputerze na podstawie danych pochodzących z analizy akustycznej zostaje wpierw wpisany do pamięci memoskopu. Jej wielkość wyrażona iloczynem 63 x 256 odpowiada liczbie punktów, z których utworzony może być obraz na ekranie telewizyjnym. Liczba 63 odpowiada maksymalnej liczbie wierszy, składających się na obraz telewizyjny, a 256 jest liczbą punktów tworzących jeden wiersz. Gradacja jasności każdego takiego punktu jest binarna, co oznacza, że może on przyjmować tylko dwa stopnie jasności, tj. biel lub czerń. Obraz akustyczny uzyskany na ekranie telewizyjnym przy użyciu memoskopu powstaje bez

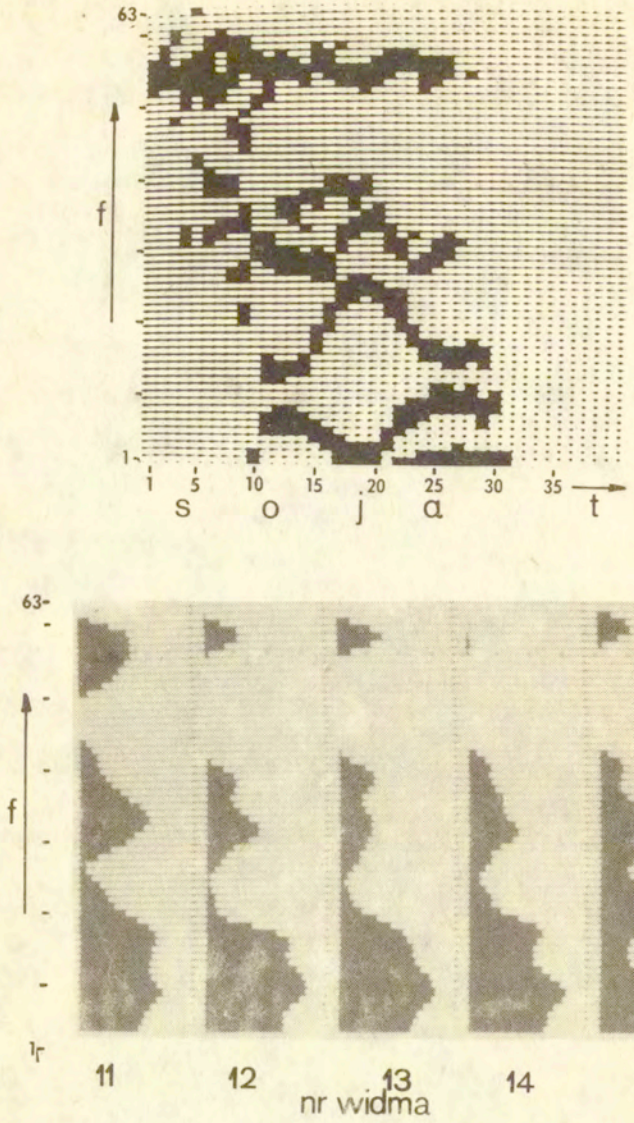


Rys. 10. Przykład spektrogramu na ekranie oscyloskopu z pamięcią

udziału konwertera cyfrowo-analogowego, który jest nieodzowny dla uzyskania podobnego obrazu na ekranie oscyloskopu z długą poświatą.

Memoskop, umożliwiając wygodną i szybką kontrolę obrazów akustycznych mowy uzyskiwanych drogą różnego rodzaju widmowych przekształceń, okazał się urządzeniem bardzo pożytecznym w badaniach nad rozpoznawaniem mowy. Na rys. 11 pokazano obrazy spektrogramu binarnego oraz widm sygnału mowy uzyskane na ekranie telewizora za pomocą memoskopu.

Przedstawiony wyżej układ analogowo-cyfrowy stworzono przede wszystkim jako narzędzie do badań nad rozpoznawaniem mowy. Niebawem okazało się, że może on być bardzo pomocny przy programowaniu syntezy mowy i w takim, między innymi, zastosowaniu jest on obecnie intensywnie używany.



Rys. 11. Przykłady obrazów spektrogramu binarnego oraz widm sygnału mowy na ekranie telewizora

4. Przesłanki fonetyczno-akustyczne właściwego wyboru parametrycznej reprezentacji sygnału mowy

Z podstaw fonetyki akustycznej wiadomo, że widma amplitudowe większości dźwięków mowy posiadają strukturę, którą określa się jako formantową. Polega ona na tym, że w pewnych zakresach częstotliwości obwiednia widma amplitudowego odwzorowuje ekstremalny fragment charakterystyki amplitudowej rezonansu akustycznego, wyrażonej wzorem:

$$\left|H(f)\right| = \left[(1 - x^2)^2 + x^2/Q^2\right]^{-\frac{1}{2}}, \qquad (4.1)$$

w którym $x = \frac{f}{F}$, $Q = \frac{F}{B}$. F oznacza częstotliwość rezonansową a B szerokość pasma rezonansu zmierzoną na poziomie -3 dB względem wierzchołka. Nierzadko jednak mają miejsce przypadki, w których podobieństwo odnośnego wycinka obwiedni widma do krzywej rezonansowej jest wielce problematyczne. Fragment widma, dla którego zachodzi podobieństwo konfiguracji obwiedni widma do krzywej rezonansowej w jej ekstremalnym zakresie nazywa się formantem.

Formant charakteryzują trzy parametry: częstotliwość, szerokość i poziom. W znacznej mierze, chociaż w bardzo różnym stopniu, w parametrach tych zawarte są cechy dystynktywne poszczególnych dźwięków mowy oraz cechy osobnicze każdego głosu. Widma większości dźwięków mowy zawierają po kilka formantów, lecz zwykle jedynie najniższe z nich wykazują znaczącą wartość dystynktywną. Szerokość i poziom formantów nie przejawiają istotnej wartości dystynktywnej, o ile tylko ich wielkość nie wykracza poza określone zakresy. Szczególna rola dystynktywna przypada natomiast częstotliwościom najniższych formantów. Struktura formantowa widoczna jest najwyraźniej w widmach samogłosek. Mniej dostrzegalna jest w widmach niektórych spółgłosek.

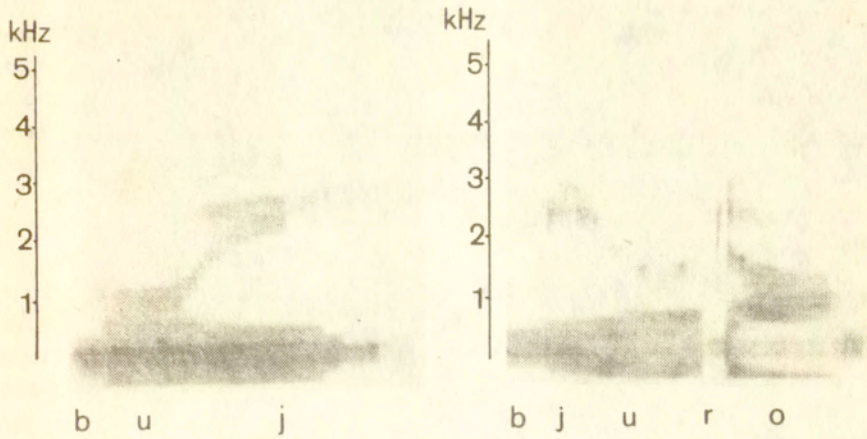
W miarę poznawania struktur widmowych poszczególnych dźwięków mowy i odkrywania szczególnej roli niektórych formatnów w

ich rozróżnianiu, tworzyc zaczęto modele rozpoznawania fonemów operujące parametrami formantowymi. Ukazało się wiele prac, w których wykazywano, że niektóre fonemy, a szczególnie samogłoski, daje się jednoznacznie wyrazic i sklasyfikowac za pomocą parametrów formantowych, głównie częstotliwości pierwszych kilku formantów. Udowadniano (przytaczając wyniki różnych eksperymentów), że na podstawie częstotliwości zaledwie dwóch pierwszych formantów można poprawnie identyfikowac samogłoski. Bogaty materiał na ten temat w odniesieniu do mowy polskiej zaprezentował w swoich pracach samodzielnych i wspólnych W.JASSEM [49], [50], [51], [52], [53], [57]. Niemal równocześnie z pracami podstawowymi poświęconymi klasyfikacji fonemów na podstawie parametrów formantowych czyniono próby automatycznej ekstrakcji częstotliwości formantów. Na przełomie lat 60-tych i 70-tych powstały w Polsce dwa modele automatycznych ekstraktorów częstotliwości formantów: EXFOR-1 i EXFOR 2. Zostały one opisane w pracach H.KUBZDELI (1970, 1973, 1976), który był równocześnie autorem obu modeli.

Idea automatycznego rozpoznawania mowy poprzez identyfikację w niej samogłosek i niektórych spółgłosek, wyraźnie zdeterminowanych częstotliwością dystynktywnych formantów, mimo swojej niewątpliwej atrakcyjności, nie doczekała się jednak pełnej realizacji. Złożyło się na to kilka powodów. Po pierwsze częstotliwości formantów dystynktywnych nie zawsze dają się ekstrahować. Dość często mają miejsce przypadki znacznej bliskości dwóch sąsiednich formantów, w wyniku czego są one po części lub całkowicie nierozróżnialne. Szczególnie niekorzystny układ występuje wtedy, gdy oprócz znacznego zbliżenia dwóch sąsiednich formantów dystynktywnych jeden z nich ma poziom znacznie niższy od drugiego. Zjawiska takie występują w przykładach zamieszczonych na rys. 12.

Ze względu na harmoniczny charakter większości dźwięków mowy, którym teoretycznie przypisuje się formantową strukturę widma, określenie położenia wierzchołka formantu jest możliwe tylko w przybliżeniu i to tym mniej dokładnie, im większa jest częstotliwość podstawowa rozpatrywanego dźwięku mowy.

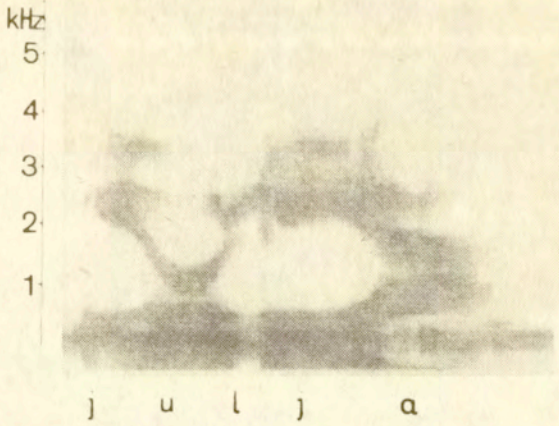
Częstotliwości formantów dystynktywnych zmieniają się w czasie artykulacji danego fonemu. Ilustrują to przykłady za-



Rys. 12. Przykłady krytycznych zbliżeń dwóch formantów

mieszczone na rys. 13. O wielkosci tych zmian decyduje kontekst artykułowanego fonemu, czyli stadia artykulacji związane z minioną artykulacją poprzedniego fonemu oraz przyszłą artykulacją następnego fonemu. Proponowane niekiedy metody rozpoznawania niektórych fonemów na podstawie częstotliwości formantów dystynktywnych zakładają natomiast, że poszczególne typy fonemów wystarczy wyrazic jedynie wartosciami częstotliwości tych formantów w stadium quasiustaloŋym, które stanowi nieraz jedynie znikomy fragment całego fonemu. Metody te przyjmują zatem rezygnację z reszty ważnej informacji, jaka zawarta jest w częstotliwościach formantów dystynktywnych w pozostałych stadiach fonemu. Warto w tym miejscu wspomniec, że w kwestii oceny ważności różnych stadiów fonemu dla potrzeb automatycznego rozpoznawania mowy konkurują z sobą dosc przeciwstawne poglądy.

FANT (1960) w swojej teorii wytwarzania mowy wykazał, że akustyczne właściwości anatomicznego aparatu artykulacyjnego człowieka znajdują swoje odbicie w widmie produktu artykulacji,



Rys. 13. Przykłady tzw. ugięć formantowych wywołanych wpływem sąsiednich fonemów

jakim jest ludzka mowa. Wytworzenie dźwięku mowy wymaga określonych ruchów efektorów artykulacyjnych. Ruchy te wywołują takie konfiguracje toru artykulacyjnego, przy których możliwe jest uzyskanie zamierzonego efektu artykulacji. Każdej konfiguracji toru artykulacyjnego odpowiada więc określone widmo wymawianego dźwięku mowy. Różnorodność konfiguracji, jakie w trakcie procesu mówienia przyjmuje tor artykulacyjny, znajduje swoje właściwe odbicie w różnorodności widm artykułowanych dźwięków.

Zagadnieniu współzależności cech akustycznych dźwięków mowy oraz geometrycznych kształtów i wymiarów toru artykulacyjnego poświęconych zostało kilka prac J.KACPROWSKIEGO (1962), (1963), (1977), (1979). Pewne cechy widmowe produktu artykulacji nadają wypowiedzi zamierzone przez osobę mówiącą znaczenie fonetyczne. Inne cechy widmowe dla znaczenia fonetycznego wypowiedzi nie odgrywają żadnej roli, a jedynie nadają jej osobniczego zabarwienia charakteryzującego głos osoby mówiącej. W przeciwieństwie do zadowalającego poziomu wiedzy o współza -

leżnościach pomiędzy widmowymi cechami mowy a akustycznymi właściwościami narządów artykulacyjnych człowieka, wiele niejasności istnieje jeszcze w kwestii roli niektórych cech widmowych dźwięków mowy w ich percepcji. Bliższe poznanie ważności poszczególnych cech widmowych dla percepcji różnych dźwięków mowy wpłynęłoby zapewne na skorygowanie niektórych obecnych poglądów na zagadnienie automatycznego rozpoznawania mowy.

Uwagi zawarte w niniejszym rozdziale prowadzą do następujących wniosków:

Bardzo ważnym zagadnieniem w automatycznym rozpoznawaniu mowy jest właściwe wyrażenie sygnału mowy za pomocą odpowiednich parametrów. Istotną rolę w tej kwestii odgrywa nie tylko trafność wyboru parametrów, lecz również stopień złożoności przekształceń sygnału mowy prowadzących do wyznaczenia tych parametrów.

Podstawowa wiedza z zakresu wytwarzania i percepcji mowy wskazuje, że widmo jest najpełniejszym obrazem akustycznym dźwięków mowy. Fakt ten skłania do poszukiwania właśnie w widmie parametrów, które najwłaściwiej opisywałyby sygnał mowy dla jej automatycznego rozpoznawania. Parametry te powinny wyrażać jedynie te cechy widmowe, które przedstawiają dużą wartość dystynktywną. Ponieważ wiadomo, że ważną cechą dystynktywną większości dźwięków mowy jest położenie niektórych formantów, cechę tę należy przede wszystkim wykorzystać w opisie parametrycznym sygnału mowy. Ekstrakcja częstotliwości wierzchołków formantowych jest z technicznego punktu widzenia zawodna i stąd unikać należy posługiwania się wprost częstotliwościami formantów w automatycznym rozpoznawaniu mowy.

Spodziewać się należy, że liczba parametrów koniecznych dla jednoznacznego opisu sygnału mowy jest uzależniona od rodzaju tych parametrów ściślej od ich wartości dystynktywnej. Dlatego badania prowadzące do ustalenia, ile parametrów wystarcza do poprawnego rozpoznawania określonych jednostek mowy, zacząć należy od możliwie szerokiego opisu parametrycznego opartego o widmo z odpowiednio dużą rezolucją częstotliwości.

5. Widma binarne jako forma reprezentacji sygnału mowy

5.1. Wygładzanie widma cyfrowego

Widmo sygnału mowy, jakim posługujemy się zwykle w praktyce, jest ciągiem danych wyrażających poziom sygnału mowy w kolejnych pasmach częstotliwości, na jakie podzielono zakres analizy. Gdy szerokości tych pasm są mniejsze od częstotliwości podstawowej sygnału mowy, wówczas w strukturze widma widoczny jest harmoniczny charakter dźwięku, przez co zamaskowaniu ulegają różne ważne cechy widmowe. Dla wyeliminowania wpływu harmonicznego charakteru dźwięku na obraz jego widma stosuje się wygładzanie widma. W widmie wygładzonym uwypuklone zostają cechy dźwięku nie związane z częstotliwością harmonicznych.

Model rozpoznawania wyrazów, któremu poświęcona jest niniejsza praca, korzysta z analizatora widma WAAW opisanego w rozdziale 2.3. Pasma przepustowe filtrów w większości kanałów tego analizatora mają szerokość 80 Hz, dzięki czemu uzyskuje się widma sygnału mowy, które z jednej strony cechuje dobra rozdzielczość częstotliwości, a z drugiej niekorzystny wpływ harmonicznego charakteru dźwięku mowy na obraz widma. Trudno jest uzyskać w takich widmach parametry, który wyrażałyby jednoznacznie sygnał mowy w późniejszym procesie rozpoznawania. Z tego powodu widmo wyznaczone przez analizator po wpisaniu go do pamięci minikomputera zostaje programowo wygładzone. Wygładzenia dokonuje się przez sumowanie wartości parametrów wyokienkowanego fragmentu widma dla każdej kolejnej pozycji okienka wagowego przesuwającego się wzdłuż osi widma reprezentującej częstotliwość. Dla każdej pozycji okienka wartość rzędnej widma wygładzonego wyliczona zostaje ze wzoru:

$$y_{iw} = \sum_{j=-k}^{+k} y_{i+j} \cdot W_j \qquad (5.1)$$

w którym y_{i+j} reprezentuje rzędną widma niewygładzonego, a W_j rzędną okienka wagowego.

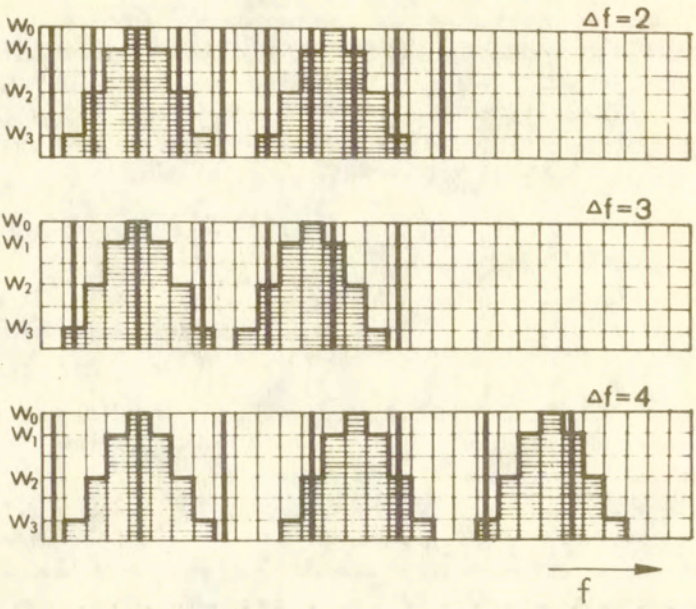
Wymiary okienka wagowego określono w oparciu o kryterium, według którego przekształcenie wygładzające ma zapewnić idealne wygładzenie widma sygnału periodycznego, złożonego z harmonicznych o jednakowej amplitudzie i mającego częstotliwość podstawową równą całkowitej krotności szerokości $\triangle f$ pasm, w jakich wyznaczone zostało widmo poddane wygładzeniu. Okienko jest symetryczne i ma kształt schodkowy. Równania:

$$\left.\begin{aligned} 2W_1 - 2W_2 + 2W_3 &= W_0 \\ -W_1 - W_2 + 2W_3 &= W_0 \\ W_1 + W_3 &= W_0 \\ 2W_2 &= W_0 \end{aligned}\right\} \qquad (5.2 \div 5.5)$$

na podstawie których wyliczone zostały wymiary okienka 7-elementowego, wyrażają powyższe kryterium dla trzech wartości częstotliwości podstawowej sygnału mowy, równych 2-, 3- i 4-krotnej szerokości $\triangle f$ pasm, w jakich dane jest widmo oraz dla trzech wybranych położeń okienka względem harmonicznych sygnału. Rysunek 14 ilustruje przypadki, do jakich odnoszą się równania, z których wyliczono wymiary okienka 7-elementowego. Kolejne rzędne tego okienka tworzą następujący ciąg liczb: 0, 1/6, 1/2, 5/6, 1, 5/6, 1/2, 1/6, 0. W procesie uśredniania widma liczby te odgrywają rolę współczynników wagowych.

W podobny sposób wyznaczyć można okienko o mniejszej lub większej liczbie elementów. Ustalono w trakcie badań, że wystarczająco dobre wygładzenie widm wyznaczonych przez WAAW uzyskuje się przy użyciu okienka mniejszego, 5-elementowego, którego wymiary są następujące: 1/4, 3/4, 1, 3/4, 1/4.

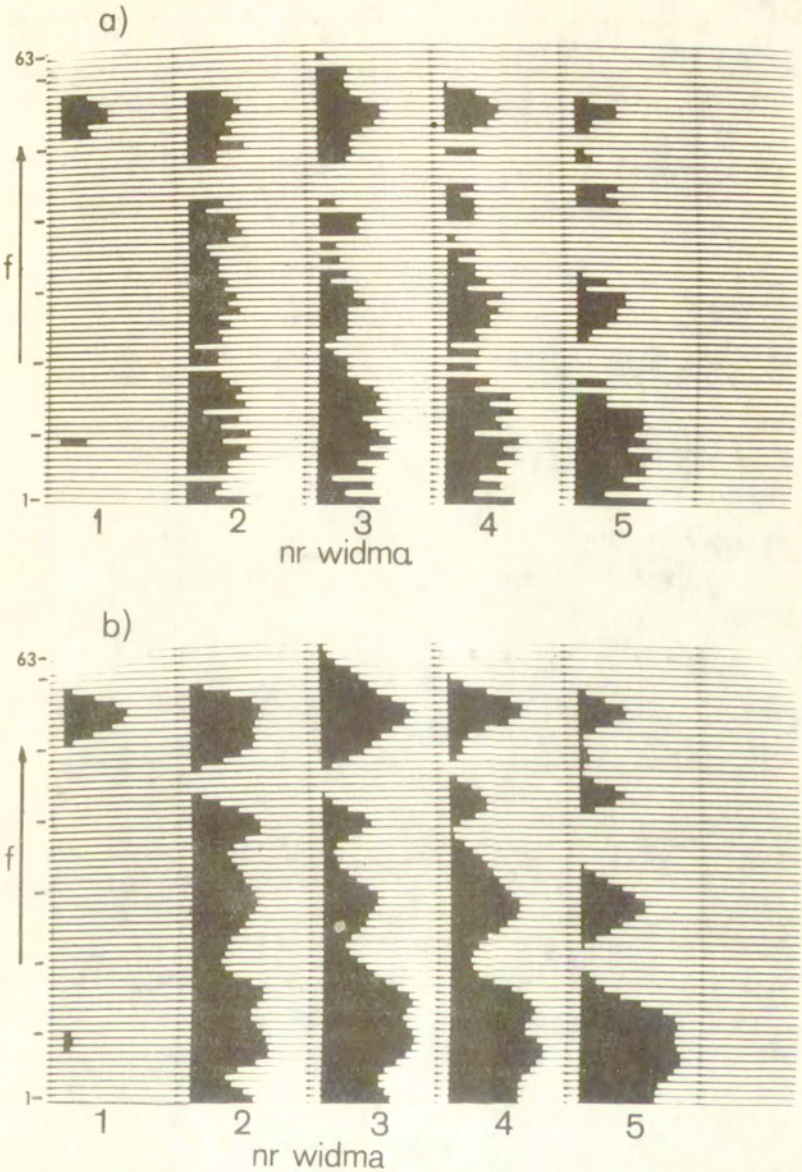
Na rys. 15 pokazano kilka przykładów widm przed wygładzeniem i po wygładzeniu przedstawioną wyżej metodą. Widma wygładzone nie wykazują wahań poziomu spowodowanych harmonicznym charakterem dźwięku, którego dotyczą. Wskutek uśrednienia wyraźnemu uwypukleniu ulega też formantowa struktura widm. Posługując się tak uśrednionymi widmami można łatwiej dokonać wyboru właściwych cech widmowych dla możliwie najkorzystniejszego wyrażenia sygnału mowy.



Rys. 14. Przypadki, do jakich odnoszą się kryteria optymalnego wygładzenia widma

5.2. Przekształcenie widma cyfrowego w widmo binarne

Położenia formantów są niewątpliwie wyraźną cechą dystynktywną większości dźwięków mowy. Jednakże, jak wspomniano wyżej, w miarę dokładny pomiar częstotliwości formantów nastręcza trudności. Okazuje się, że łatwiej i korzystniej jest ekstrahować zakresy formantowe. Interesującą metodę takiej ekstrakcji przedstawił RUSKE (1976). Najogólniej ujmując, polega ona na przekształceniu widma wygładzonego w osobliwe widmo binarne, w którym wyraźnie zaznaczone są zakresy poszczególnych formantów. Zasadniczą cechą widma binarnego jest to, że poszczególne jego



Rys. 15. Przykłady widm; a) przed i b) po wygładzeniu

parametry przyjmują jedynie wartości Ø lub 1. Przypisanie poszczególnym parametrom jednej z tych dwóch wartości zależy od wyniku oceny wartości kolejnych parametrów widma wygładzonego względem postawionego kryterium.

W wersji dyskretnej kryterium wartościowania parametrów p_{bij} widma binarnego, zastosowane przez RUSKE [118] jest następujące:

$$\text{Jeżeli } p_{wij} - \frac{1}{(2n+1)(2m+1)} \sum_{q=-n}^{n} \sum_{p=-m}^{m} p_{w(i-p)(j-q)} \geqslant 0, \qquad (5.6)$$

$$p_{bji} = 1,$$

$$p_{bji} = Ø \text{ w pozostałych przypadkach.}$$

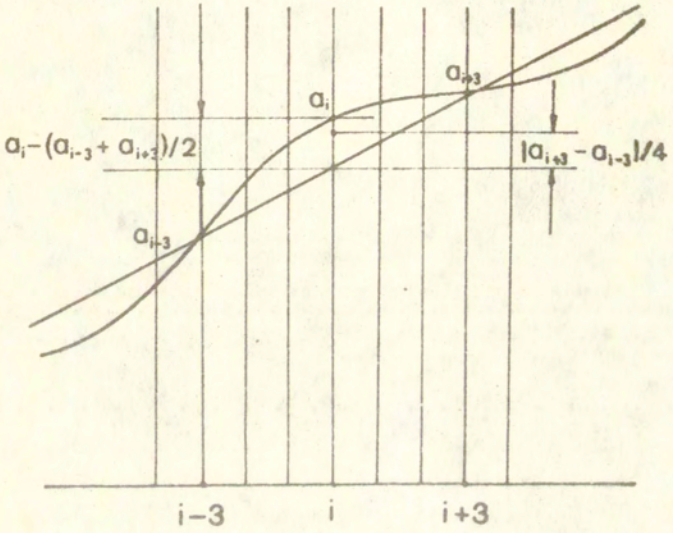
Tak więc, zgodnie z powyższym zapisem, i-ty parametr j-tego widma binarnego oznaczony przez p_{bji} przyjmuje wartość i, jeżeli wartość i-tego parametru j-tego widma wygładzonego p_{wij} jest większa lub równa średniej z wartości $(2m+1)$ parametrów widmowych $(2n+1)$ widm objętych oknem $(j-n, j+n)$, $(i-m, i+m)$.

Istotnym elementem niniejszej pracy jest zastosowanie widma binarnego jako formy wyrażenia sygnału mowy dla automatycznego rozpoznawania wyrazów. Mając na uwadze taką rolę widma binarnego, przyjęto własną zasadę jego wartościowania i zrezygnowano z kryterium podanego przez RUSKE [118]. Zamiast odnoszenia poszczególnych danych widma wygładzonego do lokalnego poziomu średniego w trakcie wyznaczania widma binarnego przyjęto regułę, według której widmo binarne przyjmuje wartość 1 w miejscach, w których obwiednia widma wygładzonego wykazuje odpowiednio wydatną wypukłość. Przyjmuje się, że wypukłość taka ma miejsce wówczas, gdy spełniona jest następująca nierówność:

$$p_{wi} - (p_{w(i-k)} + p_{w(i+k)})/2 \geqslant K\left|p_{w(i-k)} - p_{w(i+k)}\right| \qquad (5.7)$$

Wyrazami tej nierówności są odnośne parametry widma wygładzonego. Współczynnik K uzależnia ostrość kryterium wyrażonego

wzorem (5.7) od nachylenia obwiedni widma w miejscu i. Rysunek 16 przedstawia graficzną ilustrację tego kryterium.



$p_{bi}=1$, jeśli

$$a_i-(a_{i-3}+a_{i+3})/2 \geqslant |a_{i+3}-a_{i-3}|/4$$

p_w oznaczono przez a , $k=3$

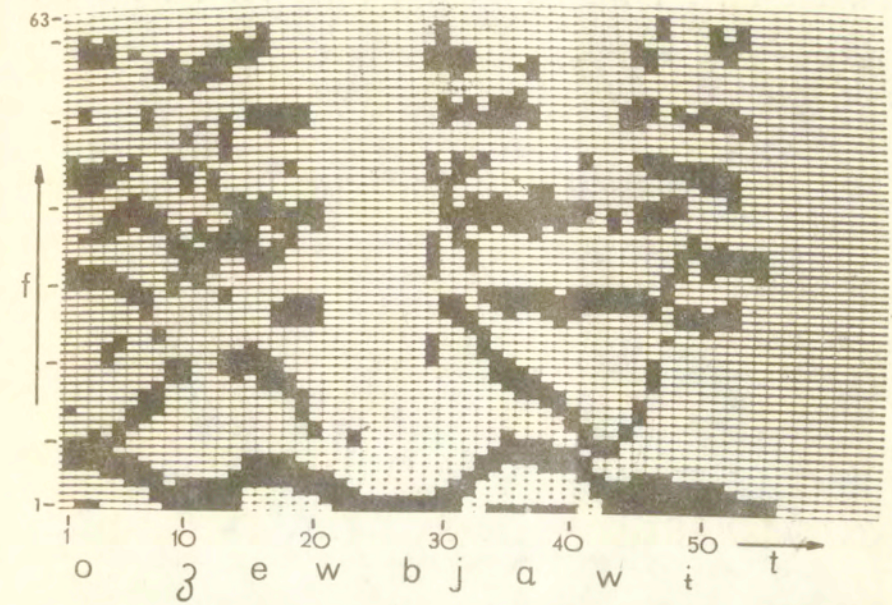
Rys. 16. Ilustracja kryterium wartościowania parametrów widma binarnego

Na rys. 17 przedstawiono kilka przykładów widma wygładzonego, przy czym u podstawy każdego z tych widm, wzdłuż osi reprezentującej częstotliwość, umieszczono odnośne widmo binarne. Zakresy, w których widma binarne przyjmują wartość „1" pokrywają się z przedziałami widma wygładzonego, w których przypadają formanty.

Rys. 17. Przykłady widm wygładzonych oraz pochodnych widm binarnych

Pojedyncze widmo binarne wyraża sygnał mowy w wąskim przedziale czasu. Z powodu ograniczonych warunków technicznych częstość otrzymywania widm binarnych wynosić może nie więcej niż około 50 na sekundę. Ciąg widm binarnych uzyskiwanych z taką częstością tworzy osobliwego rodzaju spektrogram. Spektrogram taki określa się mianem „binarny". Przykład spektrogramu binarnego w zestawieniu z klasycznym sonagramem wykonanym przy pomocy znanego analizatora Sona-Graph produkcji amerykańskiej firmy Kay Electric Company pokazano na rys. 18. Przykład ten dotyczy wypowiedzi „orzeł biały" głosem męskim.

Z porównania obu tych spektrogramów wynika, że tego typu



Rys. 18. Spektrogram binarny w zestawieniu z klasycznym sonagramem wykonanym przy pomocy Sona-Graph'u Kay'a wypowiedzi [ozewbjawł]

spektrogram binarny dobrze przekazuje te cechy mowy, które wyrażają się w rozkładzie formantów. Ponieważ przebiegom formantów przypisuje się dużą rolę dystynktywną, uważać można, że spektrogram binarny stanowi właściwą reprezentację sygnału mowy dla automatycznego rozpoznawania wyrazów. Szczególnie korzystną cechą takiej formy reprezentacji jest jej zwięzłość. Złożone z 63 parametrów widmo binarne zawiera się zaledwie w ośmiu bajtach, zaś spektrogram wypowiedzi o długości jednej sekundy zawrzeć można w czterystu bajtach.

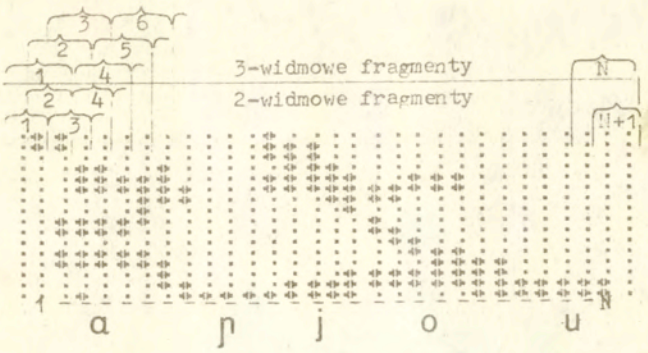
W dalszej części niniejszej pracy przytoczone zostaną dowody na to, że liczebność parametrów widma binarnego dla automatycznego rozpoznawania wyrazów może zostać ograniczona do 16. Widmo binarne 16-parametryczne zajmuje zaledwie dwa bajty, a spektrogram binarny wypowiedzi o rozciągłości 1 sekundy jedynie 100 bajtów.

6. Porównywanie spektrogramów binarnych w modelu ROWBIR 1

Przyjmując wyżej przedstawiony spektrogram binarny jako formę reprezentacji sygnału mowy, opracowano metodę automatycznego rozpoznawania wyrazów i zrealizowano ją przy użyciu minikomputera MERA 303. Powstały w ten sposób model globalnego rozpoznawania wyrazów otrzymał nazwę ROWBIR 1, która stanowi mnemoniczny skrót następującego określenia: rozpoznawanie wyrazów w oparciu o binarną reprezentację sygnału mowy.

6.1. Lokalne konfrontacje fragmentów porównywanych spektrogramów binarnych

Podstawowym i wielokrotnym działaniem w procesie globalnego rozpoznawania wyrazów jest porównywanie dwóch obrazów akustycznych. Ponieważ w modelu ROWBIR 1 obraz akustyczny ma formę spektrogramu binarnego, w grę wchodzi w tym przypadku porównywanie dwóch spektrogramów binarnych. Polega ono na określeniu podobieństw odpowiadających sobie fragmentów wzajemnie porównywanych spektrogramów, z których każdy może wyrażać konkretną wypowiedź lub być obrazem wzorcowym jakiegoś wyrazu. Fragmentem spektrogramu binarnego może być jedno widmo binarne lub grupa kilku widm bezpośrednio po sobie następujących. Przykła-



Rys. 19. Podział spektrogramu binarnego na fragmenty o jednakowej rozciągłości

dy fragmentów spektrogramu binarnego złożonego z dwóch i trzech widm pokazano na rys. 19. Jeśli spektrogram składa się z N widm to maksymalnie można go podzielić na N fragmentów jedno- lub trójwidmowych lub na N+1 fragmentów dwuwidmowych. W przypadku fragmentów dwu- i trójwidmowych spektrogram binarny zostaje poszerzony o dwa widma zerowe, jedno na początku i jedno na końcu. Jak pokazano na rys. 19, widma te wchodzą w skład pierwszego i ostatniego fragmentu dwu- i trójwidmowego.

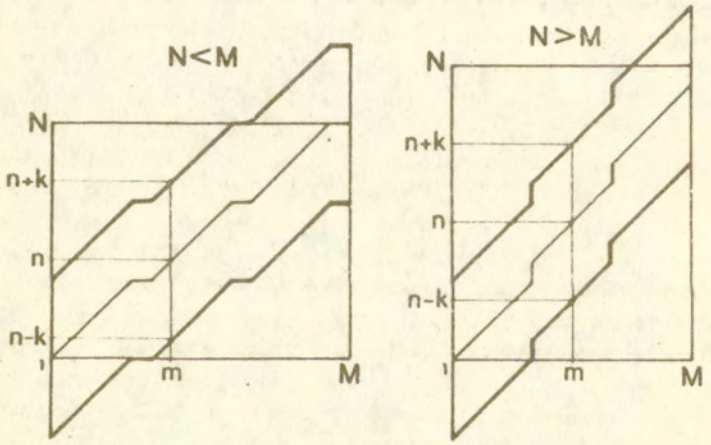
Porównywane spektrogramy binarne mają z reguły różne liczby widm i tym samym różne liczby jednakowej rozciągłości fragmentów, na jakie maksymalnie można spektrogramy te podzielić. Należy też wziąć pod uwagę, że w każdej wypowiedzi tego samego wyrazu może mieć miejsce inny rozkład czasowy poszczególnych elementów fonetycznych, z jakich wyraz ten jest zbudowany. Posłużenie się wyłącznie liniową normalizacją czasową dla uzyskania czasowej współbieżności fragmentów jednego spektrogramu z odnośnymi fragmentami drugiego spektrogramu jest zgoła niewystarczające. Wobec tego przyjęto następującą regułę poszukiwania zgodnych fragmentów w porównywanych z sobą spektrogramach binarnych. Optymalnego odpowiednika bieżącego fragmentu $Fr_m(SB_A)$ spektrogramu binarnego SB_A o długości M poszukuje się

wśród fragmentów: $Fr_{n-k}(SB_B)$, ..., $Fr_{n+k}(SB_B)$ spektrogramu binarnego SB_B o długości N. Liczby porządkowe m i n odnośnych fragmentów porównywanych spektrogramów binarnych podlegają quasiliniowej normalizacji czasowej, która stwarza pomiędzy nimi następującą współzależność:

$$n = m \pm INT\left(\frac{m}{M}|M - N|\right) \qquad (6.1)$$

Quasiliniowość normalizacji polega w tym przypadku na jej dyskretnym charakterze. Funkcja quasiliniowej normalizacji czasowej wyrażona wzorem (6.1) wyznacza oś obszaru, obejmującego zakresy poszukiwania zgodnych fragmentów we wzajemnie porównywanych spektrogramach binarnych.

Rys. 20 ilustruje, w których zakresach spektrogramu SB_B o długości N poszukiwane są fragmenty optymalnie podobne do poszczególnych fragmentów spektrogramu SB_A o długości M, w przypadku, gdy $N > M$ oraz gdy $N < M$.



Rys. 20. Obszar poszukiwania zgodnych fragmentów wzajemnie porównywanych spektrogramów binarnych

$n \leq 0$ i $n > N$ dotyczą fragmentów przypadających w marginesach ciszy występującej bezpośrednio przed i po wypowiedzi. Wykorzystanie tych marginesów służy zabezpieczeniu przed następstwami ewentualnych pomyłek w detekcji granic wypowiedzi oraz uproszczeniu procedury porównywania poprzez utrzymanie stałej szerokości zakresu, z którego dokonuje się wyboru jednego z fragmentów w trakcie kompletowania par fragmentów najbardziej podobnych. Nasuwa się zapewne w tym miejscu wątpliwość, czy nie należałoby wprowadzić dodatkowo warunku, żeby liczby porządkowe optymalnie podobnych fragmentów tworzących parę o numerze kolejnym K były odpowiednio wyższe lub co najmniej równe względem liczb porządkowych fragmentów, składających się na parę K-1. Takie kryterium obowiązuje w szeroko rozpowszechnionej metodzie wyznaczania globalnego podobieństwa dwóch obrazów akustycznych, znanej pod nazwą „Dynamic Time Warping Technique". Okazuje się, że zarówno pominięcie, jak i uwzględnienie tego warunku ma swoje zalety i wady. Bilans korzyści i strat z tym związanych przemawia jednak za jego pominięciem. Warunek taki z jednej strony gwarantowałby właściwą chronologię fragmentów w kolejnych parach, lecz z drugiej strony ograniczałby możliwość optymalnego doboru fragmentów w przypadkach, gdy do poprzedniej pary wszedł fragment pochodzący z górnego brzegu zakresu wyboru. Wykazać można, że nawet w przypadku dwóch wypowiedzi tego samego wyrazu, identycznych pod względem ilości fragmentów oraz czasowego rozkładu elementów segmentalnych, nie we wszystkich parach podobnych fragmentów obrazów akustycznych tych wypowiedzi są fragmenty o identycznej liczbie porządkowej. Fragment n spektrogramu binarnego jednej wypowiedzi może np. okazać się bardziej podobny do fragmentu n-2 lub n+2 niż do fragmentu n spektrogramu drugiej wypowiedzi. Takie przypadki mogą mieć miejsce również w zakresie stadiów ustalonych, które de facto są jedynie quasiustalonymi.

Kryterium chronologii fragmentów utrudnia osiąganie dobrych skojarzeń fragmentów w pary, gdy spektrogramy należą do wypowiedzi różnych wyrazów, ale stanowi też przeszkodę w optymalnych kojarzeniach fragmentów w przypadku spektrogramów, których podobieństwa należy się spodziewać. W obu przypadkach kryterium to zawęża wycinek spektrogramu, w którym poszukiwany

jest fragment najbardziej podobny do rozpatrywanego fragmentu innego spektrogramu. Uwzględnienie kryterium chronologii fragmentów wymagałoby znacznego rozbudowania algorytmu znajdowania par podobnych fragmentów wzajemnie porównywanych spektrogramów. Realizacja tego algorytmu przy użyciu dostępnych autorowi środków komputerowych byłaby możliwa jedynie kosztem ograniczenia zakresu badań, tzn. przede wszystkim zmniejszenia wielkości słownika rozpoznawanych wyrazów oraz liczby wypowiedzi użytych w testach rozpoznawania.

6.2. Miara podobieństwa fragmentów

Podobieństwo odnośnych fragmentów wzajemnie porównywanych spektrogramów binarnych musi być określone ilościowo. W rozpoznawaniu mowy miary podobieństwa są przeważnie dostosowane do formy, w jakiej reprezentowany jest sygnał mowy. Na przykład najbardziej rozpowszechniona miara Itakury opiera się na wyrażeniu sygnału mowy za pomocą współczynników predykcji liniowej. Przedstawienie sygnału mowy w formie spektrogramu binarnego jest dość szczególne i w zastosowaniu do rozpoznawania mowy rzadko spotykane [36]. Jak już wykazano, widmo binarne używane w tej pracy zawiera jedynie informacje o zakresach częstotliwości, w których występują dominujące składowe sygnału mowy. Zakresy te podane są ciągami liczb binarnych, z których każda odnosi się do określonego pasma częstotliwości. Ciąg jedynek wskazuje na ważność tej części sygnału mowy, która objęta jest zakresem reprezentowanym przez ów ciąg. Założono, że podobieństwo dwóch sygnałów mowy powinno wyrażać się stopniem pokrywania się zakresów widmowych, w których występują dominujące składowe. W przypadku, gdy dźwięki mowy reprezentowane są przez ich widma binarne, podobieństwo tych dźwięków określać może stopień pokrywania się ciągów jedynek w nałożonych na siebie pojedynczych widmach binarnych lub ciągach kilku kolejnych widm binarnych. Na takim założeniu oparto miarę podobieństwa przeznaczoną do porównywania dwóch fragmentów różnych spektrogramów binarnych. Miara ta wyraża się stosunkiem liczby jedynek zgodnie występujących na odpowiadających sobie pozycjach w obu porównywanych fragmentach spektrogramów binarnych, do sumy wszystkich jedynek w obu tych fragmentach. Określa ją wzór:

$$m_p = \frac{2 \sum_{j=1}^{q} \sum_{i=1}^{k} \left[p_{bi}\left(WB_{n+j}(SB_A)\right) \cdot p_{bi}\left(WB_{m+j}(SB_B)\right) \right]}{\sum_{j=1}^{q} \sum_{i=1}^{k} \left[p_{bi}\left(WB_{n+j}(SB_A)\right) + p_{bi}\left(WB_{m+j}(SB_B)\right) \right]} \qquad (6.2)$$

w którym przez $p_{bi}\left(WB_{n+j}(SB_A)\right)$ i $p_{bi}\left(WB_{m+j}(SB_B)\right)$ oznaczono odpowiednio i-te parametry (n+j)-tego widma binarnego spektrogramu A i (m+j)-tego widma binarnego spektrogramu B. Miara m_p przyjmować może wartość w granicach od 0 do 1. Wartość zero oznacza zupełny brak podobieństwa, a wartość jeden pełną identyczność porównywanych parametrów. Na rys. 21 zilustrowano zasadę porównywania fragmentów spektrogramów binarnych, przyjętą w niniejszej rozprawie.

Niewątpliwą zaletą przedstawionej tutaj miary jest jej prostota. Wyliczenie na podstawie wzoru (6.2) podobieństwa dwóch równolicznych ciągów widm binarnych, z których każdy złożony jest z q k-parametrycznych widm, sprowadza się do wykonania następujących działań:

1/ q-krotnego mnożenia logicznego k-tego rzędu liczb dwójkowych,

2/ zliczania jedynek w (3 x q) liczbach dwójkowych również k-tego rzędu,

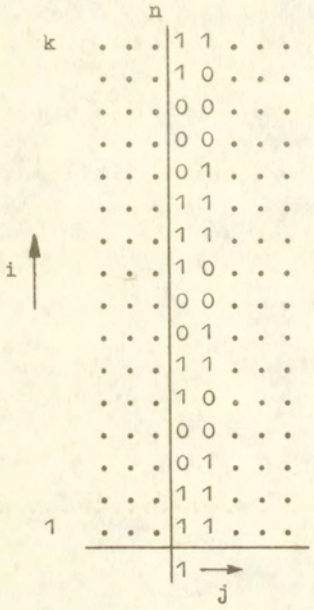
3/ zsumowania (2 x q) liczb mniejszych od (k x q) oraz

4/ jednokrotnego podzielenia.

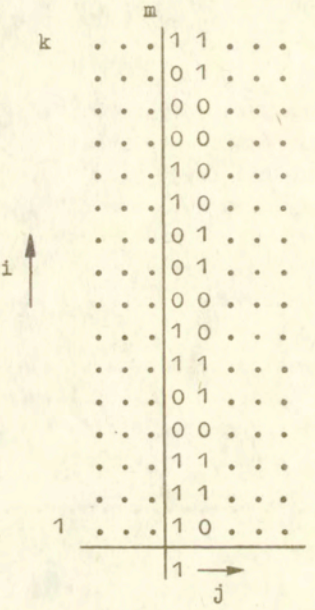
W końcowej wersji modelu ROWBIR 1 k = 16, q = 2. Z wyjątkiem dzielenia wszystkie te operacje są bardzo proste i bardzo krótki jest czas ich wykonywania nawet w przypadku posługiwania się tak prostym komputerem jak MERA 303.

Wyniki rozpoznawania wyrazów, uzyskiwane dotychczas przy zastosowaniu tej miary, świadczą o niej korzystnie. Może ona jednak okazać się niewystarczająca w razie późniejszej adaptacji przedstawionej tutaj metody globalnego rozpoznawania izolowanych wyrazów do rozpoznawania elementów segmentalnych w mowie, takich np. jak połączenia fonemowe. Omówieniu koncepcji takiej adaptacji poświęcony jest jeden z dalszych rozdziałów pracy. Wcześniej zbadano dwie pokrewne wersje wyżej przedsta-

Fragment spektrogramu binarnego A zaczynający się od widma o numerze n + 1



Fragment spektrogramu binarnego B zaczynający się od widma o numerze m + 1



$k = 16, \quad q = 2$

$$\sum_{j=1}^{2} \sum_{i=1}^{16} \left(p_{bi}\, WB_{n+j}(SB_A) \cdot p_{bi}\left(WB_{m+j}(SB_B)\right) \right) = 8$$

$$\sum_{j=1}^{2} \sum_{i=1}^{16} \left[p_{bi}\left(WB_{n+j}(SB_A)\right) + p_{bi}\left(WB_{m+j}(SB_B)\right) \right] = 32$$

$m_p = 0.5$

Rys. 21. Ilustracja zasady porównywania fragmentów spektrogramów binarnych

wionej miary podobieństwa dwóch ciągów widm binarnych. Według pierwszej z nich, wyrażającej się ilorazem:

$$m_p = 2 \frac{\sum(1)_z + \sum(0)_z}{\sum(1) + \sum(0)} \qquad (6.3)$$

o podobieństwie decydują nie tylko zgodnie występujące jedynki, lecz także i zera.

We wzorze (6.3) przez $\sum(0)_z$ i $\sum(1)_z$ oznaczono odpowiednio liczby zgodnie występujących zer i jedynek, a przez $\sum(1)$, $\sum(0)$ sumy wszystkich zer i wszystkich jedynek w obu porównywanych z sobą ciągach widm binarnych.

Zbadano też miarę zbliżoną do poprzedniej, gdyż różniącą się od niej jedynie zwiększeniem roli zgodnych jedynek. Wzór wyrażający tę miarę ma postać:

$$m_p = 2 \frac{k \cdot \sum(1)_z + \sum(0)_z}{\sum(1) + \sum(0)} \qquad (6.4)$$

i różni się od poprzedniego jedynie tym, że zawiera współczynnik wagowy k, przez który pomnożona zostaje suma zgodnych jedynek. Ponieważ suma jedynek i zer w porównywanych fragmentach spektrogramów binarnych jest stała, wyliczenie miary podobieństwa w wersjach podanych wzorami (6.3) i (6.4) nie wymaga dzielenia. Mimo tej korzystnej okoliczności obie wspomniane miary ustąpiły miejsca mierze określonej wzorem (6.2), która okazała się właściwszą przy porównaniach fragmentów o małej liczbie jedynek.

W następnym rozdziale będzie mowa o zgodnych widmach binarnych. Są one elementami zgodnych fragmentów. Wyszukiwanie zgodnych widm binarnych następuje zatem poprzez znajdowanie podobnych fragmentów dwóch wzajemnie porównywanych spektrogramów binarnych. Przyjęto, że gdy fragment składa się z trzech kolejnych widm, wówczas dwa widma są podobne, jeśli zajmują środkowe pozycje w dwóch podobnych fragmentach. Gdy fragment spektrogramu binarnego składa się z dwóch widm, wówczas za podobne uważa się dwa widma, zajmujące identyczne pozycje w dwóch podobnych fragmentach. W skróconej formie regułę powyższą zapi-

sać można następująco:

$$\left.\begin{array}{l} WB_i \sim WB_j, \text{ jeżeli } \left\{WB_{i-1}, \ldots, WB_{i+1}\right\} \sim \left\{WB_{j-1}, \ldots, WB_{j+1}\right\} \\ WB_i \sim WB_j, \text{ jeżeli } \left\{WB_{i-1}, WB_i\right\} \sim \left\{WB_{j-1}, WB_j\right\} \vee \\ \qquad \text{jeżeli } \left\{WB_i, WB_{i+1}\right\} \sim \left\{WB_j, WB_{j+1}\right\} \end{array}\right\} \quad (6.5)$$

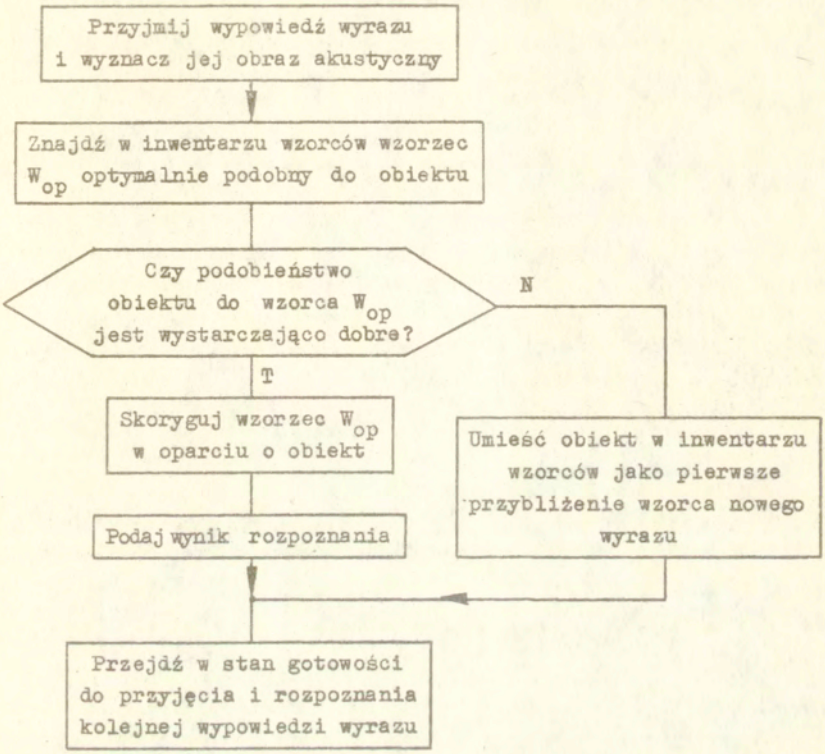
Przez WB_i oznaczano pojedyncze widmo binarne, a przez wyrażenie w klamrze fragment spektrogramu binarnego.

7. Adaptacja i wzorcowe spektrogramy binarne wyrazów

Automatyczne rozpoznawanie wyrazów wymaga znajomości wzorcowych obrazów akustycznych poszczególnych wyrazów. Przygotowanie automatu do rozpoznawania następuje drogą adaptacji (uczenia). Działanie to polega na wyznaczeniu wzorcowych obrazów poszczególnych wyrazów zgodnie z przyjętą formą reprezentacji sygnału mowy. Rozpoznawane być mogą tylko te wyrazy, dla których uprzednio utworzono wzorcowe obrazy akustyczne, nazywane przez autora w skrócie wzorcami. Jeśli dla automatu rozpoznającego reprezentacją wypowiedzi wyrazu jest spektrogram binarny, to wzorce wyrazów wyrażone zostają także w tej formie.

7.1. Adaptacja poprzez rozpoznawanie

W dotychczas spotykanych systemach automatycznego rozpoznawania wyrazów adaptacja jest procesem wstępnym, poprzedzającym właściwy proces rozpoznawania. Idealnym rozwiązaniem byłby taki automat rozpoznający wypowiedzi wyrazów, który równocześnie uczyłby się i rozpoznawał. Procedura jego działania w najogólniejszym ujęciu przedstawiona została na rys. 22. Automat ten próbuje rozpoznać każdą wypowiedź porównując jej obraz z wzorcami różnych wyrazów przechowywanymi w pamięci. Jeśli porównanie to nie wykaże wystarczającego podobieństwa do żadnego z wzorców, wówczas obraz badanej wypowiedzi przyjęty zostanie



Rys. 22. Schemat ogólnej procedury adaptacji w trybie rozpoznawania

do inwentarza wzorców jako zaczątek wzorca nowego wyrazu. Operator winien w takim przypadku podać symbol, którym wyraz ten w późniejszym kontakcie maszyny z człowiekiem ma być etykietowany. Symbol taki może być również nadany w sposób automatyczny z poinformowaniem o tym fakcie operatora. Jeśli wystąpi podobieństwo obrazu wypowiedzi do jednego z wzorców, wówczas automat skoryguje ten wzorzec, wnosząc do niego nowe elementy pochodzące z obrazu bieżącej wypowiedzi, i poinformuje równocześ-

nie o wyniku rozpoznania podając symbol wyrazu, do którego ów wzorzec przynależy.

W znanych systemach automatycznego rozpoznawania wyrazów adaptacja i rozpoznawanie przebiegają oddzielnie, chociaż mają one wiele wspólnych procedur. Model tutaj prezentowany także nie adaptuje się w procesie rozpoznawania, gdyż niemożliwe było osiągnięcie takiej zdolności przy użyciu minikomputera MERA 303. W przypadku zastąpienia tego minikomputera jednym z obecnie dostępnych mikrokomputerów stworzenie takiej zdolności stałoby się wykonalne. Sprzyja temu bowiem okoliczność, że w metodzie rozpoznawania wyrazów, jakiej poświęcona jest niniejsza praca, algorytmy adaptacji i rozpoznawania są w zasadniczych fragmentach identyczne.

7.2. Wyznaczanie wzorcowego spektrogramu binarnego wyrazu

Wzorcowy spektrogram binarny każdego wyrazu utworzony zostaje w oparciu o spektrogramy binarne pewnej liczby wypowiedzi. Wzorcowi nadana zostaje długość jednej z wypowiedzi. Teoretycznie może to być długość którejkolwiek z wypowiedzi tego samego wyrazu, praktycznie zaś wygodnie jest przyjąć jako długość wzorca długość jednej z dwóch pierwszych wypowiedzi. Dla modelu globalnego rozpoznawania wyrazów ROWBIR 1, którego dotyczy niniejsza rozprawa, przyjęto następującą definicję wzorcowego spektrogramu binarnego wyrazu:

> Wzorcowy spektrogram binarny wyrazu jest ciągiem widm binarnych, z których każde zawiera cechy widmowe, jakie wystąpiły w większości wzajemnie odpowiadających sobie widm binarnych spektrogramów wypowiedzi adaptacyjnych.

Definicja ta określa wzorcowy spektrogram binarny obrazu dość ogólnie i dlatego różne rodzaje wzorców mogą być z nią zgodne. Jest rzeczą zrozumiałą, że o reprezentatywności wzorca decyduje sposób jego wyznaczenia. Poszukując odpowiedniego sposobu adaptacji dla modelu rozpoznawania wyrazów w oparciu o spektrogramy binarne autor opracował dwie różne metody oznaczone umownie przez A1 i A2. Każda z nich respektowała wyżej

podaną definicję wzorcowego spektrogramu binarnego.

Metoda A1

Na rys. 23, zamieszczonym na końcu pracy, podano algorytm adaptacji według metody A1. Wyznaczenie wzorca wyrazu w oparciu o tę metodę wymaga uprzedniego zgromadzenia spektrogramów binarnych wszystkich wypowiedzi, na podstawie których wzorzec ma być zbudowany. Spośród zgromadzonych spektrogramów wyselekcjonowany zostaje ten, do którego najwięcej pozostałych spektrogramów wykazuje dobre podobieństwo. Spektrogram ten nazwano umownie centralnym. Przy jego wyborze obowiązuje zasada, według której dwa spektrogramy są podobne, jeżeli wyrażone liczbowo podobieństwo między nimi nie wykracza poza ustalony poziom minimalny. Miarą podobieństwa dwóch wzajemnie porównywanych spektrogramów binarnych może być średnia z podobieństw lokalnych, do wyznaczenia których służy metoda przedstawiona w rozdziale 6.2. W kolejnym etapie, także przy użyciu metody wyznaczania podobieństw lokalnych, wyselekcjonowany uprzednio spektrogram zostaje jeszcze raz porównany z każdym z pozostałych spektrogramów. W trakcie tej operacji utworzone zostają równoliczne zbiory widm binarnych podobnych do poszczególnych widm binarnych uprzednio wyselekcjonowanego spektrogramu. Zbiory te otrzymują numery porządkowe identyczne z numerami porządkowymi widm spektrogramu centralnego. Następnie dla każdego zbioru wyznaczone zostaje wypadkowe widmo binarne. Wartości poszczególnych jego parametrów są liczebnie przeważającymi wartościami tychże samych parametrów w obrębie wszystkich elementów zbioru. Oznacza to, że jeśli w większości widm binarnych zbioru dany parametr posiada wartość 1, to ten sam parametr w widmie wypadkowym otrzymuje także wartość 1. Ciąg widm wypadkowych ułożonych w kolejności zgodnej z kolejnością zbiorów, z których widma te pochodzą, przyjmuje się za wzorcowe widmo binarne.

Przedstawiona metoda adaptacji uwzględnia różne wymogi ważne dla uzyskania właściwego wzorca. Dopuszcza ona do udziału w tworzeniu wzorca dowolną liczbę wypowiedzi wyrazu. Dzięki zastosowaniu w niej selekcji statystycznej istnieje gwarancja, że do wzorca nie przedostaną się przypadkowe lub sporadyczne

cechy widmowe, które z różnych powodów znalazły się w spektrogramach binarnych niektórych wypowiedzi adaptacyjnych. Działania adaptacyjne według tej metody składają się z prostych operacji. Mimo wielu korzystnych zalet tej metody, z przyczyn wyłącznie technicznych nie została ona jednak zastosowana w opisanym tutaj modelu rozpoznawania wyrazów.

Metoda A2

Dla modelu ROWBIR 1 opracowano uproszczoną metodę adaptacji, dającą się łatwo zrealizować przy zastosowaniu komputerowych środków dostępnych na etapie wykonywania niniejszej pracy. Metodę tę oznaczono umownie symbolem A2. Wzorcowy spektrogram binarny wyrazu uzyskany metodą A2 jest także zgodny z wyżej podaną definicją. Rys. 24 zamieszczony na końcu pracy przedstawia schemat algorytmu adaptacji według tej metody.

Spektrogramy binarne wyrazu, przeznaczone do zbudowania wzorca, pogrupowane zostają w pary. Parę tworzą spektrogramy binarne dwóch kolejnych wypowiedzi, przy czym każdy spektrogram może wchodzić tylko w skład jednej pary. Przyjęto, że cechy widmowe, które występują przynajmniej w jednym ze spektrogramów każdej pary, powinny znaleźć się we wzorcu. Oznacza to, że wzorzec przejmuje cechy widmowe, które pojawiły się w spektrogramach binarnych trzech, a w niektórych przypadkach jedynie dwóch wypowiedzi na każde kolejne cztery. Innymi słowy - cechy widmowe, które wystąpią już w 75 procentach, a w pewnych przypadkach nawet tylko w 50% spektrogramów binarnych wypowiedzi adaptacyjnych, przeniesione zostaną do wzorca. Realizacja tego rodzaju adaptacji jest znacznie prostsza niż wyznaczanie wzorca metodą A1.

Spektrogramy SB_{v1} i SB_{v2}, tworzące pierwszą parę, są lokalnie porównywane w sposób przedstawiony w rozdziale 6.2. Indeks v przy symbolu spektrogramu SB_{vi} oznacza numer porządkowy wyrazu, a indeksy liczbowe i są numerami kolejnymi wypowiedzi adaptacyjnych. W wyniku porównania, dla poszczególnych fragmentów jednego spektrogramu zostają znalezione podobne fragmenty w drugim spektrogramie.

Przyjmuje się, zgodnie z tym co powiedziano wcześniej, że

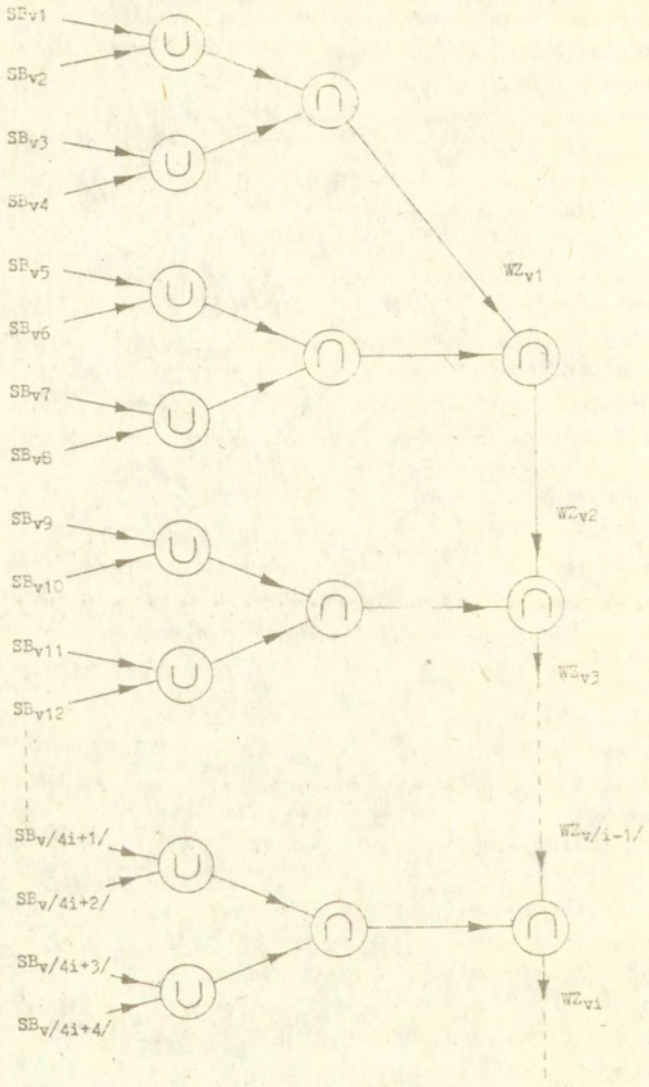
jeśli podobne są dwa fragmenty: $FR_n(SB_{vi})$ i $FR_m(SB_{v(i+1)})$, to podobne są także odpowiednie widma binarne $WB_j \in (FR_n(SB_{vi}))$ i $WB_j \in (FR_m(SB_{v(i+1)}))$ należące do tych fragmentów. Dla adaptacji wybiera się tylko po jednym widmie binarnym z każdego fragmentu. Parę widm pobranych z pary podobnych fragmentów poddaje się logicznemu sumowaniu, którego wynik jest jednym z widm wzorca pośredniego, utworzonego z pary spektrogramów binarnych dwóch kolejnych wypowiedzi. Według tej procedury zostają wyznaczone wszystkie widma wzorca pośredniego. W analogiczny sposób, z pary wzorców pośrednich, uzyskanych z dwóch kolejnych par spektrogramów binarnych wypowiedzi, utworzony zostaje przybliżony wzorzec. Kompilacja wzorców pośrednich polega jednak nie na logicznym sumowaniu, lecz mnożeniu podobnych widm. Dla następnych dwóch par wypowiedzi można także utworzyć przybliżony wzorzec. Kompilując wzorce przybliżone, podobnie jak uczyniono to z wzorcem pośrednim w trakcie wyznaczania wzorca przybliżonego, otrzymuje się następną wersję wzorca przybliżonego. Opisane kompilacje prowadzące do uzyskania wzorca v-tego wyrazu przedstawić można w sposób ogólny następującym wzorem:

$$WZ_v = \bigcap_{i=0}^{n} \bigcap_{j=0}^{1} \bigcup_{k=1}^{2} SB_{v(4i+2j+k)}, \qquad (7.1)$$

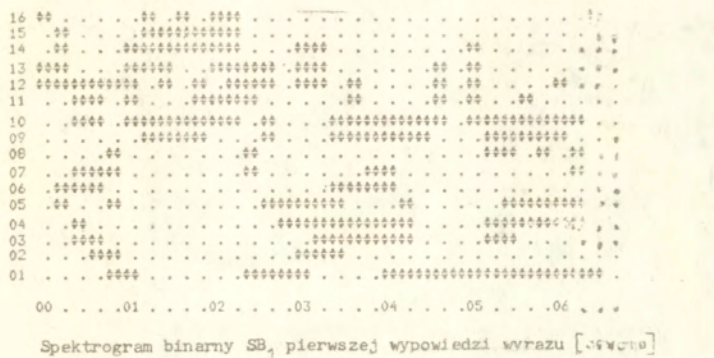
w którym symbole $\bigcup$ i $\bigcap$ oznaczają odpowiednio sumę i iloczyn logiczny, WZ_v wzorzec v-tego wyrazu, $SB_{v(\,)}$ spektrogramy binarne kolejnych wypowiedzi v-tego wyrazu.

Liczba wypowiedzi, na podstawie których utworzony zostaje wzorzec, musi być całkowitą krotnością 4. Krotność ta oznaczona jest we wzorze (7.1) przez n. Na rys. 25 zamieszczono graf działań wyrażonych wzorem (7.1).

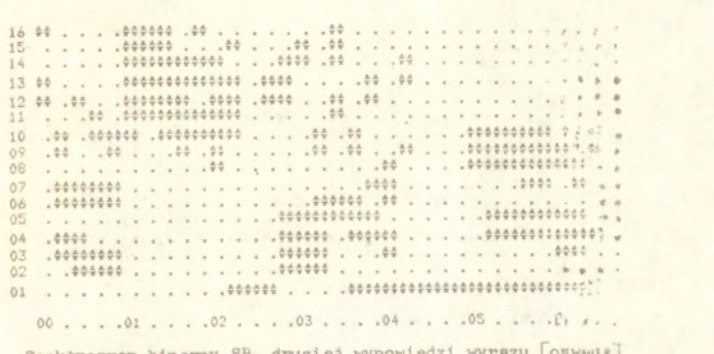
Dla modelu ROWBIR 1, realizującego globalne rozpoznawanie wyrazów w oparciu o spektrogramy binarne, przyjęto n = 1. W modelu tym wzorzec wyrazu powstaje więc na podstawie tylko czterech wypowiedzi. Na rys. 26 przedstawiono przykład powstawania wzorca wyrazu [oswona] z czterech spektrogramów różnych wypowiedzi tego wyrazu jednym głosem.



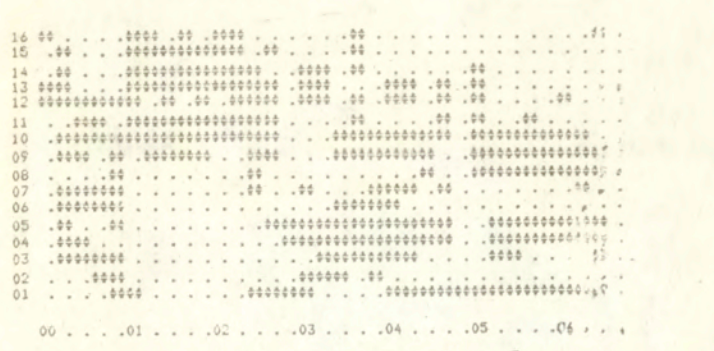
Rys. 25. Graf działań wyrażonych wzorem 7.1.



Spektrogram binarny SB_1 pierwszej wypowiedzi wyrazu [oswona]

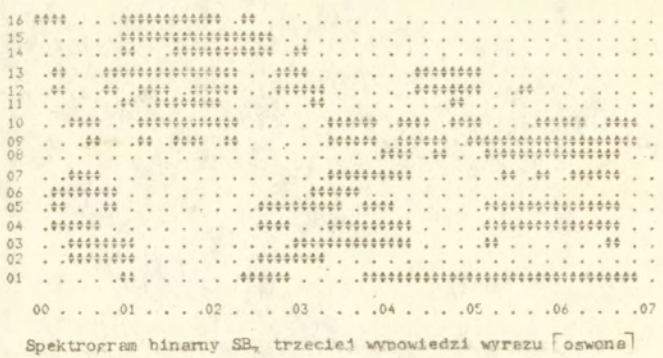


Spektrogram binarny SB_2 drugiej wypowiedzi wyrazu [oswona]

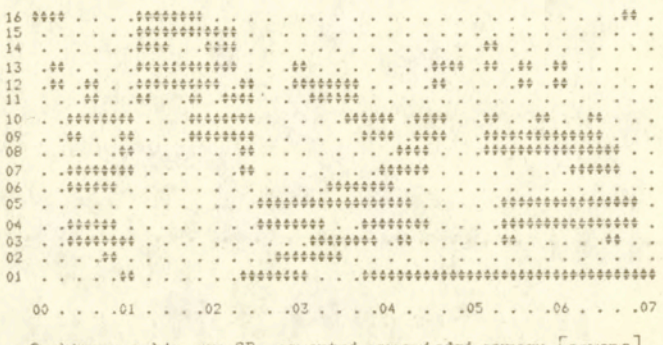


Pierwszy wzorzec pośredni WZP_1 wyrazu [oswona]

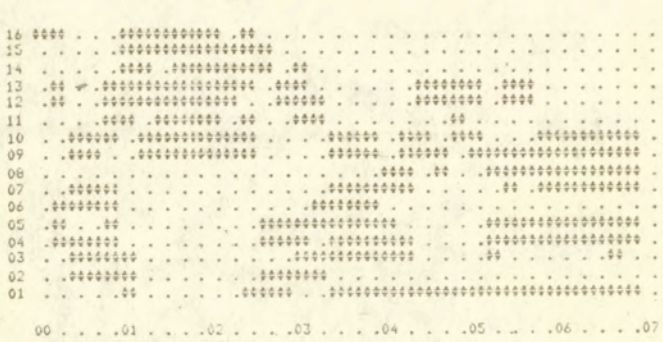
Rys. 26 a) Przykład powstawania pierwszego wzorca pośredniego WZP_1 wyrazu [oswona]



Spektrogram binarny SB_3 trzeciej wypowiedzi wyrazu [oswona]

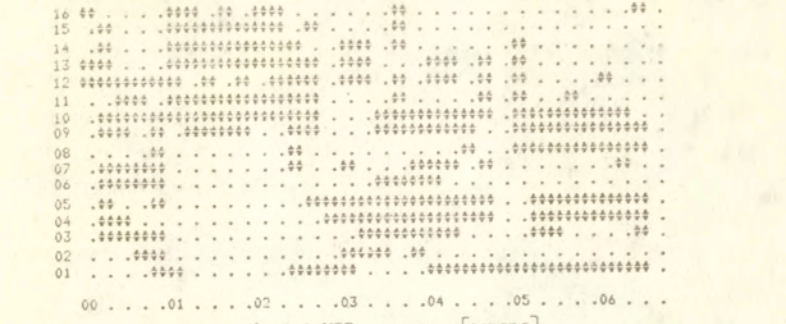


Spektrogram binarny SB_4 czwartej wypowiedzi wyrazu [oswona]

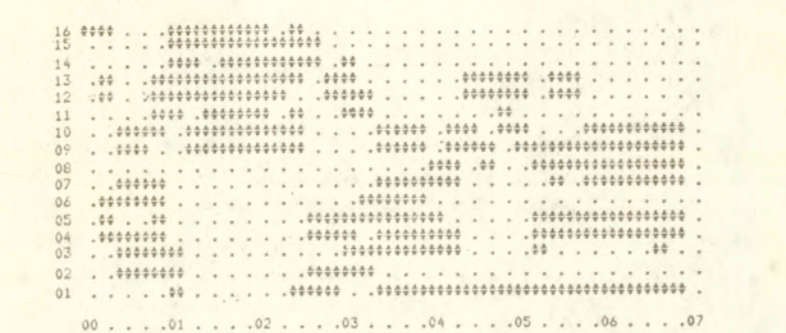


Drugi wzorzec pośredni WZP_2 wyrazu [oswona]

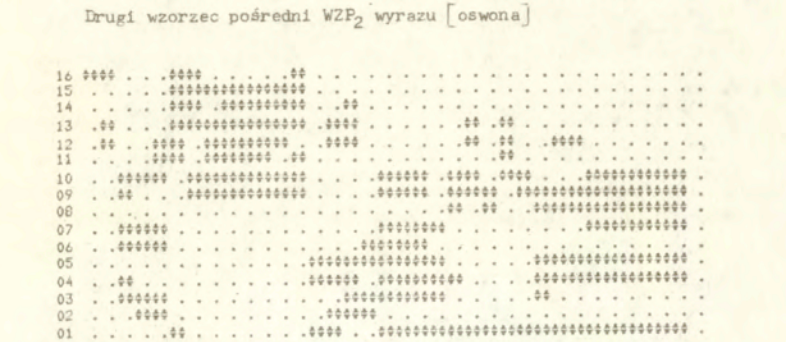
Rys. 26 b) Przykład powstawania drugiego wzorca pośredniego WZP_2 wyrazu [oswona]



Pierwszy wzorzec pośredni WZP_1 wyrazu [oswona]



Drugi wzorzec pośredni WZP_2 wyrazu [oswona]



Wzorzec WZ wyrazu [oswona]

Rys. 26 c) Przykład utworzenia wzorca przybliżonego WZ wyrazu [oswona] na podstawie dwóch wzorców pośrednich: WZP_1 i WZP_2

7.3. Techniczne szczegóły adaptacji w modelu ROWBIR 1

Wyznaczenie wzorcowego spektrogramu binarnego wyrazu przez model ROWBIR 1 jest całkowicie zautomatyzowane i mimo bardzo skromnych możliwości operacyjnych minikomputera MERA 303 przebiega w czasie zbliżonym do rzeczywistego. Wzorzec gotowy jest po upływie kilku sekund od momentu zakończenia czwartej wypowiedzi adaptacyjnej.

Jeśli operator czuwający nad przebiegiem adaptacji nie życzy sobie kontroli obrazów spektrograficznych poszczególnych wypowiedzi, czy też obrazów wzorców pośrednich i wzorca przybliżonego, pomiędzy kolejnymi wypowiedziami adaptacyjnymi tego samego wyrazu wystarczają krótkie przerwy, mniejsze niż 5 sekund. W adaptacji rutynowej kontrola taka nie jest na ogół konieczna. Zwykle korzysta się z niej w eksperymentach badawczych zmierzających np. do określenia czynników mogących wpływać ujemnie na jakość wzorca, lub w przypadkach, gdy wątpliwa jest jakość wypowiedzi przeznaczonej do utworzenia wzorca. Obrazy wypowiedzi, dla których wyznaczony zostaje wzorzec pośredni, mogą być wyświetlane na ekranie oscyloskopu z poświatą równocześnie, w celu przeprowadzenia ich subiektywnej oceny porównawczej. Podobnie w późniejszej fazie procesu adaptacji na tym samym ekranie mogą być wyświetlane równocześnie obrazy dwóch kolejnych wzorców pośrednich jednego wyrazu, a następnie rezultat adaptacji, czyli wzorzec wyrazu utworzony na podstawie czterech wypowiedzi. Operator posiadający podstawową wiedzę empiryczną na temat struktur widmowych poszczególnych elementów fonetycznych mowy potrafi ocenić jakość wzorca wyrazu, jeśli dysponuje jego obrazem. Przyjęto jako zasadę, że dla każdego wyrazu produkt adaptacji podlega tego rodzaju ocenie i od jej pozytywnego wyniku uzależnia się akceptację wzorca.

Wzorzec nie budzący zastrzeżeń zostaje dołączony do ciągu wzorców innych wyrazów już zapisanych w pamięci komputera łącznie z etykietą, którą stanowi kod ortograficzny wyrazu. Mimo, iż spektrogram binarny jest nadzwyczaj zwięzłą reprezentacją wypowiedzi wyrazu, w pamięci operacyjnej minikomputera MERA 303 zmieścić można maksymalnie zaledwie 26 wzorcowych spektrogramów binarnych o widmach 16-parametrycznych, a więc znacznie

skomprymowanych w porównaniu z widmami 63-parametrycznymi stosowanymi pierwotnie. Gdy liczebność słownika rozpoznawanych wyrazów jest większa niż 26, konieczne jest przechowywanie wzorców wyrazów poza pamięcią operacyjną minikomputera MERA 303. W trakcie rozpoznawania wyrazu wzorce są przesyłane z pamięci zewnętrznej do pamięci operacyjnej minikomputera w grupach po 26 wzorców. Każda taka transmisja wzorców jest dodatkową czynnością w procesie rozpoznawania wypowiedzi wyrazu i czas jej trwania powiększa łączny czas oczekiwania na wynik rozpoznawania.

7.4. Modyfikacja inwentarza wzorców i wzorce wspólne

W celu uniknięcia konieczności uwzględniania wszystkich wzorców w procesie rozpoznawania należałoby wzorce poszczególnych wyrazów połączyć w grupy według stopnia ich wzajemnego podobieństwa. Każdą grupę reprezentowałby jeden ze wzorców do niej należących, wyłoniony według zasady MINIMAX, zgodnie z którą podobieństwo tego wzorca do najmniej podobnych jemu elementów grupy jest największe w porównaniu z innymi wzorcami grupy rozpatrywanymi w podobny sposób.

Osobną grupę wzorców stanowiliby reprezentanci poszczególnych grup. Przy takim układzie wzorców można by uniknąć konieczności transmisji wszystkich wzorców z pamięci zewnętrznej do pamięci operacyjnej minikomputera w czasie rozpoznawania wyrazu i tym samym skrócić czas oczekiwania na wynik rozpoznania. Połączenie wzorców w grupy według określonego kryterium podobieństwa byłoby celowe i pożyteczne nie tylko w przypadku, gdy inwentarz wzorców musi być przechowywany w pamięci zewnętrznej ze względu na małą pamięć operacyjną minikomputera, lecz także wówczas, gdy pamięć operacyjna jest wprawdzie wystarczająco duża, ale równocześnie słownik rozpoznawanych wyrazów jest dość obszerny. Bowiem przy tego rodzaju rozmieszczeniu wzorców można się spodziewać w ogóle znacznego skrócenia czasu rozpoznawania wyrazu. Realność takiej konstrukcji inwentarza wzorców wymaga potwierdzenia w drodze doświadczeń, które należałoby poprzedzić właściwym zdefiniowaniem podobieństwa dwóch wzorców różnych wyrazów. Jak wiadomo, podobieństwo takie może być jedynie fragmentaryczne, a zatem należy się spodziewać, że nie-

które wzorce wystąpią w kilku różnych grupach wzorców. Zaproponowany tutaj układ inwentarza wzorców nie został uwzględniony w modelu ROWBIR 1, gdyż pierwszoplanowym celem relacjonowanych badań było stwierdzenie, czy możliwe jest poprawne rozpoznawanie wyrazów w sposób globalny, posługując się zdefiniowaną przez autora binarną reprezentacją widmową sygnału mowy. Zagadnienie optymalizacji struktury inwentarza wzorców zalicza się zaś do sfery usprawnienia rozpoznawania wyrazów metodą już doświadczalnie zweryfikowaną i zaakceptowaną, i jako takie zostanie zbadane w drugiej kolejności.

Symboliczne przedstawienie wzorców

grupy 1	grupy 2
a b x z	v b x z
a b w z	a c x y
a b x y	a b x y

Rys. 27. Ilustracja przynależności wzorca do dwóch grup

Na rys. 27 zilustrowano symbolicznie dwie grupy wzorców oraz ich reprezentantów. Literami oznaczono elementy wzorców, które odpowiadać mogą np. ciągom fonemów lub półsylabom. Wzorzec abxy należy do obu grup i z tego powodu nie może on reprezentować każdej z tych grup. W inwentarzu wzorców występuje on dwukrotnie. Także wzorce będące reprezentantami grup występują dwukrotnie: w swojej grupie i w grupie reprezentantów. Zatem pogrupowanie wzorców na zasadzie skupienia ich wokół swoich reprezentantów, z którymi łączy ich częściowe podobieństwo, prowadzi do poszerzenia inwentarza wzorców. Opłacone to jednak może zostać znacznym skróceniem czasu oczekiwania na wynik rozpoznania wyrazu.

Przeprowadzone zostały między innymi próby użycia modelu

ROWBIR 1 do rozpoznawania wyrazów wymawianych przez dwa głosy mające wspólne wzorce. Wyniki tych prób podane są w jednym z dalszych rozdziałów. Wzorzec wyrazu wspólny dla pary głosów tworzono na podstawie dwóch wypowiedzi jednym głosem i dwóch drugim. W sposób wyżej opisany z każdej pary wypowiedzi tym samym głosem wyznaczano wzorzec pośredni, zaś z wzorców pośrednich dotyczących różnych głosów tworzono wzorzec wspólny. Tego rodzaju kompilacja cech widmowych różnych głosów jest bardzo uproszczona. Wybrano ją, gdyż zamiarem było uzyskanie jedynie wstępnej informacji, czy cechy osobnicze poszczególnych głosów pozwalają na uzyskanie wspólnego wzorcowego spektrogramu binarnego wyrazu. Problem tworzenia wzorców wspólnych dla dowolnej populacji głosów jest w ogóle bardzo złożony, a dla przypadku, gdy wzorzec ma formę spektrogramu binarnego, nie był dotychczas szerzej rozpatrywany i będzie niebawem przedmiotem osobnych studiów.

8. Rozpoznawanie wyrazów w sposób globalny w oparciu o spektrogramy binarne

8.1. Procedura rozpoznawania

W pierwszym etapie procesu rozpoznawania wyrazu wypowiedź zostaje wyrażona w formie spektrogramu binarnego w sposób wyżej przedstawiony. Spektrogram rozpoznawanego wyrazu jest nieznanym obiektem, który należy zidentyfikować.

Identyfikacja wymaga konfrontacji obiektu z wzorcami poszczególnych wyrazów, uzyskanymi w wyniku uprzednio przeprowadzonej adaptacji modelu rozpoznającego. Konfrontacja obiektu z jednym wzorcem ma na celu określenie stopnia wzajemnego podobieństwa obiektu i tego wzorca i polega na porównaniu dwóch spektrogramów binarnych. Przebieg tego porównania jest identyczny, jak podczas adaptacji. Dla poszczególnych fragmentów jednego spektrogramu (obiektu) poszukiwane są fragmenty najbardziej do nich podobne w drugim spektrogramie (wzorcu). Zgodnie z tym, co powiedziano wcześniej i co zilustrowano na rys. 20, obszar tych poszukiwań jest zawężony do przedziałów: $(m - \Delta, m + \Delta)$, obejmujących najbliższe otoczenie wartości funkcji $m =$

= f(n) normalizującej quasiliniowo długości obu spektrogramów. Ocenę podobieństwa dwóch fragmentów różnych spektrogramów binarnych przeprowadza się przy użyciu miary podobieństwa wyrażonej wzorem (6.2).

Na rys. 28 przedstawiono macierz wartości podobieństw q_{mn} każdego fragmentu spektrogramu binarnego SB_A z każdym fragmentem spektrogramu binarnego SB_B. Podobieństwa q_{mn} są uzupełnieniami do 1 podobieństw wyliczonych ze wzoru (6.2) i wyrażone zostały liczbami ósemkowymi z przesuniętym o dwa miejsca w prawo przecinkiem. W macierz wrysowano krzywą schodkową quasiliniowej normalizacji długości obu spektrogramów oraz zaznaczono obszar obejmujący najbliższe otoczenie tej krzywej. W obszarze tym przypadają podobieństwa najlepiej odpowiadających sobie fragmentów porównywanych spektrogramów binarnych. Liczby wyrażające wartości podobieństw tych fragmentów zostały podkreślone. Numery porządkowe fragmentów wykazujących najlepsze wzajemne podobieństwo uznać można za współrzędne wyznaczające funkcję optymalnej normalizacji długości porównywanych spektrogramów pod warunkiem, że spełniona jest nierówność:

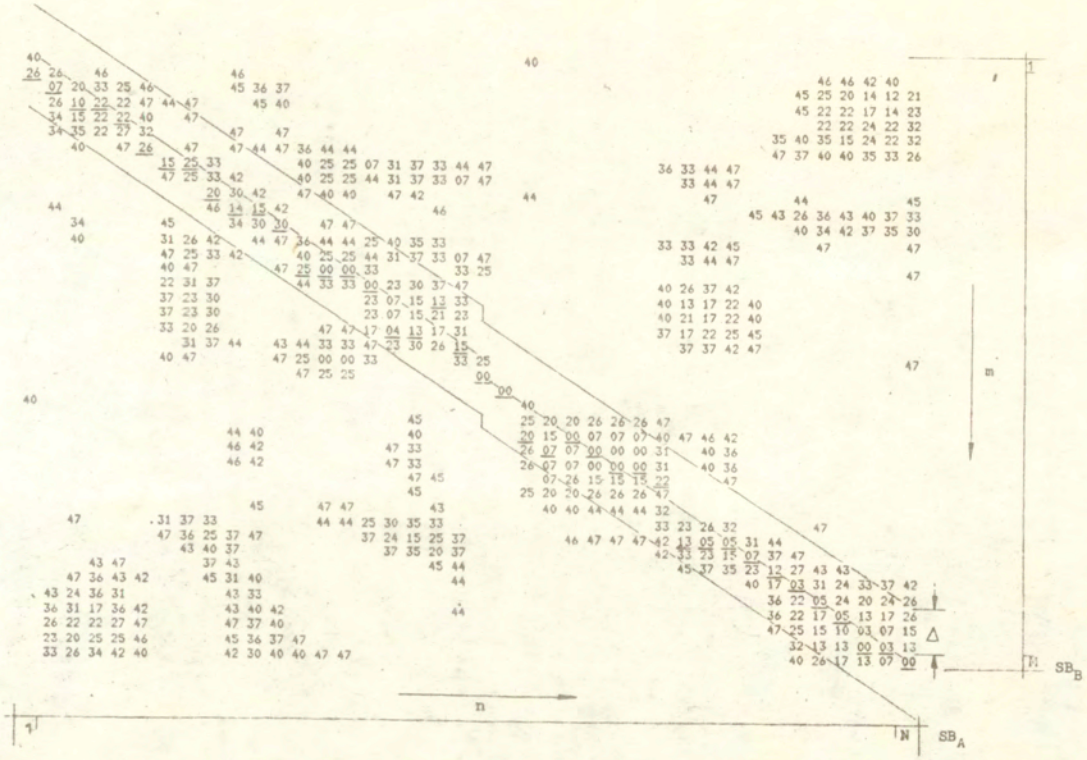
$$m(k) - m(k+1) \geqslant 0 \qquad (8.1)$$

oraz, że porównywane spektrogramy należą do różnych wypowiedzi tego samego wyrazu.

W rozdziale poświęconym porównywaniu spektrogramów binarnych podano, że mogą zaistnieć przypadki nie spełniające warunku (8.1). W przedstawionej tutaj metodzie rozpoznawania wyrazów dopuszcza się, iż w ciągu par $\{P_k[Fr_n(SB_A), Fr_m(SB_B)]\}$ wzajemnie podobnych fragmentów dwóch porównywanych spektrogramów binarnych numer porządkowy m = i(k) jednego z fragmentów może dla pewnych k spełniać zależność:

$$i(k) - i(k-6) < 0, \qquad (8.2)$$

zamiast ze wzrostem numeru k pary fragmentów utrzymywać się w tendencji wzrostowej lub przynajmniej nie zmieniać się. Ze względów oczywistych numer porządkowy drugiego fragmentu zawsze spełnia warunek:



Rys. 28. Macierz wartości podobieństw fragmentów porownywanych spektrogramów binarnych

$$j(k) - j(k-1) \geqslant 0, \qquad (8.3)$$

przy czym $j(k) = n$ wyraża zależność pomiędzy kolejnością k par fragmentów i kolejnością n fragmentów jednego z porównywanych spektrogramów binarnych.

Wynikiem porównania dwóch spektrogramów binarnych, w tym przypadku obiektu z jednym ze wzorców, jest ciąg:

$$Q_{vN} = \left\{ q_{v1}, \ldots, q_{vi}, \ldots, q_{vN} \right\}, \qquad (8.4)$$

wyrażający podobieństwa q_{vi} kolejnych fragmentów obiektu i najbardziej podobnych do nich fragmentów v-tego wzorca, wybranych w sposób wyżej opisany spośród fragmentów skupionych w otoczeniu trajektorii quasiliniowej normalizacji długości obu spektrogramów. Na rys. 28 ciąg ten tworzą podkreślone elementy macierzy.

Pożyteczną informacją dodatkową, pomocną dla rozpoznania obiektu, jest ciąg:

$$D_{vN} = \left\{ d_{v1}, \ldots, d_{vi}, \ldots, d_{vN} \right\}, \qquad (8.5)$$

którego poszczególne wyrazy d_{vi} są odchyleniami numerów porządkowych wybranych fragmentów wzorca od numerów porządkowych odnośnych fragmentów wzorca, leżących na trajektorii quasiliniowej normalizacji długości porównywanych spektrogramów.

Rozpoznanie wyrazu sprowadza się do znalezienia jego wzorca, który z założenia powinien być najbardziej podobny do obiektu. Procedurę działań z tym związanych ilustruje schemat algorytmu przedstawiony na rys. 29 (na końcu pracy). Znalezienie wzorca najbardziej podobnego do obiektu przebiega drogą eliminacji. Jeśli kolejny wzorzec W_v wykazuje lepsze podobieństwo do obiektu niż wzorzec najbardziej podobny do obiektu spośród wzorców $W_1 \ldots W_{v-1}$, v-ty wyraz staje się nowym kandydatem do wyniku rozpoznawania, w miejsce kandydata dotychczasowego rekrutującego się z części słownika obejmującego wyrazy od pierwszego do (v-1)-tego. Innymi słowy, v-ty wyraz pretenduje do roli wyniku identyfikacji obiektu, jeżeli spełniony jest wa-

runek:

$$P(OB, W_v) \geqslant P_r(OB, W_{v-r}) \quad \text{dla} \quad r \in \{1, \ldots, v-1\} \tag{8.6}$$

Przez $P(OB, W_i)$ oznaczono podobieństwo spektrogramu binarnego obiektu i wzorcowego spektrogramu binarnego i-tego wyrazu, wyrażone przez ciąg podobieństw lokalnych poszczególnych fragmentów obiektu z wybranymi fragmentami wzorca W_i i-tego wyrazu. Przyjęto zasadę, że obiekt jest bardziej podobny do tego z dwóch wzorców, do którego jest on bardziej podobny w większej liczbie fragmentów. Precyzyjniej tę zasadę można sformułować następująco:

O tym, który z dwóch wzorców W_{v-r} i W_v jest bardziej podobny do obiektu OB, decydują liczebności dwóch klas do jakich zaliczone zostały elementy ciągu:

$$\Delta_{v(v-r)N} = \left\{(q_{v1} - q_{(v-r)1}), \ldots\ldots, (q_{vN} - q_{(v-r)N})\right\}, \tag{8.7}$$

różnic lokalnych podobieństw q_{vi} i $q_{(v-r)i}$ obiektu OB do wzorców W_v i W_{v-r}. Podstawą klasyfikacji jest tu kryterium wyrażone następującymi nierównościami:

$$\left.\begin{aligned} q_{vi} - q_{(v-r)i} &\geqslant \Delta\,(q_{(v-r)i}) \\ \left|q_{vi} - q_{(v-r)i}\right| &< \Delta\,(q_{(v-r)i}) \end{aligned}\right\} \tag{8.8}$$

We wzorze tym $\Delta(q_{(v-r)i})$ oznacza nieistotną różnicę między lokalnymi podobieństwami q_{vi} i $q_{(v-r)i}$, która powinna być parabolicznie rosnącą funkcją $q_{(v-r)i}$, czyli wartości podobieństwa i-tego fragmentu obiektu i odpowiadającego mu fragmentu (v-r)-tego wzorca, najbardziej podobnego do obiektu spośród wzorców (W_1 .... W_{v-1}). Elementy ciągu (8.7), spełniające tylko pierwszą z nierówności (8.8), zaliczone zostają do klasy KL_{v-r} (lepsze podobieństwo lokalne obiektu w miejscu i do wzorca W_{v-r}). Elementy tego ciągu nie spełniające obu nierówności zaliczone zostają do klasy KL_v (lepsze podobieństwo lokalne obiektu w miejscu i do kolejnego wzorca W_v). Elementy spełnia-

jące tylko nierówność drugą trafiają do obu klas.

Jeśli do klasy KL_v zaliczonych zostało więcej elementów niż do klasy KL_{v-r}, oznacza to, że większe jest podobieństwo obiektu do wzorca W_v niż do wzorca W_{v-r}. Gdy proporcja liczebności elementów w klasach KL_v i KL_{v-r} wypadła odwrotnie, świadczy to że obiekt pozostaje bardziej podobny do wzorca W_{v-r}. W przypadku jednakowej lub zbliżonej liczebności obu klas, o tym, do którego z dwóch wzorców obiekt jest bardziej podobny rozstrzygnąć może porównanie wartości sum elementów ciągów $D_{(v-r)N}$ i D_{vN}, czyli:

$$\sum_{i=1}^{N} |d_{vi}|$$

oraz

$$\sum_{i=1}^{N} |d_{(v-r)i}|.$$

Suma mniejsza wskazuje na lepsze podobieństwo obiektu i wzorca. Gdy ma miejsce przypadek, że:

$$\sum_{i=1}^{N} |d_{vi}| < \sum_{i=1}^{N} |d_{(v-r)i}|, \qquad (8.9)$$

wzorzec W_v uznać można za bardziej podobny do obiektu niż wzorzec $W_{(v-r)}$.

Na rys. 30 przedstawiono macierze wartości podobieństw kolejnych fragmentów obiektu do wybranych fragmentów dwóch różnych wzorców. Pod macierzami wypisane są ciągi Q_{vN}, $Q_{(v-r)N}$, $-Sign(\Delta_{v(v-r)N})$, D_{vN}, $D_{(v-r)N}$. Elementy ciągów Q_{vN} oraz $Q_{(v-r)N}$ zaznaczono w macierzach podkreśleniem. Elementy ciągów D_{vN} i $D_{(v-r)N}$ są pionowymi odległościami podkreślonych elementów macierzy od trajektorii quasiliniowej normalizacji długości obiektu i wzorca. Odległości te mierzone są liczbą odstępów pomiędzy wierszami, przez które przechodzi wspomniana trajektoria, i w których przypadają podkreślone elementy macierzy. Przynależność elementów ciągu $\Delta_{v(v-r)N}$ do jednej z dwóch klas zaznaczono znakami + i -. Znaki + dotyczą klasy, której liczebność wyraża podobieństwo wzorca W_v do obiektu. Przewaga

ilości elementów w jednej z klas rozstrzyga o tym, który z dwu rozpatrywanych wzorców jest bardziej podobny do obiektu. Na rys. 30 jest nim wzorzec WZ_{v-r}. Uzyskany tą drogą werdykt zostaje potwierdzony w wyniku porównania sum elementów ciągów $D_{(v-r)N}$ i D_{vN}. Suma elementów ciągu D_{v-r} jest znacznie mniejsza. Ze względu na małe możliwości obliczeniowe minikomputera MERA 303 nie użyto procedury typowania kandydata do wyniku w oparciu o ciągi $D_{(v-r)N}$ i D_{vN}.

Klasyfikowanie elementów ciągu $\Delta_{v(v-r)N}$ następuje w trakcie porównywania obiektu z kolejnym wzorcem W_v. Jeżeli stawka na korzyść aktualnie rozpatrywanego wzorca, wyrażająca się liczebnością L_v jednej z klas, przyrasta niewiele i przewaga liczebności L_{v-r} drugiej klasy staje się znaczna, przerywa się dalsze porównywanie obiektu i wzorca na j-tym fragmencie obiektu na tyle odległym od jego końca, że nawet pełna zgodność jego pozostałych N-j fragmentów z odnośnymi fragmentami rozpatrywanego wzorca nie spowodowałaby naruszenia powstałej przewagi liczebności jednej z klas, przemawiającej za utrzymaniem dotychczasowego kandydata do wyniku rozpoznawania.

Numer fragmentu obiektu, na którym w takim przypadku można przerwać dalsze porównywanie obiektu z rozpatrywanym wzorcem, wynika z zależności:

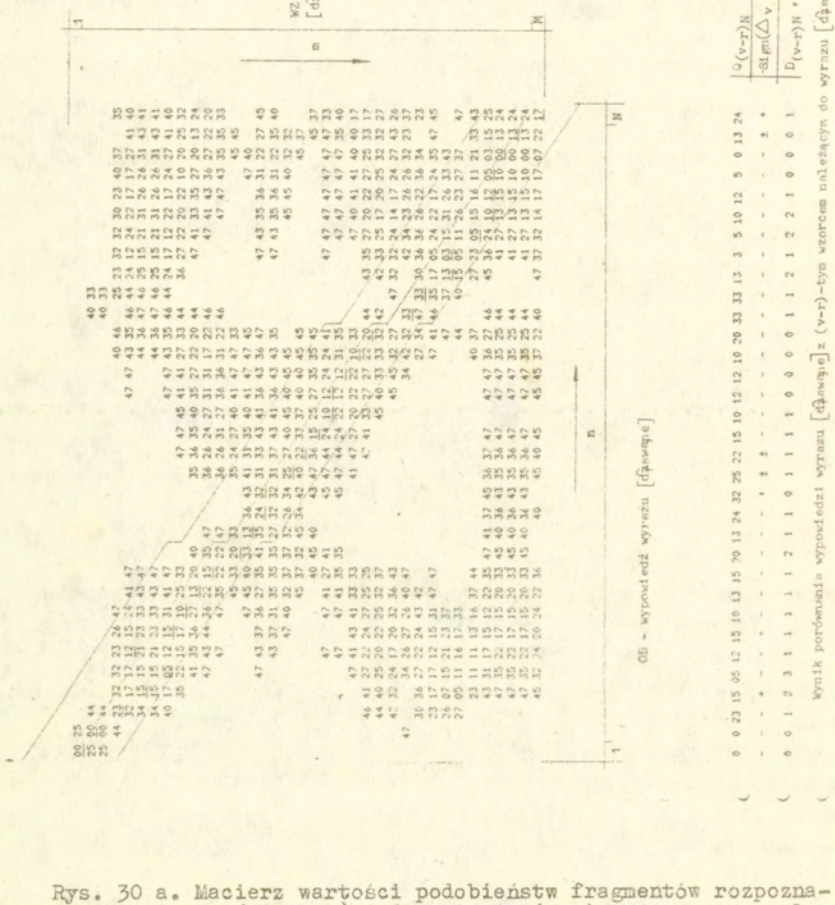
$$j \geqslant L_v - L_{v-r} + N + \delta , \qquad (8.10)$$

gdzie N jest liczbą wszystkich fragmentów rozpatrywanego obiektu, a δ wymaganą przewagą liczebności jednej z klas.

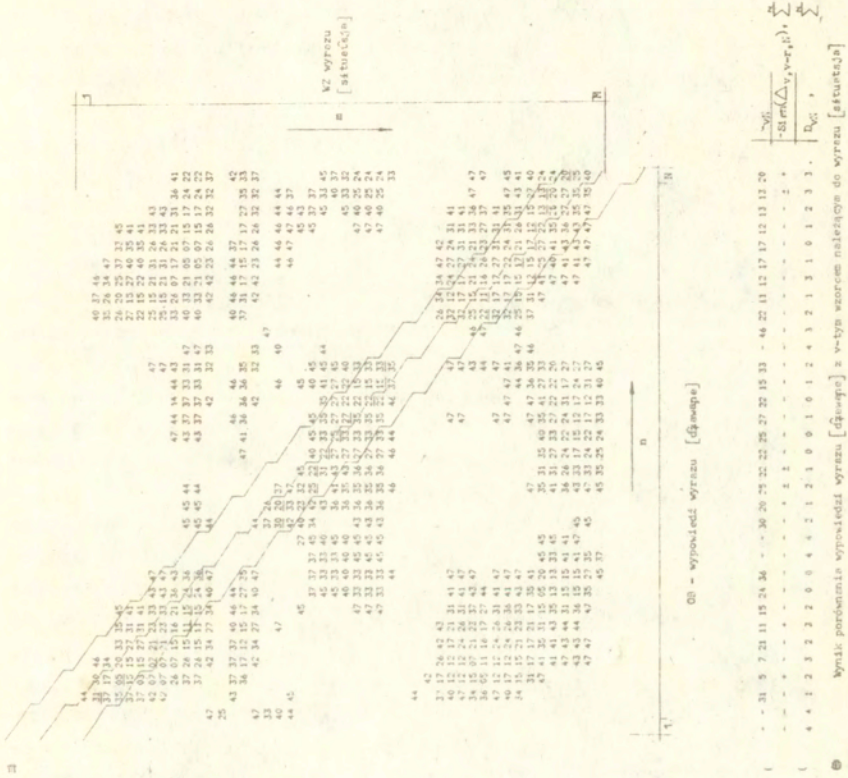
Przerwanie porównywania obiektu z wzorcem następuje także w przypadku odwrotnym, tzn. wówczas, gdy wystąpi znaczna przewaga liczebności tej klasy, która wyraża kwalifikację aktualnie rozpatrywanego wzorca do przyjęcia nowego kandydata do wyniku rozpoznania. Porównanie obiektu ze wzorcem przerywa się wtedy na fragmencie obiektu, którego numer określa, analogiczna jak poprzednio, zależność:

$$j \geqslant L_{v-r} - L_v + N + \delta . \qquad (8.11)$$

Inny sposób uniknięcia zbędnych operacji w procesie rozpozna-



Rys. 30 a. Macierz wartości podobieństw fragmentów rozpoznawanego wyrazu (obiektu) i fragmentów (v-r)-tego wzorca należącego do tego wyrazu oraz tabela elementów ciągów: $Q_{(v-r)N}$, $-Sign\,(\Delta_{v(v-r)N})$, $D_{(v-r)N}$



Rys. 30 b. Macierz wartości podobieństw fragmentów rozpoznawanego wyrazu (obiektu) i fragmentów v-tego wzorca nie należącego do tego wyrazu oraz tabela elementów ciągów: Q_{vN}, -Sign $(\Delta_{v(v-r)N})$, D_{vN}

wania wyrazu polega na nieporównywaniu obiektu z tymi wzorcami, których długość znacznie odbiega od długości obiektu. Można przyjąć, że wzorzec i obiekt dotyczą różnych wyrazów, gdy długość obiektu jest co najmniej dwukrotnie większa lub mniejsza od długości wzorca. Zbyteczne jest w takich przypadkach porównywanie obiektu z wzorcem.

W modelu ROWBIR 1 wzorce wyrazów ułożone są w pamięci w porządku alfabetycznym. Rozpoznanie wyrazu następuje przez porównanie jego spektrogramu binarnego kolejno ze wszystkimi wzorcami bez względu na to, którą pozycję w inwentarzu wzorców zajmuje wzorzec rozpoznawanego wyrazu. Tryb i czas rozpoznawania są w przybliżeniu takie same zarówno wtedy, gdy rozpoznawany wyraz ma swój wzorzec na pierwszym miejscu, jak i wówczas, gdy jego wzorzec jest w inwentarzu wzorców ostatnim. Rozpoznawanie w takim trybie określić można jako szeregowe. Dla uproszczenia i przyśpieszenia procesu rozpoznawania tryb szeregowy można by ograniczyć jedynie do początku obiektu w celu wyłonienia wpierw potencjalnych kandydatów do wyniku. Spośród nich następnie, poprzez rozpoznawanie globalne, zostałby uzyskany wynik ostateczny. W ten sposób w dość czasochłonnym rozpoznawaniu globalnym brałoby udział przypuszczalnie zaledwie parę wzorców.

Dysponując inwentarzem wzorców w układzie grupowym, który zaproponowano w rozdziale poświęconym adaptacji, można by zastosować tzw. rozpoznawanie kaskadowe. Obiekt byłby wówczas wpierw porównywany z reprezentantami grup podobnych wzorców. Po zidentyfikowaniu reprezentanta najbardziej podobnego do obiektu, dalsze rozpoznawanie sprowadziłoby się do porównania obiektu jedynie z wzorcami wskazanymi wcześniej przez wybranego reprezentanta. Poszukiwanie wzorca najbardziej podobnego do obiektu byłoby więc ukierunkowane, dzięki czemu czas oczekiwania na wynik rozpoznawania uległby znacznemu skróceniu w porównaniu z trwaniem rozpoznawania szeregowego.

Skrócenie czasu rozpoznawania uzyskuje się nie tylko dzięki wyżej omówionym udoskonaleniom metodycznym. Można je także osiągnąć stosując układy wyspecjalizowane. Sprzyja tej tendencji prostota działań, składających się na operacje rozpoznawania wyrazu zaprezentowaną tutaj metodą. Jak wiadomo, są to w przeważającej części proste działania logiczne. W oparciu o

technikę mikroprocesorową można by stworzyć układ rozpoznawania równoległego. Układ taki umożliwiałby porównywanie obiektu z kilkoma wzorcami naraz, dzięki czemu kilkakrotnie krótszy byłby czas oczekiwania na wynik rozpoznania.

8.2. Próby testowe kolejnych wersji modelu ROWBIR 1

Metoda głobalnego rozpoznawania wyrazów na podstawie spektrogramów binarnych, której poświęcono niniejszą pracę, przeszła wielokrotne weryfikacje w drodze różnych prób testowych. Wyniki kolejnych prób okazywały się zgodne z oczekiwaniami, a pod niektórymi względami nawet je przekraczały w sensie dodatnim. Poszczególne testy dotyczyły różnych wersji parametrycznych modelu rozpoznającego, a także różnych słowników i głosów. Cele i zakres kolejnych prób wyznaczały często wyniki prób wcześniejszych. Na zakres prób znaczny wpływ wywarły też zmienne w czasie warunki techniczne, w jakich badania były przeprowadzone. Niedostatek środków technicznych zmuszał do różnych modyfikacji modelu rozpoznającego. Modyfikacje te okazywały się nieraz bardzo pożyteczne. Pierwsze próby opisane w pracy autora [83] z roku 1981 miały na celu ukazanie wartości metody rozpoznawania wyrazów na podstawie spektrogramów binarnych. Nie dysponowano bowiem wtedy szerszą wiedzą na ten temat. Różne przesłanki wynikające z doświadczenia w zakresie fonetyki akustycznej pozwalały rokować, że metoda ta powinna okazać się skuteczna i opłacalna, i dlatego warta rozwijania. Badania, o których mowa, miały to przypuszczenie potwierdzić lub obalić.

Jakość rozpoznawania oceniano na podstawie wyrażonego w procentach ilorazu poprawności rozpoznawania. Brano pod uwagę przede wszystkim tzw. globalną poprawność rozpoznawania, tzn. procentowy stosunek poprawnie rozpoznanych wypowiedzi do wszystkich wypowiedzi biorących udział w teście. Czasem rozpatrywano także tzw. wyrazową poprawność rozpoznawania, czyli stosunek poprawnie rozpoznanych wypowiedzi tylko jednego wyrazu do wszystkich wypowiedzi tego wyrazu w teście.

8.2.1. Próby testowe pierwszej wersji modelu ROWBIR 1

W pierwszych próbach model rozpoznający cechowały następujące parametry: Widmo binarne było 43 parametryczne, gdyż tyle kanałów miał wówczas analogowy analizator widma, będący w modelu podstawowym instrumentem analizy akustycznej sygnału mowy. Zakres analizy wynosił od 120 do 3560 Hz. Bliższe szczegóły na temat tego analizatora zawiera rozdział 3.2. Funkcja wygładzająca widmo sygnału mowy była prostokątna, przez co nie uzyskiwano dostatecznego wygładzenia w przypadkach głosów o wyższych częstotliwościach podstawowych.

Uwzględniano wszystkie parametry widma binarnego. Fragmenty, na jakie dzielono porównywane spektrogramy binarne dla wyznaczenia lokalnych podobieństw między nimi, obejmowały po 3 kolejne widma binarne.

Warunek lepszego podobieństwa fragmentów, stanowiący podstawę oceny globalnych podobieństw porównywanych spektrogramów binarnych, nie uwzględniał zmiennego przedziału równego podobieństwa fragmentów. Innymi słowy $\triangle(q_{(v-r)i})$ we wzorach (8.8) było liczbą stałą. W trakcie pierwszych prób dokonano też wyboru najwłaściwszego, spośród trzech postulowanych, wariantu miary wyrażającej podobieństwo fragmentów porównywanych spektrogramów binarnych.

Słownik rozpoznawanych wyrazów składał się z zaledwie 13 słów, którymi były polskie nazwy dziesięciu cyfr oraz wyrazy: kropka, przecinek, kreska. Ten zestaw nie jest zbyt łatwy z punktu widzenia globalnego rozpoznawania wyrazów, gdyż występują w nim wyrazy stosunkowo krótkie, głównie jedno- i dwusylabowe. Wyrazy takie rozpoznaje się w sposób globalny najtrudniej, gdyż ich obrazy akustyczne tworzą krótkie ciągi widm. W próbach brał udział jeden głos męski. W teście rozpoznawania poszczególne wyrazy słownika wymawiane były w porządku losowym, każdy wyraz słownika 20-krotnie. Przy takiej ilości materiału testowego nierozpoznanie jednej wypowiedzi obniżało wartość ilorazu wyrazowej poprawności rozpoznawania o 5%, a ilorazu globalnej poprawności o 0,4%. Uzyskane wyniki rozpoznawania zamieszczone są w tablicy 1. Na 13 wyrazów słownika 8 roz-

poznawanych było ze 100% poprawnością, a pozostałe 5 z poprawnością w granicach 85-95%. Globalna poprawność rozpoznawania wyniosła 96%.

Tablica 1. Wyniki testu rozpoznawania 13 wyrazów dla pierwszej wersji modelu ROWBIR 1

R \ N	0	1	2	3	4	5	6	7	8	9	.	,	-
0	20												
1		17						3					
2			18	2									
3				20									
4		1			18			1					
5						20							
6							20						
7		1						19					
8									20				
9										20			
.											18		2
,												20	
-													20

Wyniki te pozwoliły stwierdzić, że metoda rozpoznawania wyrazów na podstawie spektrogramów binarnych zasługuje na akceptację i warta jest dalszych badań. Okazało się, że metoda ta przy swojej prostocie daje dobre rezultaty. Stosunkowo nieliczne błędy, jakie wystąpiły w rozpoznawaniu, spowodowane zostały po części zawężeniem zakresu analizy akustycznej do częstotliwości 3560 Hz, a po części przez pewne uchybienia w wygładzaniu widma dyskretnego i następnie przekształcaniu go w

widmo binarne. Pewne błędy wytłumaczono wystąpieniem wyjątkowych zjawisk artykulacyjnych przy wypowiedziach niektórych wyrazów.

8.2.2. Próby testowe drugiej wersji modelu ROWBIR 1

Następne próby przedstawiono w pracy H.KUBZDELI (1982). Przeprowadzono je już w odmiennych warunkach technicznych. Zastosowano w nich bowiem 63-kanałowy analizator widma o zakresie analizy poszerzonym do częstotliwości 8310 Hz, udoskonalony układ logarytmujący dane wyjściowe z analizatora oraz urządzenie pamięci zewnętrznej na dyskach elastycznych. O ile w pierwszych próbach wyznaczenie spektrogramu binarnego przebiegało równocześnie z trwaniem wypowiedzi wyrazu, o tyle w drugiej serii prób nie dało się tego osiągnąć ze względu na powiększoną liczbę parametrów widmowych. Dane widmowe, napływające z analizatora widma, po odpowiednim uśrednieniu umieszczano w pamięci operacyjnej minikomputera MERA 303, która pierwotnie przewidziana była wyłącznie do przechowywania wzorców. Wyznaczenie spektrogramu binarnego następowało dopiero po zakończeniu wypowiedzi. Spektrogramy binarne zarówno wypowiedzi adaptacyjnych, jak i testowych (przeznaczonych do rozpoznania) zgromadzono w pamięci zewnętrznej na dyskach elastycznych. Tam też po nie sięgano podczas adaptacji i rozpoznawania. Celem drugiej serii prób było uzyskanie poglądu na efektywność rozpoznawania wyrazów metodą spektrogramów binarnych w zmienionych warunkach, tzn. po rozszerzeniu zakresu analizy widmowej i wprowadzeniu kilku dodatkowych udoskonaleń technicznych oraz dla zwiększonego słownika wyrazów. Pod wpływem trudności, jakie podczas prób wynikały z niedostatku pamięci operacyjnej i ograniczonych możliwości obliczeniowych minikomputera MERA 303, zbadano też, w jakim stopniu zmniejszenie o połowę liczby parametrów widma binarnego poprzez wyeliminowanie parametrów o numerach parzystych, a także zawężenie do dwóch widm szerokości fragmentu rozpatrywanego przy wyznaczaniu podobieństw lokalnych porównywanych spektrogramów pogarsza wyniki rozpoznawania. W tej serii prób automatycznego rozpoznawania wyrazów posłużono się słownikiem wyrazowym złożonym z nazw 48 miast polskich, głównie wojewódzkich. Poniżej zamieszczono ich listę,

na której zapisane są one w tej samej kolejności, w jakiej zostały uszeregowane ich wzorce. Testy rozpoznawania przeprowadzono dla trzech głosów, dwóch męskich i jednego żeńskiego, przy zastosowaniu ich indywidualnych wzorców.

Słownik wyrazów użytych w drugiej serii prób automatycznego rozpoznawania

1. Gniezno	13. Kraków	25. Częstochowa	37. Sieradz
2. Gdynia	14. Krosno	26. Elbląg	38. Wadowice
3. Kutno	15. Ostrołęka	27. Gdańsk	39. Słupsk
4. Jarocin	16. Piła	28. Gorzów	40. Suwałki
5. Kościan	17. Płock	29. Legnica	41. Szczecin
6. Warszawa	18. Poznań	30. Leszno	42. Tarnobrzeg
7. Białystok	19. Przemyśl	31. Lublin	43. Tarnów
8. Kalisz	20. Radom	32. Łomża	44. Toruń
9. Katowice	21. Rzeszów	33. Łódź	45. Wałbrzych
10. Konin	22. Chełm	34. Opole	46. Wrocław
11. Kielce	23. Bydgoszcz	35. Olsztyn	47. Włocławek
12. Koszalin	24. Ciechanów	36. Siedlce	48. Zamość

Wzorcowy spektrogram binarny każdego z wyrazów 48-wyrazowego słownika był zatem utworzony oddzielnie dla każdego głosu na podstawie szterech wypowiedzi. Rozpoznawaniu poddano co najmniej ośmiokrotne wypowiedzi każdego z wyrazów wybranego słownika przez każdy z trzech głosów. Dla każdego głosu próba testowa składała się z co najmniej 392 wypowiedzi. Testy rozpoznawania przeprowadzono w trzech wariantach. Testy w ramach tego samego wariantu określa się też przez "eksperyment".

Pierwszy wariant dotyczył rozpoznawania wypowiedzi wyrazów tylko jednego głosu przy 63-parametrycznej reprezentacji widmowej sygnału mowy i z zachowaniem dotychczasowej szerokości fragmentu porównawczego, a więc obejmującej 3 kolejne widma. Test rozpoznawania w tym wariancie służył sprawdzeniu efektywności rozpoznawania w nowych warunkach, a mianowicie po rozszerzeniu zakresu częstotliwości analizy widmowej i powiększe-

niu słownika rozpoznawanych wyrazów.

Drugi wariant charakteryzował się tym, że reprezentację wypowiedzi stanowił spektrogram binarny 32-parametryczny zamiast jak dotychczas 63. Jak już wspomniano, redukcję ilości parametrów uzyskano w drodze wyeliminowania z pierwotnego widma binarnego parametrów o parzystych numerach porządkowych. W teście według tego wariantu brał udział również tylko jeden głos męski.

W trzecim wariancie testu rozpoznawania zawężono fragment porównawczy z trzech do dwóch kolejnych widm, zachowując równocześnie 32-parametryczną reprezentację widmową. W próbach testowych według tego wariantu brały udział 3 głosy, 2 męskie i 1 żeński.

Wyniki rozpoznawania wyrazów uzyskane w poszczególnych testach przedstawione są poniżej w tablicach 2-4. Tablica 2 zawiera wartości globalnej poprawności rozpoznawania oraz ilości wyrazów słownika rozpoznanych bezbłędnie i z różnymi liczbami błędów uzyskane w poszczególnych testach rozpoznawania. Tablica 3 stanowi wykaz błędnych rozpoznań w każdej z trzech wyżej opisanych prób rozpoznawania. Tablica 4 dotyczy wyłącznie wariantu trzeciego i zawiera wykaz ilości błędów popełnionych w rozpoznawaniu niektórych wyrazów wypowiedzianych przez niektóre głosy.

Tablica 2. Wyniki rozpoznawania uzyskane w eksperymentach 1, 2, 3

Eksperyment	Globalna poprawność rozpoznawania %	Liczebność haseł słownika rozpoznanych				
		bez błędu	z 1 błędem	z 2 błędami	z 3 błędami	z 4 i więcej błędami
1	96	42	2	1	2	1
2	97.25	42	3	2	-	1
3	96.8	42	4	-	-	2
	96.9	41	3	3	1	-
	97.1	41	4	2	-	1

Tablica 3. Wykaz błędnych odpowiedzi w eksperymentach 1, 2, 3

Eksperyment 1	
Nadano	Odebrano
1. Kraków	Tarnów
2. Kraków	Tarnów
3. Kraków	Ciechanów
4. Kutno	Poznań
5. Kutno	Leczno
6. Kutno	Krosno
7. Kutno	Krosno
8. Kutno	Krosno
9. Kutno	Krosno
10. Tarnów	Kraków
11. Kościan	Olsztyn
12. Kościan	Olsztyn
13. Słupsk	Suwałki
14. Wadowice	Katowice
15. Wadowice	Katowice
16. Wadowice	Katowice

Eksperyment 2	
Nadano	Odebrano
1. Kraków	Ciechanów
2. Kraków	Tarnów
3. Kutno	Łomża
4. Kutno	Łomża
5. Kutno	Leszno
6. Kutno	Krosno
7. Tarnów	Kraków
8. Słupsk	Suwałki
9. Poznań	Koszalin
10. Wadowice	Katowice
11. Wadowice	Katowice

Eksperyment 3

Głos M1	
Nadano	Odebrano
1.Kutno	Poznań
2.Kutno	Krosno
3.Kutno	Krosno
4.Kutno	Łomża
5.Kutno	Leszno
6.Kraków	Suwałki
7.Tarnów	Kraków
8.Słupsk	Suwałki
9.Kościan	Leszno
10.Wadowice	Katowice
11.Wadowice	Katowice
12.Wadowice	Katowice
13.Wadowice	Katowice

Głos M2	
Nadano	Odebrano
1. Toruń	Konin
2.Wadowice	Zamość
3.Wadowice	Katowice
4.Wadowice	Katowice
5.Tarnów	Kalisz
6.Tarnów	Sieradz
7.Szczecin	Konin
8.Szczecin	Przemyśl
9.Wrocław	Krosno
10.Siedlce	Sieradz
11.Kraków	Radom
12.Kraków	Radom

Głos Ż1	
Nadano	Odebrano
1.Słupsk	Siedlce
2.Ciechanów	Rzeszów
3.Toruń	Gorzów
4.Toruń	Rzeszów
5.Kraków	Tarnów
6.Kraków	Tarnów
7.Białystok	Gniezno
8.Białystok	Gniezno
9.Białystok	Wrocław
10.Białystok	Wałbrzych
11.Jarocin	Tarnobrzeg
12.Łomża	Łódź

Tablica 4. Zestaw ilości błędów dla poszczególnych wyrazów i głosów

Hasło	Głosy		
	M1	M2	Ż1
Słupsk	1	-	1
Ciechanów	-	-	1
Toruń	-	1	2
Kraków	1	2	2
Białystok	-	-	4
Łomża	-	-	1
Jarocin	-	-	1
Wadowice	4	3	-
Tarnów	1	2	-
Szczecin	-	2	-
Wrocław	-	1	-
Siedlce	-	1	-
Kutno	5	-	-
Kościan	1	-	-
R a z e m	13	12	12

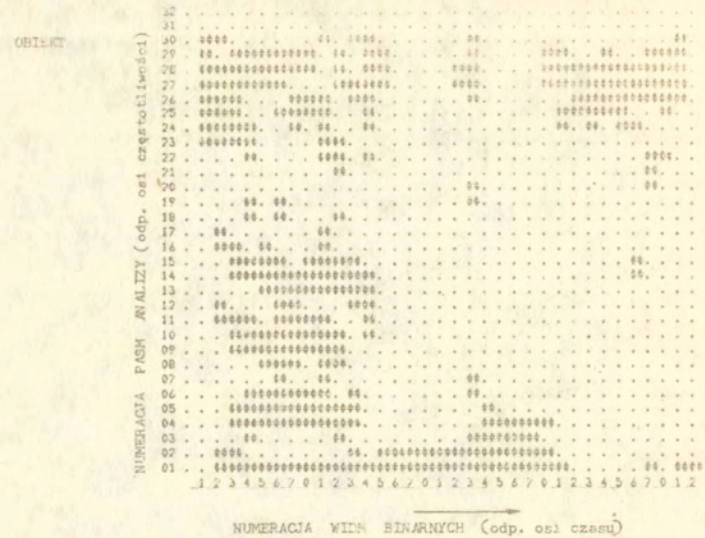
Z analizy wyników uzyskanych w tej serii prób wypływają następujące wnioski:

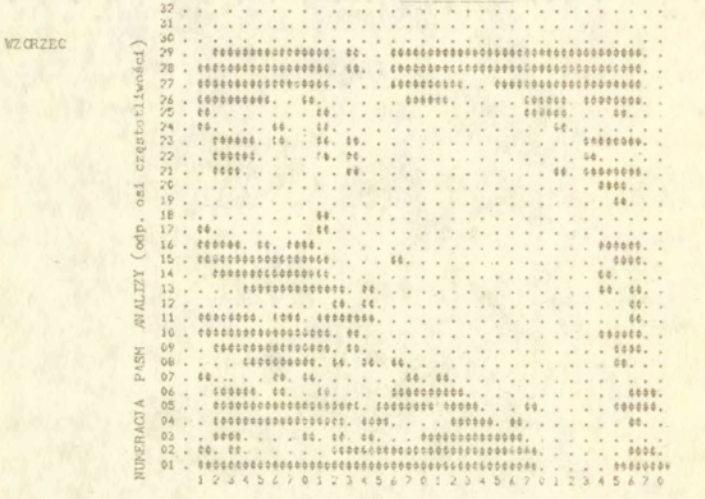
1. 32-widmowe parametry binarne, opisujące sygnał mowy w zakresie częstotliwości od 80 do 8310 Hz, wystarczają do niemal jednoznacznego wyrażenia w formie spektrogramu binarnego każdego z 48 różnych wyrazów wymówionych przez jeden głos. Zredukowanie liczby parametrów binarnych o połowę nie pogorszyło jakości rozpoznawania.
2. Podstawowy fragment spektrogramu, którym operuje się w procesie wzajemnego porównywania dwóch spektrogramów binarnych, może składać się z dwóch zamiast z trzech kolejnych widm binarnych.
3. Zbieżność wyników uzyskanych dla trzech różnych i przypadkowo dobranych głosów świadczy korzystnie o przydat-

ności użytej metody rozpoznawania wyrazów dla różnych głosów. Dla każdego z głosów, które wzięły udział w testach, otrzymano dobre i zbliżone wyniki rozpoznawania. Potwierdzona została w ten sposób pozytywna opinia o rozpoznawaniu wyrazów na podstawie spektrogramów binarnych, wyrażona po testach przeprowadzonych dla pierwszej wersji modelu rozpoznającego.

Analizując przypadki błędnych rozpoznań ustalono kilka rodzajów ich przyczyn. Jedną z nich były zbyt duże różnice w treści widmowej odpowiadających sobie fragmentów obiektu i jego wzorca. Na rys. 31 zamieszczono dla przykładu zestaw obiektu i wzorca wyrazu [tarnuf]. Różnią się te dwa 32-parametryczne spektrogramy binarne treścią widmową w obrębie fragmentu [nuf]. Tak znaczne różnice, jak w zacytowanym przykładzie powodują, że niektóre fragmenty obiektu są mniej podobne do odnośnych fragmentów właściwego wzorca niż innych wzorców. Podobne przyczyny błędów występowały w przypadkach, gdy rozpoznawany obiekt i niektóre obce mu wzorce miały zbliżone długości i identyczne fragmenty początkowe lub końcowe. Na przykład zdarzyło się, że końcowy fragment spektrogramu binarnego rozpoznawanego wyrazu [kutno] był bardziej podobny do zakończenia wzorca wyrazu krosno niż wzorca wyrazu [kutno]. Podobnie zdarzyło się większe podobieństwo końcowej części spektrogramu binarnego rozpoznawanego wyrazu [krakuf] z analogicznym zakończeniem wzorca wyrazu [tarnuf], niż z końcówką wzorca wyrazu [krakuf]. Fakty te świadczą o konieczności skorygowania wzorców niektórych wyrazów na podstawie dodatkowych wypowiedzi.

Inne przyczyny, które po części zaważyły na błędach w rozpoznawaniu wyrazów, polegały na zbyt dużych dysproporcjach w rozkładach czasowych odpowiadających sobie segmentów obiektu i dotyczącego go wzorca. Różnice te dotyczyły np. przerw przedplozyjnych, segmentów afrykacji i szumowych. Zauważono przypadki wydłużenia końca wypowiedzi lub braku początku wypowiedzi, gdy wyraz rozpoczynał się od głoski zwartej dźwięcznej lub dźwięcznej trącej. Stwierdzono, że dla uniknięcia błędów powodowanych przez znaczne dysproporcje w rozkładach analogicznych segmentów obiektu i dotyczącego go wzorca należy nieco poszerzyć zakres poszukiwania podobnego fragmentu, który dotychczas obejmował 5 kolejnych fragmentów. Niektóre błędy w rozpo-





Rys. 31. Obiekt i wzorzec wyrazu [tarnuf]

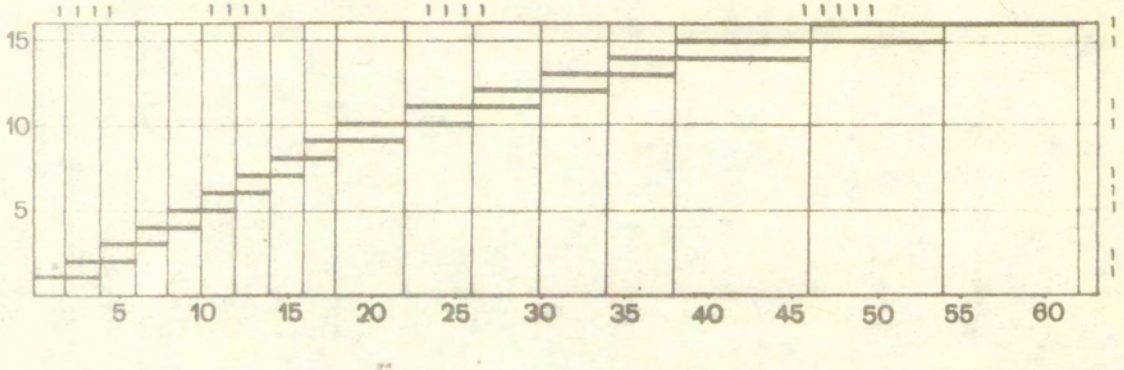
znawaniu pochodziły od zakłóceń, które sporadycznie pojawiły się przed lub po zakończeniu wypowiedzi. Dla jednego wyrazu błędy w rozpoznawaniu wynikały z wadliwego wzorca.

8.2.3. Kompresja widma binarnego oraz próby testowe 3-ciej wersji modelu ROWBIR 1

Wnioski jakie wynikały z drugiej serii prób rozpoznawania wyrazów stały się wskazówką do dalszego zmodyfikowania modelu rozpoznającego. Przede wszystkim uwzględniono uwagę o konieczności poszerzenia zakresu poszukiwania podobnego fragmentu. Na stałe przyjęto krótszą rozciągłość fragmentu. Skoro zmniejszenie liczby binarnych parametrów widmowych z 63 do 32 nie pogorszyło jakości rozpoznawania, należało zbadać, czy możliwa jest dalsza ich redukcja. Odstąpiono jednak od przeprowadzenia jej drogą dalszego pominięcia niektórych parametrów widma binarnego. W zamian zastosowano pewnego rodzaju kompresję widma binarnego pierwotnego, która polega na zastąpieniu odpowiednich ciągów parametrów widma binarnego pierwotnego parametrami pojedynczymi. W wyniku takiego przekształcenia powstaje nowe widmo binarne, którego parametry wyrażają obecność lub brak par kolejnych jedynek w odnośnych pasmach widma binarnego pełnego. Istotny jest podział na owe pasma, chociaż brak jest ścisłych wskazówek, jak należy go przeprowadzić. Zapewne szerokości poszczególnych pasm winny być adekwatne do stopnia dystynktywnej roli, jaką w pierwotnym widmie binarnym odgrywają informacje zawarte w tych pasmach. W tych zakresach widma, w których pojawia się informacja bardzo istotna i jednakowo ważna, pasma te muszą być równe i najwęższe. Odnosi się to przede wszystkim do zakresu występowania pierwszych dwóch formantów samogłoskowych. W pozostałych zakresach widma pasma te mogą być odpowiednio szersze. Odbicie w obrazie widma binarnego znanych cech widmowych sygnału mowy jest zdeterminowane głównie przez rozdzielczość częstotliwością, z jaką wykonana została analiza widmowa. Pełne widmo binarne, uzyskane w oparciu o analizę widmową stosowaną w modelu ROWBIR 1, zawiera swoiste formanty, z których każdy wyrażony jest przeważnie ciągiem trzech do pięciu jedynek bezpośrednio po sobie następujących. Wziąwszy pod uwagę powyższe względy przyjęto, że przekształcenie widma

binarnego w dolnym zakresie opierać się będzie o podział na przedziały obejmujące 4 kolejne parametry widmowe i pokrywające się odpowiednio dwoma parametrami, natomiast dla górnego zakresu widma binarnego przyjęto przedziały o szerokościach 6, 8, 12 i 16 parametrów, zachodzące na siebie odpowiednio dwoma, czterema i ośmioma parametrami. Rezultatem opartego o taki podział przekształcenia 63-parametrycznego pierwotnego widma binarnego jest widmo binarne 16-parametryczne. Zasadę tego rodzaju przekształcenia zilustrowano na rys. 32. Uznano, że tego rodzaju przekształcenie odpowiada dążeniu do zredukowania liczby parametrów widma binarnego i przez to do uproszczenia operacji rozpoznawania i zminimalizowania roli cech indywidualnych głosu w spektrogramie binarnym.

Dla pomniejszenia wpływu cech osobniczych głosu na obraz wyrazu wprowadzono do modelu rozpoznającego dodatkową innowację. Dotyczy ona wyznaczania zredukowanego spektrogramu binarnego i polega na dynamicznym ograniczeniu jego zakresu. Granicę wyznacza się osobno dla poszczególnych widm binarnych, a ustala ją dla danego widma ostatnia z pierwszych n jedynek licząc od początku widma zredukowanego. Przyjęto, że n będzie liczbą stałą, chociaż można by ją uczynić malejącą funkcją numeru porządkowego parametru, z którego pochodzi kolejna jedynka. Dynamiczne ograniczanie zakresu widma dokonuje się w trakcie wyznaczania widma binarnego pierwotnego i jednoczesnego przekształcania go w widmo binarne zredukowane. Przerwanie wyznaczania widma binarnego pierwotnego z powodu naliczenia n jedynek w pochodnym widmie binarnym zredukowanym może nastąpić dopiero na parametrze zerowym, a nie na niezakończonym ciągu jedynek. W związku z tym ilość jedynek w widmie binarnym o dwojako zredukowanej liczbie parametrów może niekiedy przekraczać liczbę n. Na rys. 32 zilustrowano przypadek, w którym z powodu wyżej wymienionego warunku następuje przekroczenie liczby n ustalonej na 6. Dzięki zastosowaniu tzw. dynamicznego ograniczania zakresu widma binarnego niektóre z widm spektrogramu binarnego zredukowane zostają do pewnej stałej liczby jedynek. I tak np. widma binarne samogłosek zostają ograniczone do najistotniejszego zakresu dwóch pierwszych formantów, a pominięty zostaje zakres wyższy, w którym obraz widma jest



Rys. 32. Ilustracja zasady przekształcenia widma binarnego 63-parametrycznego w widmo binarne 16-parametryczne

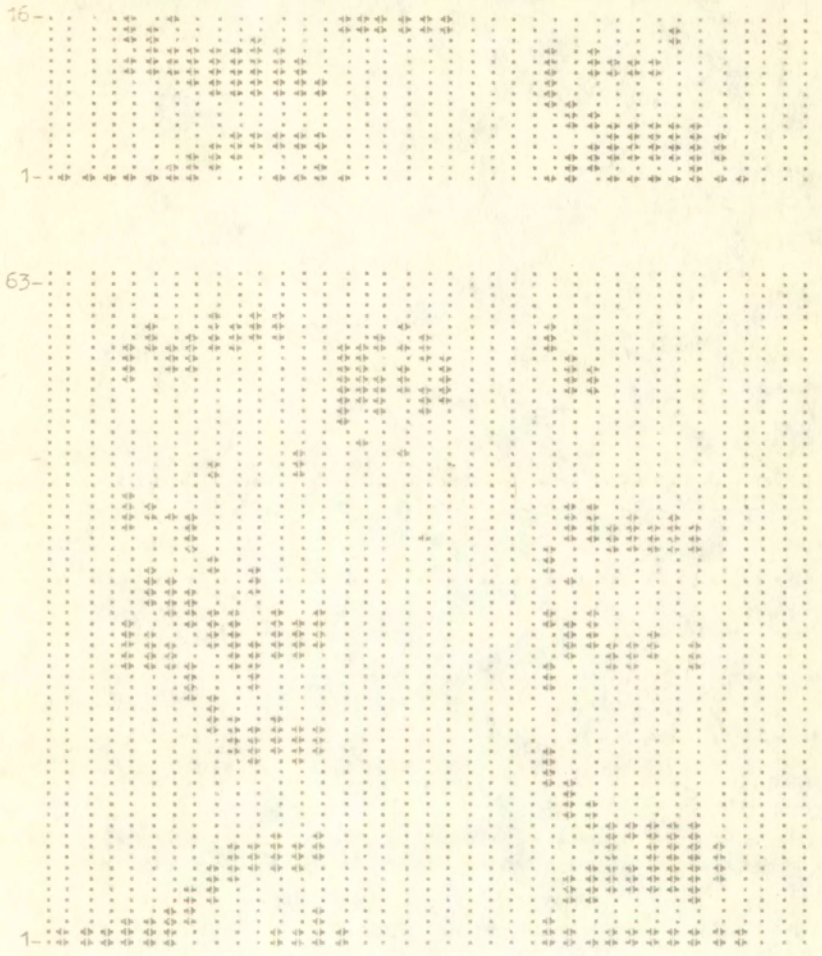
najbardziej osobniczo zindywidualizowany. Natomiast w przypadku spółgłosek trących do ograniczenia widma ze zrozumiałych względów zwykle nie dochodzi. Na rys. 33 zamieszczono spektrogram binarny ze zredukowaną liczbą parametrów oraz dla porównania spektrogram binarny pierwotny.

Posługując się modelem w nowej zmodyfikowanej wersji, przeprowadzono trzecią serię prób rozpoznawania wyrazów. Próby te miały przede wszystkim na celu sprawdzenie, czy zastosowanie nowej zredukowanej reprezentacji wyrazu nie wpłynęło pogarszająco na jakość rozpoznawania. Dzięki znacznemu pomniejszeniu objętości informacyjnej reprezentacji wyrazu powstały też warunki do poszerzenia słownika rozpoznawanych wyrazów. W porównaniu z pełnym spektrogramem binarnym spektrogram zredukowany zajmuje bowiem znacznie mniej miejsca w pamięci komputera. Następnym celem prób była więc ocena zdolności modelu do rozpoznawania dwukrotnie większej liczby wyrazów. W części prób poświęconych wymienionym dwom celom uczestniczył jeden głos męski i użyto słownika złożonego ze stu wyrazów. Wyrazami tego słownika było 100 pierwszych rzeczowników w przypadkowo napotkanym artykule gazetowym, którego tytuł brzmiał: „18 dni twierdzy Modlin", i który ukazał się w lokalnym dzienniku poznańskim „Głos Wielkopolski". Rzeczowniki sprowadzono do mianownika, lecz zachowano ich gramatyczną liczbę, w jakiej zostały użyte w artykule.

Zbadaniu poddano również rozpoznawanie wyrazów na podstawie wzorców wspólnych dla dwóch głosów. W tej części prób brały udział cztery głosy męskie, a słownik rozpoznawanych wyrazów ze względów formalno-technicznych liczył tylko 26 słów. Poniżej zamieszczono oba słowniki.

W pierwszej części prób jeden głos męski wypowiedział każde ze 100 słów słownika pięciokrotnie. Najpierw czterokrotnie każde w kolejności występowania w słowniku, a następnie jednorazowo każde w porządku przypadkowym - łącznie 500 wypowiedzi. Podczas wypowiedzi wyznaczany był spektrogram binarny, a bezpośrednio po jej zakończeniu następowało przepisanie tegoż spektrogramu do pamięci zewnętrznej w celu późniejszego wykorzystania go do adaptacji lub w teście rozpoznawania.

400 spektrogramów uzyskanych w pierwszej turze wypowiedzi



Rys. 33. Spektrogram binarny pierwotny 63-parametryczny oraz zredukowany 16-parametryczny

posłużyło do wyznaczenia wzorców wszystkich stu wyrazów słownika. Wzorce tworzono według zasady podanej w rozdziale 7.2. niniejszej pracy, poświęconym adaptacji. W części pamięci operacyjnej minikomputera MERA 303 zarezerwowanej dla inwentarza wzorców mieści się jedynie 26 wzorców. Z tego powodu cały inwentarz przechowywano w pamięci zewnętrznej na dysku elastycznym.

S Ł O W N I K 100-wyrazowy
(użyty w próbie testowej zmodyfikowanego modelu rozpoznającego)

amunicja	garnizon	obrońca	rozkaz
armia	generał	obręcz	schron
artyleria	głony	oddział	siła
atak	godziny	odwrót	system
batalion	groźba	obiekt	sytuacja
bateria	karabin	odcinek	suma
bitwa	kierunek	oficer	świadek
bohaterstwo	kilometr	okop	świt
brygada	konwencja	osłona	szpital
brzeg	krater	pirat	tabor
bunkier	lata	piechota	tyły
czołgi	linia	ptak	twierdza
cel	lot	południe	uderzenie
determinacja	lotnictwo	pozycja	uwaga
dni	ludność	północ	wojska
dostęp	ławka	próba	wsparcie
dowódca	łańcuch	przeszkoda	wschód
dowództwo	magazyn	przyczółek	wybuch
droga	marsz	przygotowanie	wulkan
drut	mech	punkt	zasiek
dywizja	nadwyżka	pułk	znaczenie
działo	natarcie	przykład	ziemia
działanie	noc	rejon	żołnierz
fort	obszar	rok	żelazo
fortyfikacja	obrona	rów	żar

S Ł O W N I K 26-wyrazowy
(użyty w próbie rozpoznawania wyrazów w oparciu o grupowe zbiory wzorców)

amunicja	groźba	odcinek	siła
batalion	karabin	pirat	świadek
brygada	lata	piechota	tyły
cel	ludność	próba	wojska
dowódca	marsz	punkt	wsparcie
dywizja	noc	rok	zasiek
garnizon	obręcz		

W trakcie rozpoznawania każdego wyrazu przepisywano wzorce z pamięci zewnętrznej do pamięci operacyjnej w trzech grupach po 26 wzorców i w jednej grupie liczącej 22 wzorce. Konieczność wielokrotnego wykonywania takich transmisji jest przykładem niedogodności wynikających z niskich parametrów technicznych minikomputera MERA 303.

Rozpoznawaniu poddano wpierw te wypowiedzi, na podstawie których wcześniej utworzono wzorce, czyli każdy wyraz słownika wypowiedziany 4-krotnie, co stanowiło łącznie 400 wypowiedzi. Wynik rozpoznawania wypadł bardzo dobrze. Wszystkie 400 wypowiedzi zostało rozpoznanych bezbłędnie. Ten pozytywny rezultat miał podwójną wartość. Po pierwsze stanowił dobre świadectwo o jakości wzorców, a po drugie złagodził wątpliwości w kwestii, czy wydatne uproszczenie binarnej reprezentacji wypowiedzi wyrazu nie popsuje jakości rozpoznawania.

Następny test rozpoznawania dotyczył 100 wypowiedzi, których nie wykorzystano przy tworzeniu wzorców. Były to wszystkie wyrazy słownika wypowiedziane w porządku losowym. W pierwszej turze rozpoznanych zostało poprawnie 86 wyrazów, co stanowiło jednocześnie 86-procentową poprawność rozpoznawania. Zbadano przyczyny niektórych błędnych odpowiedzi i stwierdzono, że powód niektórych pomyłek tkwi w niewłaściwie postawionym kryterium oceny różnic w podobieństwie obiektu do poszcze-

gólnych wzorców. Koniecznym wydaje się przypomnieć w tym miejscu, na czym owo kryterium polega i w jakiej fazie procesu rozpoznawania zostaje ono użyte. Wynikiem porównania każdego wzorca z rozpoznawanym obiektem, w tym przypadku uproszczonym spektrogramem binarnym rozpoznawanego wyrazu, jest ciąg Q_v wartości ilorazu podobieństwa poszczególnych fragmentów obiektu z odnośnymi fragmentami wzorca o numerze porządkowym v. Jak wiadomo z rozdziału 8.1, o większym podobieństwie jednego z dwóch wzorców (np. v-tego i (v-r)-tego) do niewiadomego obiektu decyduje to, który z tych wzorców w większej liczbie fragmentów wykazuje lepsze podobieństwo do obiektu. Rozstrzygnięcie tego wymaga uprzedniego poklasyfikowania elementów ciągu różnic ilorazów lokalnych podobieństw poszczególnych fragmentów obiektu do odpowiednich fragmentów wzorców v-tego i (v-r)-tego. Z przeprowadzonej przez autora analizy wartości podobieństw fragmentów spektrogramów binarnych tego samego wyrazu oraz wyrazów różnych wynika, iż zakres ten powinien być malejącą funkcją ilorazu lokalnego podobieństwa obiektu i (v-r)-tego wzorca aktualnie pretendującego do roli wyniku. Jednakże uwzględnienie tego wniosku uniemożliwił brak miejsca w wykorzystanej już do maximum niewielkiej pamięci operacyjnej minikomputera MERA 303. Pod wpływem ostatnio przytoczonych wyników rozpoznawania dokonano modyfikacji, która częściowo rozwiązała problem wyboru zakresu nieistotnych różnic w podobieństwach lokalnych obiektu do dwóch wzorców. Zakres ten nieznacznie poszerzono, utrzymując jednak jego szerokość stałą w przedziale wartości ilorazu podobieństwa fragmentów, rozciągającym się od pewnego progu q_n do jedności. Odnośne podobieństwa lokalne o wartości poniżej q_n przyjęto uznawać za nieróżniące się istotnie między sobą. Przy takim założeniu badanie różnicy ilorazów podobieństw q_{vi} i $q_{(v-r)i}$ i-tego fragmentu obiektu do odpowiednich fragmentów wzorców v-tego i (v-r)-tego następuje jedynie wtedy, gdy:

$$q_{vi} \vee q_{(v-r)i} = x, \quad \left\{ x \in X: q_n < x < 1 \right\} \qquad (8.12)$$

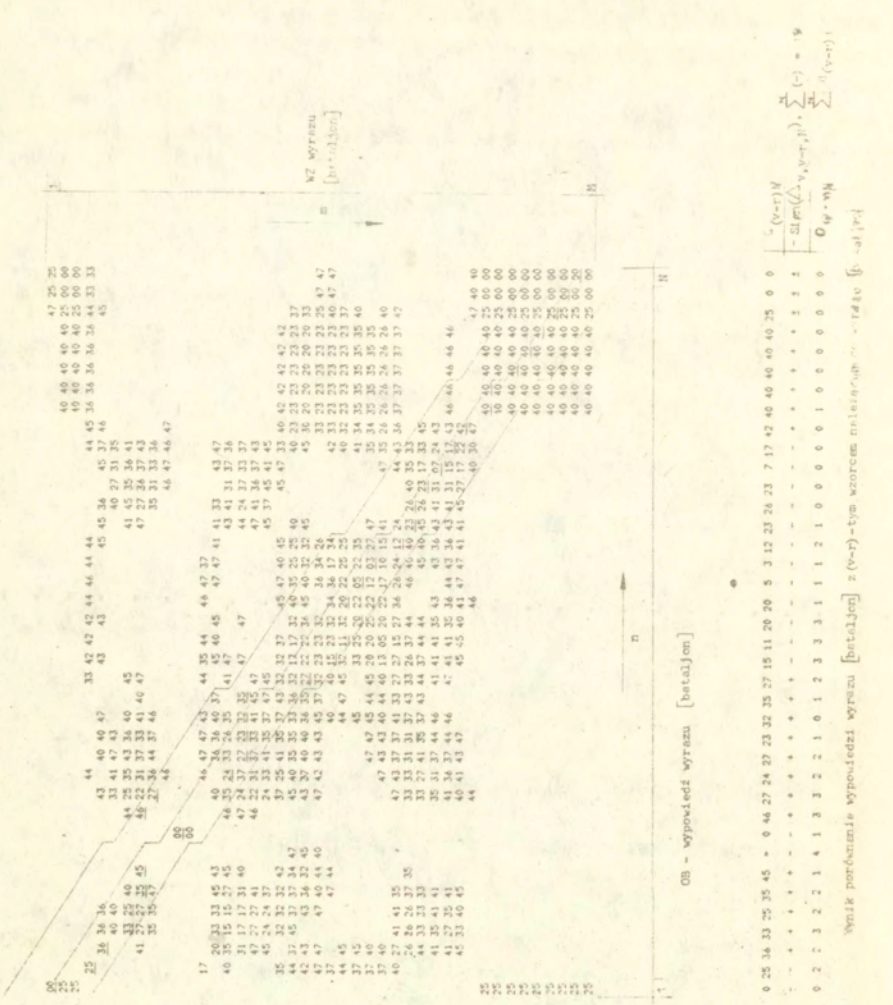
W przeciwnym razie i-ty fragment obiektu uważa się za równo podobny do odpowiednich fragmentów obu rozpatrywanych wzorców, a odnośne ilorazy podobieństw zalicza się zarówno do klasy KL_v

i $KL_{(v-r)}$, które jak wiadomo skupiają odpowiednio lepsze (klasa KL_v) i gorsze (klasa KL_{v-r}) podobieństwa poszczególnych fragmentów obiektu do odnośnych fragmentów kolejnego v-tego wzorca. Poziom progu q_n określono empirycznie. Po wprowadzeniu tej kolejnej modyfikacji z 14 wyrazów, które w poprzedniej turze nie zostały rozpoznane, 10 zostało obecnie rozpoznanych poprawnie, a 4 ponownie błędnie. Nierozpoznanymi wyrazami były: batalion, bateria, dowódca, punkt.

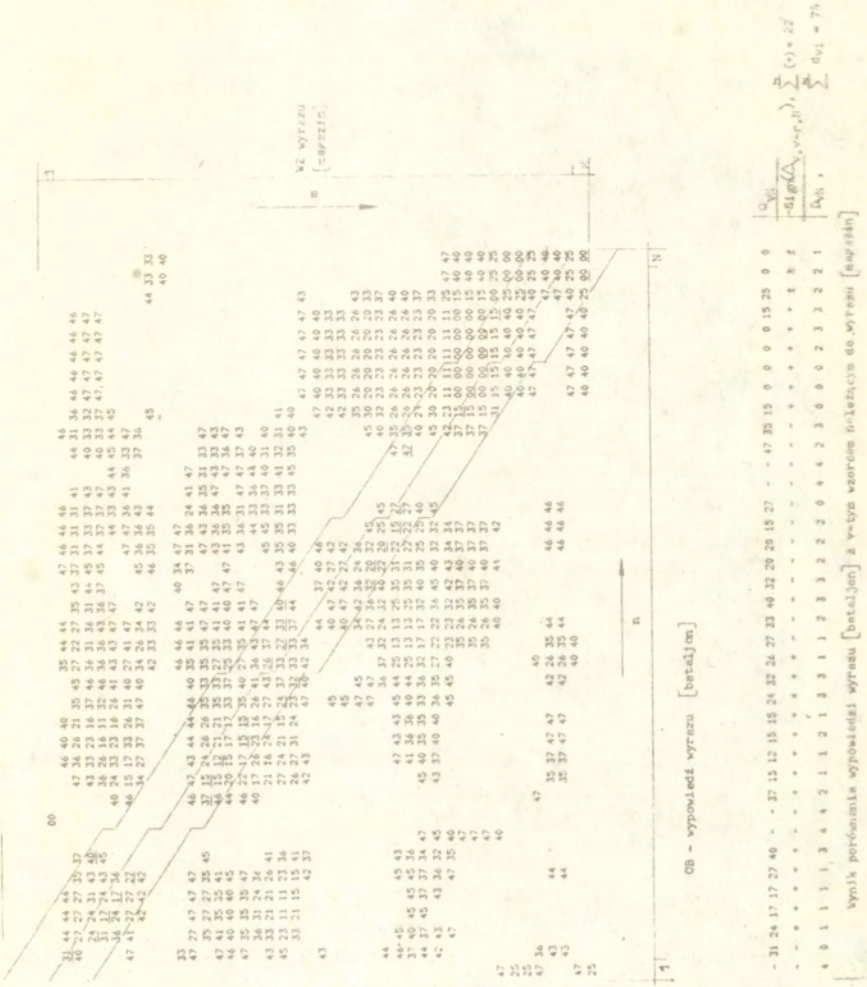
Niestety z braku miejsca w pamięci operacyjnej nie można było zastosować wspomnianego już wcześniej badania sum tzw. dewiacji kolejności czyli bezwzględnych różnic numerów porządkowych fragmentów wzorca optymalnie podobnych do odnośnych fragmentów obiektu i tych fragmentów wzorca, które powinny odpowiadać odnośnym fragmentom obiektu po quasiliniowym znormalizowaniu długości obiektu i wzorca. Dysponując procedurą badania tego rodzaju sum można by potwierdzić lub zakwestionować kandydaturę danego wzorca do wyniku rozpoznawania. Podstawę decyzyjnej roli takiej procedury stanowi zasada, że fragmenty wzorca optymalnie podobne do odpowiednich fragmentów obiektu nie powinny zbyt licznie pochodzić z obrzeży obszaru, z którego dokonuje się ich wyboru. Dla wyjaśnienia, w jakim stopniu wspomniana procedura badania sum różnic numerów porządkowych pomogłaby uniknąć czterech błędów powstałych w ostatnio omawianym teście rozpoznawania 100 wyrazów, wyznaczono macierze lokalnych podobieństw obrazu jednego z błędnie rozpoznanych wyrazów i dwóch wzorców, a mianowicie wzorca tego wyrazu i wzorca wyrazu będącego błędnym wynikiem rozpoznania. Obie macierze pokazano na rys. 34 zaznaczywszy w nich:

- obszary poszukiwania podobnych fragmentów - PPF,
- trajektorie quasiliniowej normalizacji długości obiektu i wzorca,
- elementy macierzy podkreślone, wyrażające najlepsze podobieństwa lokalne w obszarach PPF.

Ponadto w tabelce poniżej każdej macierzy, w rubryce oznaczonej przez $Q_{(\ )N}$, wypisano podkreślone elementy macierzy, a w rubryce $D_{(\ )N}$ odległości tych elementów od wspomnianej trajektorii normalizacji długości, mierzone liczbą odstępów między odpowiednimi wierszami macierzy. Suma tych odległości dla pier-



Rys. 34 a. Macierz wartości lokalnych podobieństw obiektu i dotyczącego go wzorca oraz tabela elementów ciągów: $Q_{(v-r)N}$, -Sign $(\Delta_{v(v-r)N})$, $D_{(v-r)N}$



Rys. 34 b. Macierz wartości lokalnych podobieństw obiektu i wzorca błędnej odpowiedzi oraz tabela elementów ciągów: Q_{vN}, $-\mathrm{Sign}\,(\Delta_{v(v-r)N})$, D_{vN}

wszej macierzy jest zdecydowanie mniejsza od analogicznej sumy dla drugiej macierzy. Wskazuje to, że wzorzec, którego lokalne podobieństwa do obiektu reprezentuje macierz pierwsza, lepiej przystaje do obiektu. Fakt ten nie stanowi potwierdzenia uzyskanego wyniku rozpoznawania, który był błędny, podpowiada natomiast wynik poprawny. Z analizy tego przypadku wynika więc potwierdzenie słuszności wcześniejszej sugestii o opłacalności użycia kryterium selekcjonującego kandydatów do wyniku rozpoznania wyrazu na podstawie wartości względnej sumy tzw. dewiacji kolejności podobnych fragmentów. Względnej w tym sensie, że odniesionej do długości obiektu.

W ramach trzeciej serii prób, przeprowadzonych po dokonaniu przedstawionych wyżej modyfikacji modelu, zbadano możliwość rozpoznawania wypowiedzi dwóch głosów w oparciu o ich wspólne wzorce. Ze względu na takie ukierunkowanie testów rozpoznawania użyto słownika zredukowanego do takiej ilości słów, aby wszystkie wzorce mogły zmieścić się jednocześnie w pamięci operacyjnej minikomputera. Przyjęto zatem podany wyżej słownik złożony z 26 wyrazów zaczerpniętych z poprzednio ułożonego słownika 100-wyrazowego. W testach wzięły udział 4 głosy męskie, oznaczone przez G1, G2, G3, G4. Głosy te połączono w 6 następujących par: G1/G2, G1/G3, G1/G4, G2/G3, G2/G4, G3/G4. Wzorzec danego wyrazu słownika dla danej pary głosów tworzono na podstawie dwóch wypowiedzi tego wyrazu przez każdy z dwóch głosów według następującej procedury: Z dwóch wypowiedzi tego samego v-tego wyrazu przez ten sam głos G_i utworzono wpierw wzorzec pośredni zgodnie z regułą:

$$WZP_{vGi} = SB_{vGi1} \cup SB_{vGi2} \qquad (8.13)$$

Analogicznie na podstawie wypowiedzi drugiego głosu G_j utworzono drugi wzorzec pośredni:

$$WZP_{vGj} = SB_{vGj1} \cup SB_{vGj2} \qquad (8.14)$$

W wyniku kompilacji logicznej obu wzorców pośrednich powstał wzorzec v-tego wyrazu dla pary głosów Gi/Gj:

$$WZ_v = WZP_{vGi} \cap WZP_{vGj} \qquad (8.15)$$

W teście rozpoznawania wyrazy wypowiedziane przez głosy należące do jednej pary były rozpoznawane w oparciu o wzorce utworzone dla tej pary. Rozpoznawaniu poddano zarówno te wypowiedzi, które wykorzystano do utworzenia wzorców, jak i te, które do tego nie zostały użyte. Rozpoznawanych było 104 wypowiedzi każdego głosu, w tym 52 wypowiedzi użyte do utworzenia wzorców. Ponieważ w testach brały udział 4 głosy, wszystkich rozpoznawanych wypowiedzi było razem 416.

Poniżej zestawiono otrzymane wyniki. Poszczególne głosy oznaczono przez: G1, G2, G3, G4, a pary w jakie głosy te połączono - przez Gi/Gj. Indeksy i oraz j symbolizują numer głosu wchodzącego w skład pary. Litera A w nawiasie obok notacji głosu oznacza, że wyniki rozpoznawania dotyczą wypowiedzi, które posłużyły do zbudowania wzorców. W tablicy 5 zamieszczono uzy-

Tablica 5. Procentowa poprawność rozpoznawania wyrazów w oparciu o wzorce wspólne dla pary głosów

Gi / Gi/Gj	G_1/A/	G_1/R/	G_2/A/	G_2/R/	G_3/A/	G_3/R/	G_4/A/	G_4/R/
G_1/G_2	98.80	82.70	100	98.80				
G_1/G_3	96.16	96.16			100	96.16		
G_1/G_4	94.23	88.47					100	96.16
G_2/G_3			96.16	98.80	100	96.16		
G_2/G_4			90.39	90.39			100	90.39
G_3/G_4					100	98.80	98.80	96.16

skane wyniki rozpoznawania wyrażone w procentach. W zestawieniu nietabelarycznym natomiast w pierwszej kolumnie cyfr podane są liczebności błędnych rozpoznań (liczby bez nawiasów), a w drugiej ilości nierozpoznanych słów (liczby w nawiasach), jeśli błąd dotyczył dwukrotnie tego samego słowa. W zestawieniu tym zamieszczono też wykaz wypowiedzi (zapisy w nawiasach), które zostały błędnie rozpoznane. Błędną odpowiedź podano obok zapisu w nawiasie nadanego wyrazu. I tak np. zapis w rodzaju: /marsz/-ludność oznacza, że wyraz /marsz/ nie został rozpoznany, a błędną odpowiedzią był wyraz /ludność/. W formie bardziej syntetycznej przedstawione zostały wyniki rozpoznawania w tablicy 6. Podaje ona mianowicie, które słowa wypowiedziane przez które głosy ilokrotnie nie zostały prawidłowo rozpoznane. Na boku tej tablicy podano procentową ilość przypadków nierozpoznania poszczególnych wyrazów słownika na 48 testów rozpoznawania każdego wyrazu łącznie dla wszystkich głosów. Wyrazy, których w tablicy 6 nie wykazano, były zawsze rozpoznawane poprawnie. W ostatnim wierszu tablicy 6 figurują procentowe ilości przypadków nierozpoznania wypowiedzi należących do poszczególnych głosów, na 312 testów rozpoznawania wyrazów przypadających na każdy głos.

ZESTAWIENIE BŁĘDÓW W ROZPOZNAWANIU WYRAZÓW
w oparciu o grupowe zbiory wzorców

Grupa G1/G2

G1/A/ 1 /karabin/-wsparcie
G2/A/ 0
G1/R/ 9 /8/ /noc/-amunicja, /noc/-amunicja, /świadek/-pirat, /marsz/-tyły, /garnizon/-groźba, /dowódca/-wojska, /wojska/-groźba, /rok/-piechota, /karabin/--batalion
G2/R/ 1 /noc/-amunicja

Grupa G1/G3

G1/A/ 2 /marsz/-ludność, /punkt/-pirat

G3/A/ 0
G1/R/ 2 /punkt/-obręcz, /noc/-amunicja
G3/R/ 2 /1/ /punkt/-amunicja, /punkt/-ludność

Grupa G1/G4

G1/A/ 3 /2/ /punkt/-piechota, /punkt/-ludność, /rok/-próba
G4/A/ 0
G1/R/ 6 /5/ /dowódca/-wojska, /marsz/-ludność, /pirat/-świadek, /punkt/-amunicja, /rok/-tyły, /rok/-lata
G4/R/ 2 /pirat/-świadek, /punkt/-ludność

Grupa G2/G3

G2/A/ 2 /pirat/-świadek, /batalion/-garnizon
G3/A/ 0
G2/R/ 1 /pirat/-świadek
G3/R/ 2 /piechota/-lata, /punkt/-ludność

Grupa G2/G4

G2/A/ 5 /4/ /groźba/-brygada, /groźba/-brygada, /lata/-piechota, /pirat/-świadek, /świadek/-odcinek
G4/A/ 0
G2/R/ 5 /3/ /groźba/-brygada, /pirat/-świadek, /pirat/-świadek, /próba/-brygada, /próba/-brygada
G4/R/ 5 /4/ /piechota/-brygada, /pirat/-brygada, /punkt/-ludność, /lata/-batalion, /lata/-batalion

Grupa G3/G4

G3/A/ 0
G4/A/ 1 /lata/-groźba
G3/R/ 1 /punkt/-pirat
G4/R/ 2 /1/ /noc/-marsz, /noc/-dowódca

Tablica 6. Sumaryczne zestawienie błędów w rozpoznawaniu wyrazów w oparciu o grupowe zbiory wzorców

Lp.	Wyraz nierozpoznany	G_1	G_2	G_3	G_4	% błędów
1	batalion		1			2.07
2	dowódca	2				4.15
3	garnizon	1				2.07
4	groźba		3			6.25
5	karabin	2				4.15
6	lata		1		3	8.30
7	marsz	3				6.25
8	noc	3	1		2	12.50
9	piechota			1	1	4.15
1o	pirat	1	5		2	16.70
11	próba		2			4.15
12	punkt	5		4	2	22.90
13	rok	4				8.30
14	świadek	1	1			4.15
15	wojska	1				2.07
% błędów		7.4	4.5	1.6	3.2	

Przedstawione wyniki dają pierwszy pogląd o możliwości rozpoznawania wyrazów na podstawie wspólnych zbiorów wzorcowych spektrogramów binarnych.

Dla trzech spośród sześciu możliwych skojarzeń czterech głosów w pary uzyskano wyniki względnie korzystne. Tymi parami są: G3/G4, G2/G3, G1/G3. Charakterystyczne jest, że we wszystkich tych parach występuje głos G3. Wyniki uzyskane dla par głosów G1/G2, G2/G4, G1/G4 są zdecydowanie gorsze. W tych pa-

rach nie ma głosu G3, z czego wnioskować można, że podobieństwa głosów G1, G2 i G4 są stosunkowo odległe, i że głos G3 zajmuje centralną pozycję wśród pozostałych głosów.

Z wyżej przedstawionych wyników wnioskować też można, że uzyskanie dla niektórych wyrazów i głosów reprezentatywnego wzorca wielogłosowego może okazać się trudne lub wręcz niemożliwe. Przykładem dowodzącym tego są wyrazy: /punkt/, /pirat/, /noc/, rozpoznawane najgorzej we wszystkich parach głosów. I tak wyraz /punkt/ na 48 wypowiedzi aż w 11 przypadkach, czyli w prawie 23%, nie został rozpoznany, wyraz /pirat/ w 17%, a wyraz /noc/ w 12,5%.

Na 26 wyrazów 11 rozpoznawanych było ze 100% poprawnością, a 3 z poprawnością bliską 98%, 4 z poprawnością około 95,8%, 3 z poprawnością 93,7%, a pozostałe 4 (wśród nich wspomniane już wyrazy /punkt/, /pirat/ i /noc/) z poprawnością w granicach od 77 do 91,7%. Przyczyny trudności w uzyskaniu reprezentatywnych wzorców grupowych tkwią w wydatnym zindywidualizowaniu niektórych głosów, przejawiającym się szczególnie w niektórych wypowiedziach. Duże zróżnicowanie międzyosobnicze występuje na przykład ze względu na rozległość czasową i częstotliwościową aspiracji w spółgłoskach zwartych bezdźwięcznych oraz położenie i przebieg formantów w samogłoskach. Niektóre głosy w trakcie wymawiania spółgłosek zwartych dźwięcznych cechuje bardzo słaba i przez to nieekstrahowalna w układach technicznych fonacja. Problem cech indywidualnych głosów i sposobów uniezależnienia się od nich w automatycznym rozpoznawaniu mowy jest bardzo złożony i pozostaje ciągle przedmiotem badań.

W odniesieniu do przedstawionej w niniejszej pracy metody globalnego rozpoznawania wypowiedzi izolowanych wyrazów kwestię tę można scharakteryzować następująco:

1. Rozpoznawanie wyrazów w oparciu o grupowy zbiór wzorcowych spektrogramów binarnych jest względnie poprawne, gdy grupa jest nieliczna i skupia właściwie dobrane głosy.
2. Jakość rozpoznawania wyrazów w oparciu o wspólny dla przypadkowo dobranych głosów zbiór wzorców może być bardzo zróżnicowana zarówno międzygłosowo, jak i międzywyrazowo.

9. Charakterystyka i ocena modelu ROWBIR 1

Używane w niniejszej pracy określenie: „model rozpoznawania wyrazów" dotyczy zarówno samej koncepcji rozpoznawania, jak i technicznego sposobu jej realizacji. Po odpowiednio szczegółowym przedstawieniu modelu ROWBIR 1 w poprzednich rozdziałach, w rozdziale obecnym przeprowadzona zostanie jego charakterystyka i ocena.

Model ROWBIR 1 zalicza się do kategorii modeli globalnego rozpoznawania krótkich wypowiedzi o rozciągłości wyrazu, ewentualnie ciągu kilku wyrazów. Rozpoznaje on więc tylko wypowiedzi oddzielone pauzami od ewentualnej wypowiedzi wcześniejszej lub późniejszej. Wymaga się też, aby zarówno podczas obu tych pauz, jak i podczas wypowiedzi nie było żadnych sygnałów zakłócających. Konieczne jest bowiem stworzenie akustycznych warunków, które umożliwią później dokładne określenie początku i końca wypowiedzi.

Cechą szczególną modelu ROWBIR 1 jest to, że posługuje się on binarną reprezentacją sygnału mowy. Wypowiedź wyrazu wyrażona zostaje przez tzw. spektrogram binarny, będący ciągiem widm binarnych, reprezentujących sygnał mowy w kolejnych momentach, następujących po sobie w stałych odstępach czasu. W trzech kolejnych wersjach modelu ROWBIR 1 posługiwano się odpowiednio reprezentacją 63-, 32- i 16-parametryczną. Dotychczas przeprowadzone testy rozpoznawania wykazały, że dla automatycznego rozpoznawania wyrazów w sposób globalny wystarcza już reprezentacja 16-parametryczna.

Binarny charakter widma polega na tym, że poszczególne jego parametry mogą być jedynie dwuwartościowe, tzn. przyjmować wartości Ø lub 1. Reprezentacja wypowiedzi wyrazu w takiej formie charakteryzuje się bardzo małą objętością informacyjną, w której jednakże przy właściwie dobranym kryterium binaryzacji zawierają się najważniejsze cechy widmowe sygnału mowy. Dzięki tej wyjątkowo korzystnej zalecie reprezentacja binarna sygnału mowy okazała się bardzo atrakcyjna w globalnym rozpoznawaniu wyrazów.

Następną cechą szczególną modelu ROWBIR 1, wynikającą już z faktu, że posługuje się on binarną reprezentacją sygnału mowy, jest to, że szereg podstawowych i bardzo licznych działań, zarówno w procesie adaptacji, jak i rozpoznawania, stanowią operacje logiczne na ciągach liczb binarnych.

Inną cechą szczególną modelu ROWBIR 1 jest dynamiczne ograniczanie ilości informacji zawartej w widmie binarnym stanowiącym reprezentację sygnału mowy. Inspiracją do zastosowania tego rodzaju redukcji widma był pogląd, że dalsze jedynki w widmie binarnym nie charakteryzują jednoznacznie segmentów mowy i przez to utrudniają ich poprawną identyfikację w procesie automatycznego rozpoznawania wyrazów. Dynamiczne ograniczanie zakresu widma binarnego polega więc na eliminowaniu z widma jedynek dalszych w kolejności i pozostawianiu jedynie określonej liczby jedynek pierwszych.

Szereg cech charakterystycznych modelu ROWBIR 1 dotyczy poszczególnych jego elementów, takich jak: wyznaczanie widma binarnego, określanie podobieństw lokalnych i globalnych porównywanych spektrogramów binarnych, wyznaczanie wzorców wyrazowych. W obecnym rozdziale cechy poszczególnych elementów modelu zostaną wymienione i po krótce skomentowane.

Tworzenie widma binarnego, będące podstawowym elementem modelu ROWBIR 1, charakteryzuje kilka cech. Pierwsza dotyczy wygładzenia dyskretnego widma amplitudowego wyznaczonego przez wielokanałowy analizator widma i umieszczonego w pamięci minikomputera. W modelu ROWBIR 1 wygładzenie to następuje przy użyciu specjalnie wyznaczonej funkcji, dzięki czemu z pierwotnego widma wyeliminowane zostają szczegóły odzwierciedlające harmoniczny charakter sygnału. Użycie wspomnianej funkcji uważać można za jedną z osobliwości modelu ROWBIR 1. Staranne wygładzenie widm amplitudowych jest konieczne dla zapewnienia reprezentatywności spektrogramów binarnych, bez której nie można by liczyć na powodzenie całego przedsięwzięcia rozpoznawania wyrazów na podstawie ich spektrogramów binarnych. Stąd na sprawę właściwego przygotowania widma amplitudowego do późniejszego przekształcenia go w widmo binarne zwrócona została szczególna uwaga.

O reprezentatywności widma binarnego decyduje także roz-

dzielczosc częstotliwościowa widma amplitudowego, z którego widmo binarne zostało wyznaczone. Model ROWBIR 1 pozwala uzyskać widma amplitudowe sygnału mowy o małej rezolucji częstotliwościowej, wynoszącej w najważniejszym zakresie jedynie 80 Hz. Przystępując do badań nad rozpoznawaniem wyrazów reprezentowanych przez spektrogramy binarne, nie było wiadomo, jaką minimalną rozdzielczość częstotliwościową spektrogramy takie w tym zastosowaniu powinny posiadać. Uznano wobec tego za konieczne przyjęcie na początek rozdzielczości lepszej od tej, jaka mogłaby okazać się później wystarczająca.

Przekształcenie widma amplitudowego w widmo binarne może opierać się o różnorakie zasady. Rzeczą wspólną widm binarnych powstałych drogą różnych przekształceń jest tylko ich forma. Cechą szczególną przekształcenia widma amplitudowego w widmo binarne, dokonującego się w modelu ROWBIR 1 jest to, że uzależnia ono wartość poszczególnych parametrów widma binarnego od istnienia wyraźnej wypukłości w kolejnych miejscach obwiedni widma amplitudowego. Tak uzyskane widmo binarne jest więc odzwierciedleniem położenia formantów w widmie amplitudowym i wyraża tym samym ważne cechy dystynktywne dźwięków mowy. Warto w tym miejscu wspomnieć, że wiele istniejących obecnie metod rozpoznawania mowy spotyka się z krytyką z powodu braku w nich motywacji fonetyczno-akustycznych.

W późniejszej wersji modelu ROWBIR 1 zastosowano dodatkowo przekształcenie redukujące liczbę parametrów widma binarnego z 63 do 16 poprzez zastępowanie ciągów parametrów widma binarnego pierwotnego pojedynczymi parametrami nowego widma zredukowanego. Ta innowacja będąca dodatkowym elementem wyznaczania widma binarnego zalicza się do cech charakterystycznych modelu ROWBIR 1.

Zachodzi pytanie, czy można będzie zrezygnować z dotychczas stosowanej rozdzielczości częstotliwościowej przy wyznaczeniu widm amplitudowych sygnału mowy, skoro przejście na widma binarne 16-parametryczne oznacza praktycznie rezygnację z tak dokładnej rozdzielczości. Na to pytanie trudno jeszcze obecnie udzielić pewnej odpowiedzi, chociaż spodziewać należałoby się odpowiedzi pozytywnej. Jak wiadomo, widmo binarne 16-parametryczne powstaje z dostarczonego przez analizator WAAW

widma amplitudowego drogą szeregu przekształceń, które są zarówno wzajemnie zależne, jak i zależne od formy, jaką ma reprezentacja sygnału mowy przed przekształceniem, czyli na wyjściu analizatora. Z tych powodów nie wchodzi w grę jedynie automatyczne zredukowanie ilości kanałów analizatora do 16 i nadanie ich pasmom szerokości zgodnych z zakresami, jakich dotyczą poszczególne parametry 16-parametrycznego widma binarnego. Dotychczasowe sposoby wygładzania widma amplitudowego i wyznaczania widma binarnego nie mogłyby wówczas zostać użyte wprost, lecz należałoby je dostosować do nowej formy widmowej reprezentacji sygnału mowy wynikającej z nowego systemu analizy widmowej, a ściślej mówiąc - z nowego podziału zakresu częstotliwości na pasma analizy.

Chociaż zasadniczym celem wprowadzenia przekształcenia redukującego liczbę binarnych parametrów widmowych do 16 było uzyskanie możliwie najmniejszej objętości informacyjnej widma binarnego, spodziewano się także, że przekształcenie to odegra rolę czynnika minimalizującego wpływ niewielkich lub nieistotnych różnic zachodzących pomiędzy sygnałami mowy na ich reprezentacje binarne. Upatrywano, że uproszczone spektrogramy binarne okażą się właściwszą reprezentacją w rozpoznawaniu wyrazów na podstawie wzorców grupowych, czyli obowiązujących dla pewnej liczby głosów w miarę podobnych.

Kilkoma charakterystycznymi cechami odznacza się model ROWBIR 1 w zakresie metody porównywania spektrogramów binarnych, które jest podstawowym działaniem zarówno w adaptacji, jak i rozpoznawaniu wyrazów. Cechy te dotyczą przede wszystkim sposobu poszukiwania odpowiadających sobie fragmentów w dwóch porównywanych spektrogramach różniących się wzajemnie zarówno pod względem długości czasowej, jak i rozkładu czasowego poszczególnych segmentów. Istotną cechę w tym zakresie stanowi zasada, że odpowiadającymi sobie są dwa fragmenty najbardziej wzajemnie podobne w dwóch quasiliniowo znormalizowanych spektrogramach, oddalone od siebie w skali ich ujednoliconej długości nie więcej niż założona dopuszczalna odległość. Sposób poszukiwania podobnych fragmentów charakteryzuje też użyta miara podobieństwa. Jest nią w modelu ROWBIR 1 iloraz liczby parametrów widmowych, mających w dwóch porównywanych fragmentach

zgodnie wartość 1, i liczby wszystkich parametrów o tej wartości w obu fragmentach. Wyznaczenie ilorazu podobieństwa fragmentów oprócz kilku prostych działań logicznych i zliczenia jedynek wymaga niestety dzielenia liczb, wykonywanego w minikomputerze MERA 303 stosunkowo wolno. Mimo, iż wyniki uzyskane w dotychczas przeprowadzonych testach rozpoznawania nie wskazują wyraźnie na zawodność tej miary, to jednak istnieją przesłanki skłaniające do zmodyfikowania jej w przyszłości. Byłoby pożądane, aby zmodyfikowana miara dokładniej wyrażała różnice w położeniu ciągów jedynek w porównywanych fragmentach w celu uniknięcia wskazania jednakowego stopnia podobieństwa jakiegoś fragmentu do dwóch innych zasadniczo różniących się nawzajem fragmentów. Przy zastosowaniu obecnej miary mogą mieć miejsce przypadki uchybień w ocenie podobieństwa fragmentów z powodu zbyt ogólnego uwzględniania tych różnic. Dowodzi tego przykład zilustrowany na rys. 35. Podobieństwa fragmentów A i B oraz A i C wyrażone obecnie stosowaną miarą są identyczne, chociaż fragmenty B i C różnią się wzajemnie w sposób istotny a większe podobieństwo fragmentów A i B wydaje się oczywiste.

Inną szczególną cechą modelu ROWBIR 1 w zakresie porównywania spektrogramów binarnych jest sposób określania ich globalnego podobieństwa. W przeciwieństwie do powszechnie stosowanej w rozpoznawaniu wyrazów zasady, że podobieństwo globalne obrazów wzorca i obiektu jest średnią z podobieństw lokalnych zachodzących między tymi obrazami, w modelu ROWBIR 1 globalne podobieństwo obiektu i wzorca pozostaje wyrażone przez ciąg wartości podobieństw lokalnych. Na podstawie wartości elementów tego ciągu nie orzeka się jednak wprost, czy rozpatrywany wzorzec dotyczy obiektu. Decyzja o wyniku rozpoznania podjęta zostaje w oparciu o zasadę, że obiekt i dotyczący go wzorzec wykazują najlepsze podobieństwo w największej liczbie fragmentów. Ten sam obiekt może wykazać lepsze podobieństwo z innymi wzorcami, lecz w mniejszej liczbie fragmentów. Wskazanie najlepszego podobieństwa lokalnego danego fragmentu obiektu z odnośnymi fragmentami poszczególnych wzorców odbywa się przy uwzględnieniu pewnej tolerancji, w zakresie której lokalne podobieństwa uważa się za identyczne. Ta metoda wyłaniania wzor-

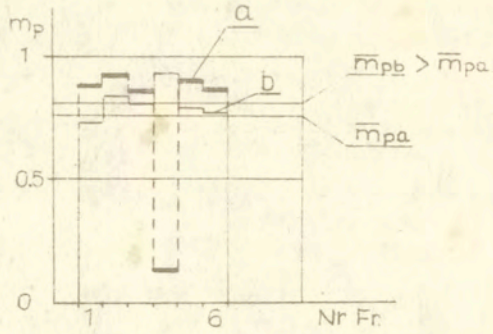
0 0	0 0	0 0
0 0	0 0	0 0
0 0	0 0	0 1
0 0	0 0	1 0
0 0	0 0	1 0
0 0	0 0	0 1
0 0	0 0	0 0
0 0	0 0	0 0
0 0	0 1	0 0
0 1	1 1	0 1
1 1	1 1	1 1
1 0	1 1	1 0
0 0	0 0	0 0
0 0	0 1	0 0
1 1	1 1	1 1
1 1	1 1	1 1
Fr A	Fr B	Fr C

$$m_p(A/C) = m_p(A/C)$$

Fr B ≠ Fr C

Rys. 35. Przykład zawodności miary podobieństwa lokalnego spektrogramów binarnych stosowanej w modelu ROWBIR 1

ca najbardziej podobnego do obiektu jest właściwsza od metod operujących średnim lokalnym podobieństwem i stosowanych najczęściej. O słuszności takiej opinii zaświadczyć może przykład zilustrowany na rys. 36. Przedstawiono na nim przykładowe wykresy wartości lokalnych podobieństw pewnego krótkiego obiektu z dwoma wzorcami X i Y. Wykres a/ wskazuje, że wszystkie z jednym wyjątkiem fragmenty obiektu wykazują dobre podobieństwo do odnośnych fragmentów wzorca X. Wspomniany jeden fragment jest zdecydowanie niepodobny do swojego odpowiednika we wzorcu X, przez co zaniżona jest wartość średniego podobieństwa lokalnego obiektu i wzorca X. Wykres b/ wyraża podobieństwa lo-

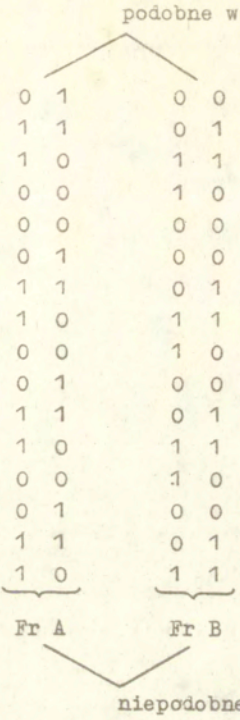


Rys. 36. Przykład zawodności najczęściej zalecanej miary podobieństwa globalnego wzorca i obiektu. Wykresy a i b przedstawiają odpowiednio lokalne podobieństwa obiektu do wzorca X i wzorca Y. Przez $\bar{m}_{pa}$ i $\bar{m}_{pb}$ oznaczono odpowiednio średnie z tych podobieństw lokalnych

kalne tego samego obiektu do wzorca Y. Są one z jednym wyjątkiem gorsze niż w przypadku wzorca X, którego dotyczy wykres a/. Średnie podobieństwo lokalne wynikające z porównania obiektu ze wzorcem Y jest jednak lepsze niż analogiczne podobieństwo uzyskane z porównania obiektu ze wzorcem X.

Zastosowaniu w modelu ROWBIR 1 odmiennej metody znajdowania wzorca optymalnie przystającego do rozpoznawanego obiektu przyświecała intencja zapobieżenia powstawaniu błędów w rozpoznawaniu wyrazów w takich przypadkach, jak wyżej zilustrowany. Autor reprezentuje pogląd, że w przedstawionym przykładzie wynikiem rozpoznawania powinien być wyraz mający wzorzec X, gdyż brak podobieństwa w obrębie tylko jednego fragmentu uważać można za przypadkowy. Fragment obejmuje krótki segment sygnału mowy.

Odniesienie porównań lokalnych do fragmentów, a nie pojedynczych widm stanowi korzystną cechę zaprezentowanego w niniejszej pracy modelu rozpoznającego wyrazy. Fragment wyraża bowiem kierunek zmian zachodzących w widmie i tym samym pewien krótkoterminowy przebieg sygnału mowy, podczas gdy pojedyncze



Rys. 37. Podobne widma binarne w dwóch wzajemnie różniących się fragmentach spektrogramów binarnych

widmo informuje jedynie o stanie chwilowym sygnału. Dzięki temu, że sąsiednie fragmenty mogą zachodzić na siebie, operowanie fragmentami zamiast pojedynczymi widmami w porównaniach lokalnych nie wymaga zmniejszenia liczby punktów, w których przyrównywane są spektrogramy binarne. Podobieństwo dwóch fragmentów jest bardziej wymowne niż podobieństwo dwóch widm, które może być przypadkowo dobre, mimo iż porównane widma pochodzą z fonetycznie różnych segmentów mowy. Dwa podobne widma binarne mogą np. należeć do dwóch porównanych fragmentów nie będących wcale podobnymi. Przykład tego pokazano na rys. 37.

Do charakterystycznych osobliwości modelu ROWBIR 1 zalicza się także zastosowana w nim metoda wyznaczania wzorcowych spektrogramów binarnych wyrazów. Dzięki temu, że reprezentacją wypowiedzi jest ciąg widm binarnych, operacje tworzenia wzorca są w swojej istotnej części działaniami logicznymi alternatywno-koniunkcyjnymi. Procedura wyznaczania wzorca na podstawie szeregu wypowiedzi podporządkowana została wymaganiu, aby wzorzec przyjął te cechy widmowe, które przynajmniej raz wystąpiły w spektrogramach binarnych pary kolejnych wypowiedzi. Każdy spektrogram wchodzi w skład tylko jednej pary, co oznacza, że pary nie zachodzą na siebie. Spełnienie wyżej wymienionego wymagania oznacza równocześnie, że do wzorca przechodzą wszystkie te cechy widmowe, które występują w spektrogramach binarnych co najmniej połowy wypowiedzi służących do jego utworzenia. Adaptacja jest w modelu ROWBIR 1 w pełni automatyczna i niezależna zarówno od rodzaju wyrazu, jak i języka, do którego wyraz ten należy. Ta cecha modelu ROWBIR 1 nie jest bynajmniej tylko jego cechą, gdyż taką samą właściwość posiadają niemal wszystkie modele rozpoznające wyrazy w sposób globalny, niezależnie od formy, w jakiej wyrażona zostaje wypowiedź i wzorce wyrazów. Jest ona niewątpliwie zaletą korzystnie wyróżniającą model globalnego rozpoznawania wyrazów i dlatego wspomniano o niej w tym miejscu.

Model rozpoznawania wyrazów charakteryzują także formy jego technicznej realizacji. Model ROWBIR 1 nie wyróżnia się pod tym względem nowoczesnością. Mimo to na podkreślenie zasługują co najmniej dwie rzeczy. Po pierwsze, zrealizowanie tego modelu wymagało kilku nietypowych rozwiązań technicznych, które w warunkach krajowych i w czasie, gdy je opracowywano, były zadaniami ambitnymi. Dotyczyły one analizy widmowej sygnału mowy, przesyłania danych widmowych w systemie "on line" do minikomputera oraz wizualizacji obrazów widmowych. Powstały w rezultacie 3 oryginalne urządzenia: 63-kanałowy analogowy analizator widma wraz z komutatorem wyjść, kanał funkcji analogowych z logarytmującym konwerterem analogowo-cyfrowym oraz memoskop. Analizator widma wraz z komutatorem wyjść zaprojektował w całości autor. Pozostałe urządzenia opracowane zostały zespołowo z inicjatywy i przy współudziale autora. Chociaż urządzenia te pow-

stały głównie dla automatycznego rozpoznawania wyrazów, wkrótce znalazły one zastosowanie w wielu innych badaniach z dziedziny fonetyki akustycznej, takich jak np. synteza mowy czy analiza parametru Fo.

Drugą sprawą wymagającą podkreślenia jest to, że model ROWBIR 1 zrealizowany został przy pomocy bardzo prymitywnego minikomputera, charakteryzującego się niskimi parametrami oraz małą zdolnością obliczeniową. Programy układane były wyłącznie w kodzie maszynowym, co było zabiegiem bardzo skomplikowanym i pracochłonnym. Kształt modelu ROWBIR 1 został w dużym stopniu zdeterminowany przez dostępne autorowi środki techniczne, które były skromne i ograniczone. Przygotowanie warunków technicznych niezbędnych do podjęcia prac nad automatycznym rozpoznawaniem wyrazów stanowiło ważny i niestety długi etap przedsięwzięcia, które doprowadziło do powstania modelu ROWBIR 1. Przy opracowywaniu modelu ROWBIR 1 brano też pod uwagę aspekty wdrożeniowe. Polegało to na akceptowaniu jedynie takich rozwiązań, które dałyby się przenieść w przyszłość do modelu użytkowego, a ten powinien być zarówno prosty i tani, jak i skuteczny w działaniu. Działania, które w modelu ROWBIR 1 wykonuje minikomputer MERA 303, są na tyle proste, że w przyszłym użytkowym modelu rozpoznawania wyrazów będzie mógł je wykonywać układ mikroprocesorowy dysponujący odpowiednio dużą pamięcią typu ROM na przechowywanie stałych programów i wzorców wyrazowych oraz pamięcią typu RAM służącą jako bufory różnych danych. Pamięci ROM przechowujące wzorce powinny być wymienne, gdyż dla każdego słownika i ewentualnie każdego głosu obowiązywać będzie oddzielny zbiór wzorców. Dla niektórych głosów będzie zapewne możliwy wspólny zbiór wzorców. Wzorce tworzone podczas adaptacji umieszczane będą wpierw w pamięci typu RAM, po czym nastąpi skopiowanie ich w pamięci typu ROM.

Model użytkowy powinien rozpoznać wypowiedź wyrazu w czasie możliwie jak najkrótszym. Ponieważ rozpoznawanie przebiegać będzie w sposób globalny, nie będzie ono możliwe w trakcie trwania wypowiedzi, lecz dopiero po jej zakończeniu. Uzyskanie krótkich czasów rozpoznawania będzie możliwe poprzez stworzenie warunków do porównywania obiektu z kilkoma wzorcami naraz, a także poprzez zastosowanie ukierunkowanego poszukiwania wzor-

ca najbardziej podobnego do obiektu. Równoczesne porównywanie kilku wzorców z obiektem będzie możliwe w przyszłym modelu użytkowym tylko wówczas, gdy będzie on wieloprocesorowy, zaś warunki do ukierunkowanego poszukiwania wzorca będzie można stworzyć np. poprzez odpowiednie pogrupowanie wzorców. Zasadę tego rodzaju adaptacji omówiono w rozdziale 7.2.

Mniej optymistycznie rysuje się perspektywa zrealizowania w modelu użytkowym tej części, którą w modelu ROWBIR 1 stanowi 63-kanałowy analogowy analizator widma. Liczyć można wprawdzie na to, że liczba pasm analizatora widma ulegnie w modelu użytkowym zredukowaniu. Nie umniejszy to jednak trudności w skonstruowaniu analizatora w wersji cyfrowej, znacznie nowocześniejszej od obecnej analogowej w modelu ROWBIR 1. Trudności te polegają na niedostępności w kraju koniecznych dla tego typu analizatora cyfrowych elementów operacyjnych, jakimi są produkowane od dawna na świecie szybkie multiplikatory. Polski przemysł mikroelektroniczny nie produkuje jeszcze obecnie tego rodzaju elementów, ani nie planuje opracowania ich produkcji w dającej się przewidzieć przyszłości. Nie ma także centralnego importu tych elementów do Polski, chociaż są one w krajach zachodnich łatwo osiągalne i niedrogie i w wielu dziedzinach nauki i techniki szeroko stosowane. Sytuacja taka utrwala niestety lukę techniki polskiej w bardzo ważnej obecnie dziedzinie jaką jest przetwarzanie sygnałów akustycznych, co wpłynie niestety hamująco także na wdrożenie polskich osiągnięć naukowych w automatycznym rozpoznawaniu mowy.

10. Propozycja metody rozpoznawania wyrazów w sposób segmentalny w oparciu o binarną reprezentację sygnału mowy

W globalnym rozpoznawaniu wyrazów pomija się fakt, że wyraz jest ciągiem podstawowych elementów fonetyczno-akustycznych i traktuje się go w całości jako jeden element. Wypowiedź każdego wyrazu, izolowana przerwami od innych wypowiedzi lub jakichkolwiek dźwięków nie będących mową, uważana jest za jednostkę podlegającą rozpoznawaniu w sposób globalny. Takie po-

dejście do rozpoznawania wyrazów pochodzi z negatywnych opinii o możliwości automatycznego segmentowania mowy na podstawowe elementy fonetyczno-akustyczne. Segmentację próbuje się zwykle przeprowadzać poprzez określenie granic międzysegmentalnych [4], [5], [58], [59]. Ponieważ wyznaczenie tych granic w wyrazie służyć ma identyfikacji segmentów, odbywać się musi w warunkach ich nieznajomości, a więc w oparciu o zasady ogólne bez odwoływania się do cech poszczególnych segmentów. Brak niestety generalnych zasad, które gwarantowałyby trafne i jednoznaczne wskazanie każdej granicy pomiędzy powszechnie uznawanymi elementami fonetyczno-akustycznymi. Wskazanie granicy między niektórymi segmentami jest proste, pomiędzy innymi może być kłopotliwe lub wręcz niemożliwe. Z tego powodu odsunięto na dalszy plan rozpoznawanie segmentalne, mimo iż nie brak wielu względów świadczących o jego bezspornej przewadze nad rozpoznawaniem globalnym.

Podstawową zaletą rozpoznawania mowy w sposób segmentalny jest to, że wymaga ono znajomości krótkich wzorców jedynie kilkuset elementów fonetyczno-akustycznych i to bez względu na zakres rozpoznawanego materiału, np. wielkość słownika w rozpoznawaniu wyrazów. Drugą zaletą tego rodzaju rozpoznawania jest możliwość użycia go zarówno do rozpoznawania izolowanych wyrazów, jak i mowy ciągłej. Na korzyść rozpoznawania segmentalnego przemawia też to, że nie wymaga ono odizolowania wypowiedzi marginesami ciszy.

Ponieważ rozpoznawanie globalne ma swoje granice i przestaje być opłacalne, gdy w grę wchodzi duży słownik wyrazów lub ciągi wyrazowe albo zdania, nieuniknione wydaje się zastąpienie go w pewnym momencie rozpoznawaniem segmentalnym. Praca niniejsza poświęcona została metodzie rozpoznawania wyrazów na podstawie spektrogramów binarnych w sposób globalny.

W obecnym rozdziale przedstawiona zostanie pewna teoretyczna koncepcja przystosowania tej metody do rozpoznawania wyrazów w sposób segmentalny. Propozycja ta określa kierunek, w jakim należałoby w przyszłości kontynuować badania nad rozpoznawaniem mowy w oparciu o spektrogramy binarne. Metoda, która zostanie przedstawiona poniżej, dotyczy rozpoznawania segmentalnego mowy bez potrzeby wyznaczania granic międzysegmental-

nych, czyli znajomości położenia początku i końca elementów fonetyczno-akustycznych w rozpoznawanym obiekcie. Zakłada się, że tą metodą rozpoznawany będzie dowolny fragment sygnału mowy o rozciągłości nie mniejszej niż 1 lub 2 fonemy, oraz że wynik będzie miał formę ciągu umownych kodów zidentyfikowanych elementów akustyczno-fonetycznych.

Załóżmy, że istnieje inwentarz wzorców fonetyczno-akustycznych elementów mowy. Niech pozostanie chwilowo sprawą otwartą, jakich konkretnie elementów inwentarz ten dotyczy. Z przyczyn podanych w rozdziale 4 nie będą tymi elementami fonemy, lecz prawdopodobnie połączenia międzyfonemowe, czyli segmenty przejściowe pomiędzy kolejnymi fonemami.

Zakłada się, że identyfikacja poszczególnych elementów w obrazie rozpoznawanej wypowiedzi przebiegać będzie drogą porównania obiektu ze wzorcami wszystkich elementów. Z liczb wyrażających lokalne podobieństwa obiektu ze wszystkimi wzorcami utworzyć można macierz. Niech elementy macierzy tworzące kolumny o numerach:

$$M_{v-1} + 1, \ldots\ldots, M_{v-1} + k_v$$

będą wartościami ilorazu podobieństwa poszczególnych fragmentów v-tego wzorca kolejno do wszystkich fragmentów obiektu.

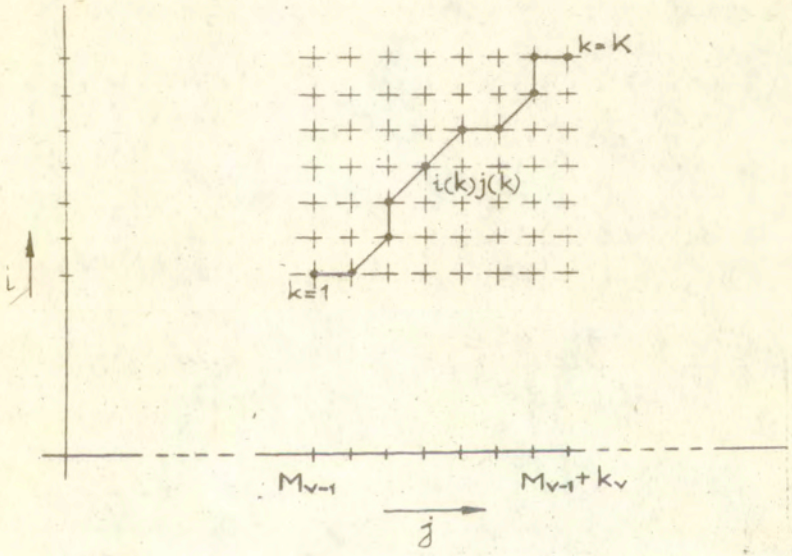
$$M_{v-1} = \sum_{j=1}^{v-1} k_j$$

wyraża numer ostatniego fragmentu wzorca zajmującego w inwentarzu pozycję v-1. k_j i k_v są odpowiednio ilościami fragmentów we wzorcach o numerach kolejnych wyrażonych przez j i v. Jeżeli w obiekcie występuje element fonetyczno-akustyczny, którego dotyczy v-ty wzorzec, wówczas w części macierzy składającej się z kolumn miejsca elementów q_{ij} o wartościach lokalnie najniższych w poszczególnych kolumnach powinny układać się w trajektorię wytyczoną następującymi zależnościami:

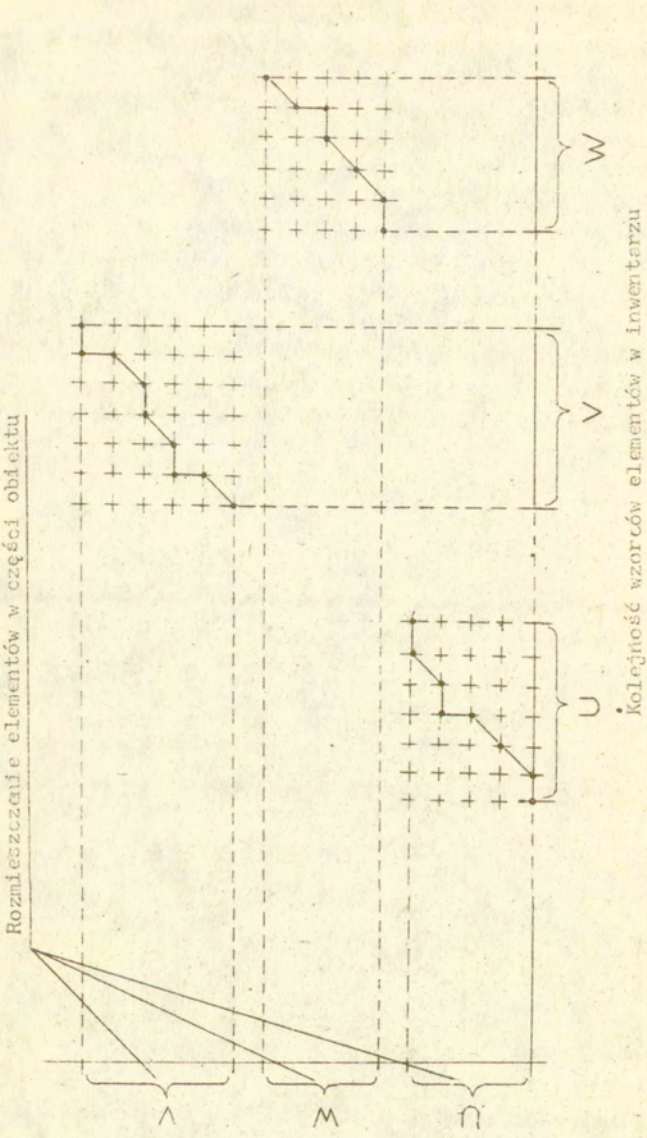
$$\left.\begin{aligned} i(k+1) &= \big(i(k)+1\big) \;\; v \;\; i(k+1) = i(k) \Longleftrightarrow i(k) \neq i(k-1) \\ j(k+1) &= \big(j(k)+1\big) \;\; v \;\; j(k+1) = j(k) \Longleftrightarrow j(k) \neq j(k-1) \end{aligned}\right\} \quad 10.1$$

gdzie k oznacza kolejne punkty wyznaczające trajektorię.

Przyjmuje się jak dotychczas, że iloraz q_{ij} jest uzupełnieniem do 1 wartości miary podobieństwa wyrażonej wzorem (6.2). Stwierdzenie istnienia takiego układu minimalnych pod względem wartości elementów w części macierzy dotyczącej porównania obiektu z jednym z wzorców jest równoznaczne z rozpoznaniem w obiekcie fonetyczno-akustycznego elementu reprezentowanego przez ten wzorzec. Położenie tego elementu w obiekcie określają numery wierszy, w których przypadają lokalnie minimalne elementy wspomnianej części macierzy spełniające zależności (10.1). Lokalnie minimalne elementy są elementami o wartościach najniższych w części kolumny obejmującej jedynie pewien zakres wierszy. Ograniczenie wyboru elementów o wartościach najniższych w kolumnie jedynie do pewnego zakresu wierszy macierzy jest wuwzględnieniem możliwości wystąpienia wielokrotnie tego samego elementu fonetyczno-akustycznego w obiekcie.



Rys. 38. Ilustracja zasad identyfikacji jednego elementu fonetyczno-akustycznego w obiekcie według proponowanej metody segmentalnego rozpoznawania mowy



Rys. 39. Uproszczona ilustracja złożenia końcowego wyniku rozpoznania części wypowiedzi

Rys. 38 jest ilustracją wyżej opisanych zasad identyfikacji jednego elementu fonetyczno-akustycznego w obiekcie. W analogiczny sposób następuje rozpoznanie pozostałych elementów.

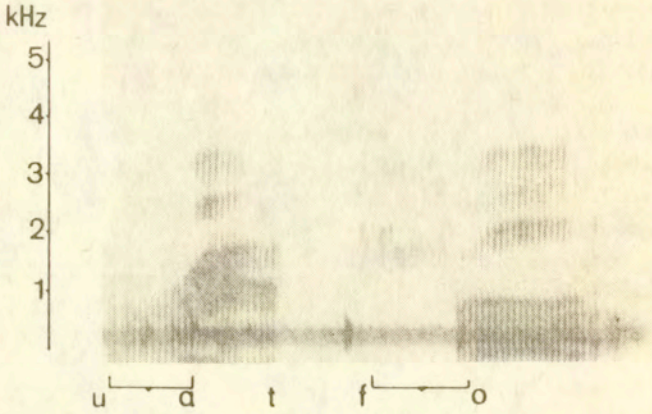
Równocześnie z rozpoznaniem w obiekcie różnych elementów fonetyczno-akustycznych następuje określenie ich wzajemnego położenia, co umożliwia zidentyfikowanie rozpatrywanej części wypowiedzi. Uproszczoną ilustracją złożenia końcowego wyniku rozpoznania jest przykład zamieszczony na rys. 39.

Wyjaśnienia wymaga kwestia, jakich elementów fonetyczno-akustycznych dotyczą wzorce w przedstawionej wyżej metodzie rozpoznawania mowy w sposób segmentalny i jak wzorce te należy tworzyć, aby uniknąć konieczności wyznaczania granic międzysegmentalnych.

Poza niektórymi wyjątkami każdy fonem występujący w kontekście innych fonemów pozostaje pod wpływem fonemu poprzedniego w swojej fazie początkowej i fonemu następnego w swojej fazie końcowej. Istnieje zatem wiele odmian kontekstowych każdego fonemu. Pomijając zjawiska łańcuchowe, teoretycznie jest ich łącznie $N \cdot (N-1) \cdot (N-2)$, gdzie N jest liczbą wszystkich fonemów. Gdyby więc przyjąć fonem jako podstawowy element w rozpoznawaniu mowy, inwentarz elementów byłby bardzo liczny i konieczne byłoby wyznaczanie granic międzyfonemowych, czego z wiadomych względów usiłuje się uniknąć. Gdyby jako podstawowy element mowy przyjąć zamiast fonemu półfonem, inwentarz elementów byłby znacznie mniejszy (teoretycznie $(N-2)$-krotnie, gdzie N jest liczbą wszystkich fonemów), lecz pozostałaby aktualna konieczność wyznaczania początku lub końca fonemu. Półfonem jest elementem zbyt krótkim i przez to trudniej rozpoznawalnym. Te okoliczności skłaniają do wniosku, że połączenie dwóch fonemów najwłaściwiej spełniać może rolę podstawowego elementu w rozpoznawaniu mowy. Za połączenie dwóch fonemów uważa się tutaj segment mowy składający się z części ustalonych oraz przypadających pomiędzy nimi części nieustalonych dwóch sąsiadujących ze sobą fonemów. Tak rozumiane połączenie dwóch fonemów, które odtąd w skrócie nazywane będzie połączeniem fonemów, zilustrowano na rys. 40.

Z pewnością ten punkt widzenia nie odpowiada dążeniu do osiągnięcia systemu rozpoznawania mowy opartego na liczbie ele-

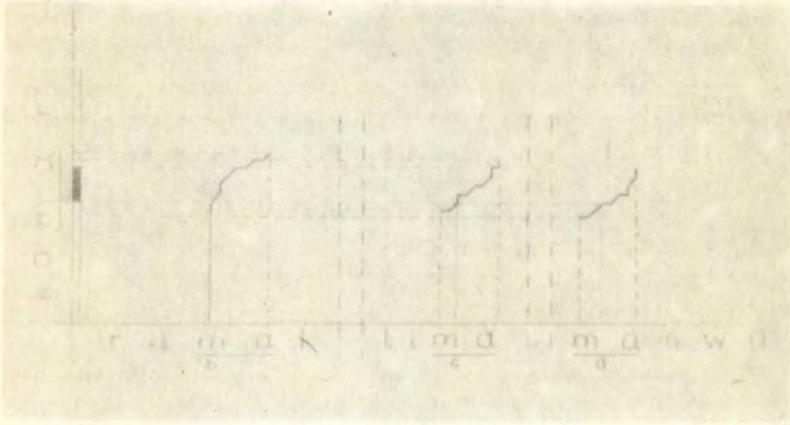
mentów zbliżonej do liczby fonemów. Uzyskanie takiego podziału mowy na elementy, jak pisma na litery, nie wchodzi jeszcze na razie w rachubę w rozpoznawaniu automatycznym mowy.



Rys. 40. Obraz spektrograficzny połączenia dwóch fonemów

Wyznaczenie wzorca połączenia dwóch fonemów poprzedzone być musi określeniem wspólnej rozciągłości różnych realizacji danego typu połączenia w różnych ciągach fonemów. Ustalenie wspólnej rozciągłości następuje poprzez wyznaczenie macierzy wartości lokalnych podobieństw jednego z ciągów fonemowych, zawierającego dane połączenie oraz kilku innych ciągów, w któtych także występuje to samo połączenie. Na przykład w celu wyznaczenia wzorca połączenia fonemów [m] i [a] należy posłużyć się szeregiem ciągów fonemowych zawierających to połączenie. Takimi ciągami mogą być zarówno logatomy, jak i wyrazy. W rozpatrywanym przykładzie niech nimi będą wyrazy: [komar], [rumak], [lima], [maswo]. Dla par tych wyrazów utworzonych według zasady, że jeden wyraz w każdej parze jest ten sam, wyznaczone zostają macierze wartości podobieństw lokalnych. Elementami takich macierzy są, jak wiadomo, wartości ilorazu podobieństwa

każdego fragmentu reprezentacji binarnej jednego wyrazu z poszczególnymi fragmentami analogicznej reprezentacji drugiego wyrazu. Oczekuje się, że te elementy macierzy, które wyrażają optymalne podobieństwa lokalne w obrębie identycznych połączeń fonemowych zajmowac będą pozycje (ij) określone kryteriami (10.1). Zakładając, że fragmenty reprezentacji wyrazu [komar] powtarzającego się we wszystkich parach mają numerację zgodną z wierszami (i), a fragmenty reprezentacji pozostałych wyrazów wchodzących w skład poszczególnych par mają numerację zgodną z kolumnami (j), przyjmuje się jako rozciągłośc wspólną połączenia fonemowego [ma] zakres wierszy macierzy, w obrębie których plasują się elementy wskazujące na największe podobieństwo odpowiadających sobie fragmentów dla wszystkich par użytych wyrazów zawierających połączenie [ma]. Zasady wyboru rozciągłości wspólnej połączenia fonemowego zilustrowano na rys. 41.



Rys. 41. Ilustracja zasady ustalania rozciągłości wzorca danego połączenia fonemowego

Wspólna rozciągłośc połączenia fonemowego staje się długością przyszłego wzorca tego połączenia.

Sam wzorzec powinien byc odzwierciedleniem cech, jakie wy-

stąpiły przynajmniej w połowie binarnych reprezentacji różnych realizacji danego połączenia fonemowego. Zostanie to spełnione jeżeli wartości poszczególnych parametrów w kolejnych widmach binarnych wzorca będą takie, jak w co najmniej połowie odpowiadających sobie widm w kolejnych grupach podobnych fragmentów, pochodzących z różnych realizacji danego połączenia fonemowego. Dla przykładu zilustrowanego na rys. 41 poszczególne parametry pierwszego widma binarnego wzorca przyjmują takie wartości, jakie mają odnośne parametry przynajmniej połowy odpowiadających sobie widm we fragmentach a, b, c, d. W ujęciu ogólnym można tę zasadę wartościowania parametrów wzorca wyrazić następująco: i-ty parametr, n-tego widma binarnego wzorca v-tego połączenia fonemowego p_{vni} przyjmuje wartość 1, jeżeli spełniony jest warunek:

$$\sum_{j=1}^{N} p_{kj\ n(i)} \geqslant \frac{N}{2}, \qquad (10.2)$$

w którego zapisie $p_{kj(n)i}$ oznacza i-ty parametr, k-tego widma binarnego. Widmo to należy do j-tej realizacji danego połączenia fonemowego i jest zgodne z n-tym widmem binarnym podstawowej realizacji tegoż połączenia. Przyjmuje się, że podstawowa realizacja danego połączenia fonemowego występuje w tym wyrazie, względem którego porównuje się inne wyrazy zawierające także to połączenie. W przykładzie na rys. 41 przyjęto, że podstawowa realizacja połączenia fonemowego [ma] zawarta jest w wyrazie [komar]. Na równi z różnymi wyrazami przedstawiającymi różne realizacje jednego połączenia fonemowego traktować też można wszelkie replikacje tych wyrazów, które warto uwzględnić dla zwiększenia reprezentatywności tworzonego wzorca danego połączenia fonemowego.

Wyznaczenie wzorców poszczególnych elementów fonetyczno-akustycznych mowy, takich np. jak połączenia fonemowe, jest zabiegiem znacznie bardziej złożonym niż wyznaczenie wzorców wyrazowych dla globalnego rozpoznawania wyrazów. Wymaga ono przede wszystkim opracowania długiej listy logatomów lub wyrazów zawierających rozmaite i wielokrotne realizacje wszystkich

rodzajów połączeń fonemowych występujących w danym języku. Ponieważ różnych połączeń dwufonemowych jest w języku polskim 1197 (W.Jassem, P.Łobacz, 1971), zaś dla dobrej reprezentatywności wzorców należy uwzględnić przy adaptacji około 4 - 6 realizacji każdego połączenia, lista adaptacyjna zawierać powinna około 5-7 tysięcy pozycji.

Przedstawioną tutaj metodę wyznaczania wzorców połączeń fonemowych jako podstawowych elementów mowy nazwać można adaptacją ukierunkowaną, gdyż prowadzi ona w sposób zamierzony do wyznaczenia wzorców z góry zadanych elementów. Prawdopodobnie możliwa byłaby również adaptacja naturalna, która polegałaby na automatycznym znalezieniu podobnych elementów w wypowiedzi tekstu odpowiednio długiego i wystarczająco reprezentatywnego dla danego języka. Ponieważ istnieje szeroki zakres rozciągłości elementów podobnych w mowie, od fragmentu fonemu począwszy aż po wyraz lub nawet ciąg wyrazów, jako elementy podstawowe przyjąć należałoby tylko te, których rozciągłość zawarta jest w określonych granicach. Adaptacja tego typu przebiegałaby bez ingerencji człowieka, a wzorce poszczególnych elementów zostałyby automatycznie zaetykietowane numerami. Ujawnienie, jakich elementów poszczególne wzorce dotyczą, mogłoby nastąpić dopiero w drugim etapie adaptacji, tzn. po utworzeniu wzorców wszystkich rodzajów elementów, które mieszczą się w założonym zakresie długości. Ten sposób adaptacji nie gwarantuje, że wyznaczone wzorce dotyczyć będą połączeń fonemowych.

O ile w globalnym rozpoznawaniu wyrazów wzorzec wyrazu reprezentuje wyraz wprost, o tyle w rozpoznawaniu wyrazów metodą segmentalną oprócz inwentarza wzorców elementów fonetyczno-akustycznych istnieć musi inwentarz ciągów kodów połączeń fonemowych dla poszczególnych wyrazów. Inwentarz taki zostaje wyznaczony w drugim etapie adaptacji, podczas którego następuje identyfikacja połączeń fonemowych (lub ogólniej elementów) występujących w wypowiedzi danego wyrazu. Ciąg zidentyfikowanych połączeń fonemowych zostaje zapamiętany w formie ciągu kodów etykietujących poszczególne połączenia, a także ciągu kodów znaków ortograficznych tworzących zapis tego wyrazu. W przypadku, gdy rozpoznawane mają być ciągi wyrazowe lub zdania, konieczny jest także inwentarz ciągów kodów połączeń fonemowych,

które we wspomnianych ciągach wyrazowych lub zdaniach występują. Ta ostatnia uwaga dotyczy uproszczonego systemu rozpoznawania wypowiedzi dłuższych niż pojedynczego wyrazu.

Przedstawiona w tym rozdziale koncepcja rozpoznawania wyrazów w sposób segmentalny w oparciu o reprezentację połączeń fonemowych za pomocą ciągu widm binarnych zostanie w przyszłości zrealizowana i poddana wnikliwej weryfikacji. Zachodzi już jednak obecnie pytanie, na ile ta nowa wersja rozpoznawania w oparciu o spektrogramy binarne może w zastosowaniu do rozpoznawania wyrazów okazać się korzystniejszą od metody globalnego rozpoznawania wyrazów, której poświęcono niniejszą pracę. Na pytanie to można częściowo udzielić odpowiedzi porównując np. dla obu metod wielkości obszarów pamięci koniecznych dla przechowania inwentarzy wzorców i ciągów kodów etykietujących różne połączenia fonemowe w poszczególnych wyrazach w rozpoznawaniu segmentalnym lub kodów etykietujących wzorce wyrazów w rozpoznawaniu globalnym, a także porównując ilości operacji składających się na proces rozpoznawania w obu metodach.

Porównanie wspomnianych wielkości obszarów pamięci sprowadza się do porównania objętości informacyjnej Kg i Ks wzorców i kodów etykietujących elementy w obu sposobach rozpoznawania. Wzorzec globalny (WG) wyrazu składa się przeciętnie z 35 16-parametrycznych widm binarnych, a zatem posiada objętość informacyjną:

$$K_{WG} = 35 \times 16 = 560 \text{ bitów.}$$

Przyjmując, że wzorzec jednego połączenia fonemowego (WPF) jest ciągiem średnio siedmiu widm binarnych, jego objętość informacyjna wynosi:

$$K_{WPF} = 7 \times 16 = 112 \text{ bitów.}$$

Kod etykietujący każdego z 1197 połączeń fonemowych wymaga słowa 11-bitowego. Wzorce wszystkich 1197 fonotaktycznie prawidłowych w języku polskim połączeń dwufonemowych oraz ich kody etykietujące posiadają łącznie objętość informacyjną:

$$K_{WWPF} = 123 \times 1197 = 147231 \text{ bitów.}$$

Przy założeniu, że przeciętny wyraz polski składa się z siedmiu połączeń fonemowych, ciąg kodów (CK) tych połączeń posiada objętość informacyjną:

$$K_{CKPF} = 11 \times 7 = 77 \text{ bitów.}$$

Stanowi on umowną etykietę wyrazu.

Podobnie szacunkowo przyjmując, że zapis przeciętnego wyrazu polskiego składa się z ośmiu liter, z których każda zakodowana jest w słowie 6-bitowym, dochodzi się do wniosku, że kod całego wyrazu będący np. ciągiem kodów znaków ortograficznych (CKZO) posiada objętość informacyjną:

$$K_{CKZO} = 8 \times 6 = 48 \text{ bitów.}$$

Dla rozpoznawania segmentalnego V wyrazów sumaryczna objętość informacyjna K_s wzorców wszystkich połączeń fonemowych i kodów etykietujących te wzorce oraz ich ciągów reprezentujących poszczególne wyrazy jest określona zależnością:

$$K_s = K_{WWPF} + (K_{CKZO} + K_{CKPF}) \cdot V .$$

Podobnie dla rozpoznawania globalnego V wyrazów sumaryczna objętość informacyjna K_G wzorców wszystkich wyrazów oraz ich kodów ortograficznych jest określona przez wyrażenie:

$$K_G = (K_{WG} + K_{CKZO}) \cdot V .$$

Stosunek objętości K_s i K_G wyraża wzór ogólny:

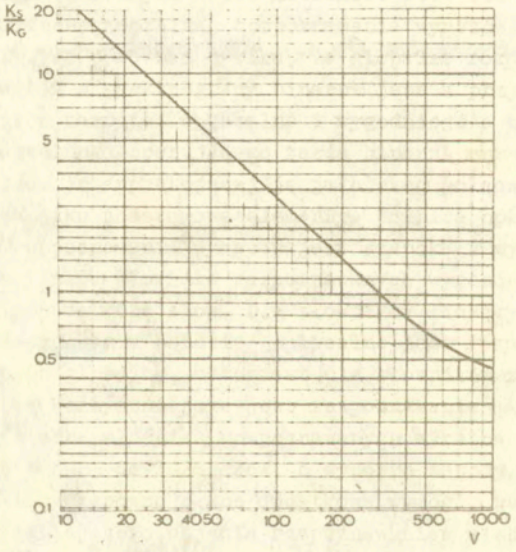
$$\frac{K_s}{K_G} = \frac{K_{CKPF} + K_{CKZO}}{K_{WG} + K_{CKZO}} \cdot \left(\frac{K_{WWPF}}{(K_{CKPF} + K_{CKZO}) \cdot V} + 1 \right), \qquad (10.3)$$

który po podstawieniu wyżej podanych wartości w miejsce literowych symboli poszczególnych objętości informacyjnych K_{WWPF},

K_{WG}, K_{CKZO}, K_{CKPF} przyjmuje postać:

$$K = 0.206 + \frac{1}{V} \cdot 242.1.$$

Na rys. 42 zamieszczony jest wykres funkcji K(V) ilustrujący która z rozpatrywanych dwóch metod rozpoznawania wyrazów i dla jak licznych słowników jest bardziej opłacalna z uwagi na zapotrzebowanie na pamięć do przechowywania inwentarza wzorców i kodów etykietujących różne elementy w rodzaju połączeń fonemowych lub ich ciągów oraz wyrazów. Gdy słownik rozpoznawanych wyrazów liczy mniej niż 305 słów, korzystniejsza jest pod względem zapotrzebowania na pamięć metoda globalnego rozpoznawania wyrazów. Dla słownika liczniejszego opłacalniejsza z tego punktu widzenia staje się metoda rozpoznawania wyrazów w sposób segmentalny.



Rys. 42. Wykres stosunku objętości informacyjnej inwentarza wzorców oraz kodów etykietujących elementy i ciągi elementów w rozpoznawaniu wyrazów na podstawie spektrogramów binarnych w sposób segmentalny i globalny

Podstawę oceny porównawczej obu metod stanowić też może stosunek ilości operacji wykonywanych w procesach rozpoznawania wyrazów według każdej z tych dwóch metod. Pewne operacje w obu rozpatrywanych metodach rozpoznawania wyrazów są identyczne, inne natomiast uważać należy za charakterystyczne dla danej metody. Do działań identycznych zalicza się porównywanie fragmentów oraz selekcja optymalnych podobieństw lokalnych. Działania te, w gruncie rzeczy proste, są w porównaniu z pozostałymi działaniami obszerniejsze i dlatego porównanie ich liczby dla obu metod pomoże wskazać, która z tych metod i w jakim zakresie jest korzystniejsza. W metodzie segmentalnej ilość tych działań L_s jest iloczynem:

$$L_s = L_{FRO} \cdot L_{FRWPF} \cdot L_{WPF} \tag{10.4}$$

liczby fragmentów L_{FRO} rozpoznawanego obiektu oraz liczby fragmentów $(L_{FRWPF} \cdot L_{WPF})$ wzorców wszystkich połączeń fonemowych. Przez L_{FRWPF} oznaczono liczbę fragmentów, z jakich składa się przeciętnej długości wzorzec połączenia fonemowego, a przez L_{WPF} liczbę wzorców wszystkich połączeń fonemowych.

W rozpoznawaniu globalnym liczba L_G działań identycznych jak w metodzie segmentalnej rozpoznawania wyrazów wyraża się iloczynem:

$$L_G = L_{FRO} \cdot L_z \cdot V \,, \tag{10.5}$$

w którym L_{FRO} oznacza to samo, co w iloczynie (10.4), L_z wyraża liczbę fragmentów mieszczących się w dopuszczalnym zakresie odchyleń od liniowej normalizacji czasowej rozpoznawanego obiektu i poszczególnych wzorców wyrazowych, zaś V symbolizuje ilość słów słownika rozpoznawanych wyrazów. Tak więc stosunek ilości działań identycznych w obu rozpatrywanych metodach rozpoznawania wyrazów określa wzór:

$$L = \frac{L_s}{L_G} = \frac{L_{FRWPF} \cdot L_{WPF}}{L_z \cdot V} \,. \tag{10.6}$$

Jak wiadomo $L_{WPF} = 1197$, $L_z = 7$. Przy założeniu, że $L_{FRWPF} = 6$, wzór ostatni przybiera konkretną postać:

$$L = \frac{1026}{V} .$$

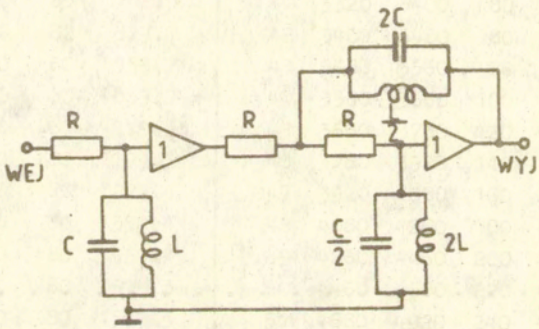
Okazuje się więc, że ze względu na liczbę operacji opłacalność rozpoznawania segmentalnego wyrazów w zaproponowanej wersji zaczyna się od 1026 słów, a więc od ilości dużą większej niż wyniósł próg opłacalności określony wcześniej w oparciu o porównanie zapotrzebowania na pamięć do przechowywania inwentarza wzorców i kodów etykietujących różne elementy dla obu rozpatrywanych metod. Próg opłacalności segmentalnego sposobu rozpoznawania wyrazów wynosi według kryterium zapotrzebowania na pamięć 305 słów, a według kryterium liczby operacji co najmniej 1026 słów. Ostatnia z tych dwóch liczb może być nawet wyższa, gdyż w przeprowadzonym porównaniu ilości operacji nie uwzględniono działań odmiennych w obu metodach. Jest ich znacznie więcej w metodzie segmentalnej rozpoznawania wyrazów.

Z powyższych porównań wynika, że globalne rozpoznawanie wyrazów w oparciu o spektrogramy binarne jest w przypadku niedużych słowników bardziej opłacalne i wygodniejsze od rozpoznawania segmentalnego operującego także binarną reprezentacją wypowiedzi, lecz interpretującej ją jako ciąg połączeń fonemowych traktowanych jako podstawowe elementy. Granica opłacalności globalnego rozpoznawania wyrazów jest szeroka i wyraża się zakresem ilości słów słownika rozpoznawanych wyrazów rozciągającym się od 305 do 1026 słów. Dla słowników liczących mniej niż 305 słów zdecydowanie bardziej opłacalne jest rozpoznawanie wyrazów w sposób globalny, natomiast dla słowników mających ponad 1026 słów korzystniejsze jest rozpoznawanie segmentalne wyrazów. Gdy liczba słownika rozpoznawanych wyrazów zawarta jest w granicach od 305 do 1026 słów, o wyborze metody właściwej decydować powinny zarówno pewne dodatkowe wymogi, takie jak np. możliwość późniejszego objęcia rozpoznawaniem większej liczby słów, krótki czas rozpoznawania, łatwość adaptacji do dowolnego głosu, jak i parametry dostępnego mikrokomputera, np. duża pamięć przy wolnym tempie działania lub też odwrotnie - mała pa-

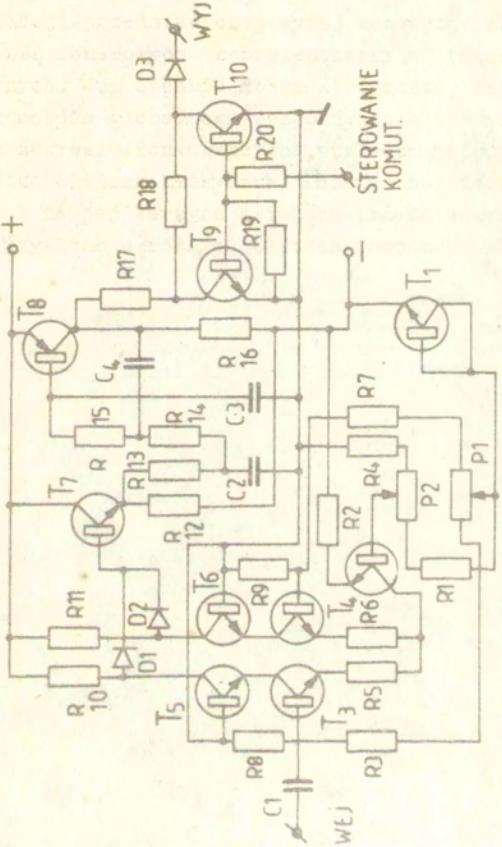
mięć i duża szybkość działania.

Przeprowadzona tutaj ocena porównawcza obu metod rozpoznawania wyrazów powinna zostać w przyszłości uzupełniona porównaniem wyników rozpoznawania uzyskiwanych dla każdej metody przy użyciu tego samego słownika. Będzie to możliwe dopiero po realizacji przedstawionej wyżej koncepcji modelu rozpoznawania połączeń fonemowych reprezentowanych także przez ciągi widm binarnych. Już obecnie można stwierdzić, że przedstawiona w tej pracy metoda globalnego rozpoznawania wyrazów, jest w dość szerokim zakresie konkurencyjna względem metody segmentalnej, uważanej za bliższą klasycznej fonetycznej teorii rozpoznawania mowy, a na jej korzyść świadczą przede wszystkim pozytywne wyniki uzyskane w różnych testach rozpoznawania.

D O D A T E K



Rys. 43 a. Schemat ideowy filtru środkowo-przepustowego w jednym kanale WAAW-u



Rys. 43 b. Schemat ideowy dedektora 1 kanałowego elementu komutatora w jednym kanale WAAW-u

Tablica 7. Zakresy pasm przepustowych oraz częstotliwości środkowe filtrów środkowo-przepustowych WAAW-u w [Hz]

Lp.	f_{-1}	f_{+1}	Δf	f_o	Lp.	f_{-1}	f_{+1}	Δf	f_o
1	120	200	80	154.9	33	2680	2760	80	2719.7
2	200	280	80	236.6	34	2760	2840	80	2799.7
3	280	360	80	317.5	35	2840	2920	80	2879.7
4	360	440	80	397.9	36	2920	3000	80	2959.7
5	440	520	80	478.3	37	3000	3080	80	3039.7
6	520	600	80	558.6	38	3080	3160	80	3119.7
7	600	680	80	638.7	39	3160	3240	80	3199.7
8	680	760	80	718.9	40	3240	3320	80	3279.7
9	760	840	80	798.9	41	3320	3400	80	3359.8
10	840	920	80	879.1	42	3400	3480	80	3439.8
11	920	1000	80	959.2	43	3480	3560	80	3519.8
12	1000	1080	80	1039.2	44	3560	3660	100	3609.6
13	1080	1160	80	1119.3	45	3660	3780	120	3719.5
14	1160	1240	80	1199.3	46	3780	3920	140	3849.4
15	1240	1320	80	1279.4	47	3920	4080	160	3999.2
16	1320	1400	80	1359.4	48	4080	4260	180	4169.0
17	1400	1480	80	1439.4	49	4260	4460	200	4356.8
18	1480	1560	80	1519.5	50	4460	4680	220	4568.7
19	1560	1640	80	1599.5	51	4680	4920	240	4798.5
20	1640	1720	80	1679.5	52	4920	5180	260	5048.3
21	1720	1800	80	1759.5	53	5180	5460	280	5318.1
22	1800	1880	80	1839.6	54	5460	5760	300	5607.9
23	1880	1960	80	1919.6	55	5760	6080	320	5917.8
24	1960	2040	80	1999.6	56	6080	6420	340	6247.7
25	2040	2120	80	2079.6	57	6420	6780	360	6597.5
26	2120	2200	80	2159.6	58	6780	7160	380	6967.4
27	2200	2280	80	2239.6	59	7160	7560	400	7357.3
28	2280	2360	80	2319.6	60	7560	7980	420	7767.2
29	2360	2440	80	2399.7	61	7980	8420	440	8197.0
30	2440	2520	80	2479.7	62	8420	8880	460	8646.9
31	2520	2600	80	2559.7	63	8880	9360	480	9116.8
32	2600	2680	80	2639.7					

BIBLIOGRAFIA

[1] ALLEN, J. : Implementation of Models for Speech Recognition with Very Large Scale Intergrated Circuit Technology, Automatic Speech Analysis and Recognition, Ed.: Haton, J.P., Dordrecht, Holland, 217-229, 1982.

[2] ATAL, B.S., SCHROEDER, M.R. : Predictive Coding of Speech Signals, Proc. 1967 Conf. Commun. and Process., 360-361, 1967.

[3] BALL, G.H., HALL, D.J. : Isodata - An Iterative Method of Multivariate Analysis and Pattern Classification, Proc. IFIPS Congr. 1965.

[4] BASZTURA, C., JURKIEWICZ, J., TYBURCY, E. : Zastosowanie fonetycznej funkcji mowy do segmentacji ciągłego sygnału mowy, Archiwum Akustyki, t. XIV, z. II, 121-130, Warszawa, 1979.

[5] BAUDRY, M., DUPEYRAT, B. : A Simple and Efficient Isolated Words Recognition System, Proc. IEEE ICASSP'82, Paris, Vol. 2, 879-882, 1982.

[6] BRIDLE, J.S., BROWN, M.D. : Connected Word Recognition Using Whole Word Templates, Proc. : Institute of Acoustics, Autum Conference, 25-28, November 1979.

[7] BRIDLE, J.S., BROWN, M.D., CHAMBERLAIN, R.M. : An Algorithm for Connected Word Recognition, Proc. IEEE ICASSP'82, Paris, Vol. 2, 899-902, 1982.

[8] BRIDLE, J.S., BROWN, M.D., CHAMBERLAIN, R.M. : An Algorithm for Connected Word Recognition, Automatic Speech Analysis and Recognition, ed. : Haton, J.P., Dordrecht, Holland, 191-204, 1982.

[9] BROWN, M.K., RABINER, L.R. : An Adaptive, Ordered Graph Search Technique for Dynamic Time Warping for Isolated

Word Recognition, IEEE Trans. on ASSP, Vol. ASSP-30, No. 4, 535--544, 1982.

[10] DAS, S.K. : Some Experiments in Discrete Utterance Recognition, IEEE Trans. on ASSP, Vol. ASSP-30, No. 5, 766-770, 1982.

[11] DAVIS, K.H., BIDDULPH, R., BALASHEK, J. : Automatic Recognition of Spoken Digits, JASA, Vol. 24, 637-645, 1952.

[12] DAVIS, S.B., MERMELSTEIN, P. : Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences, IEEE Trans. on ASSP, Vol. ASSP-28, 357-366, 1980.

[13] DENES, P., MATHEWS, M.V. : Spoken Digit Recognition Using Time-Frequency Patterns Maching, JASA, Vol. 32, 1450--1455, 1960.

[14] DERSCH, W.C. : Shoebox - A Voice Responsive Machine, Datamation, Vol. 8, 47-50, 1962.

[15] DOMAGAŁA, P. : Automatyzacja procesu segmentacji sygnału mowy w układzie analogowo-cyfrowym, Prace IPPT 5/1984, Warszawa, 1984.

[16] DREYFUS-GRAF, J. : Sonagrapf and Sound Mechanics, JASA Vol. 22, 731-739, 1949.

[17] DUDLEY, H., BALASHEK, S. : Automatic Recognition of Phonetic Patterns in Speech, JASA, Vol. 30, 721-739, 1958.

[18] FANT, G.C.M. : Acoustic Theory of Speech Production, Mouton and Co., 's-Gravenhage, The Netherlands, 1960.

[19] FANT, G.C. : Speech Sounds and Features, Cambridge, MA., M.I.T. Press, 1973.

[20] FLANAGAN, J.L. : Speech Analysis Synthesis and Perception, Springer-Verlag, Berlin, Heidelberg, New York, 1972.

[21] FLANAGAN, J.L., LEVINSON, S.E., RABINER, L.R., ROSENBERG, A.E. : Techniques for Expanding the Capabilities of

Practical Speech Recognizers, Trends in Speech Recognition, Ed. Lea, W., Englewood Cliffs, NJ, Prentice-Hall, 425-444, 1980.

[22] FORGIE, J.W., FORGIE, C.D. : Results Obtained from a Vowel Recognition Computer Program, JASA, Vol. 31, 1480-1489 1959.

[23] FORGIE, J.W., FORGIE, C.D. : Automatic Method of Plosive Identification, JASA, Vol. 34, 1979 A, 1962.

[24] FRY, D.B., DENES, P.B. : The Solution of Some Fundamental Problems in Mechanical Speech Recognition, Language and Speech, Vol. 1, 35-38, 1958.

[25] GAGNOULET, C., COUVRAT, M. : Seraphine : A Connected Word Speech Recognition System, Proc. IEEE ICASSP'82, Paris, Vol. 2, 887-890, 1982.

[26] GEMBIAK, D. : Rozpoznawanie segmentów wokalicznych i szumowych w ograniczonym zbiorze najczęstszych subiektywnie wyrazów polskich, Prace IPPT 69/1977, Warszawa, 1977.

[27] GRAY, A.H., MARKEL, J.D. : Distance Measures for Speech, Processing IEEE Trans. on ASSP, Vol. ASSP-24, 380-391, 1976.

[28] GREBLICKI, W. : Rekurencyjny, asymptotycznie optymalny algorytm uczenia rozpoznawania, Prace ICT Polit. Wrocł., nr 9, seria: Studia i Materiały, nr 8, 3-10, Wrocław, 1974.

[29] GREBLICKI, W. : Asymptotycznie optymalne algorytmy rozpoznawania i identyfikacji w warunkach probabilistycznych, Prace ICT Polit. Wrocł. nr 18, seria: Monografie, nr 3, Wrocław, 1974.

[30] GREER, K., LOWERRE, B., WILCOX, L. : Acoustic Pattern Matching and Beam Searching, Proc. IEEE ICASSP'82, Paris, Vol. 2, 1251-1254, 1982.

[31] GROCHOLEWSKI, S. : Programowanie dynamiczne w automatycznym rozpoznawaniu mowy, Elektrotechnika, t. 4, z. 1, 1985.

[32] GROCHOLEWSKI, S. : Metoda wykorzystania skrajnie ograniczonego sygnału mowy do rozpoznawania wybranych poleceń dyspozycyjnych, Rozprawa doktorska na Polit. Pozn., Instytut Automatyki, Poznań, 1981.

[33] GUBRYNOWICZ, R., KACPROWSKI, J., MIKIEL W., SKALSKI W. : Klasyfikacja spółgłosek trących metodą analizy przejść przez zero, Prace IPPT 40/1973, Warszawa, 1973.

[34] GUBRYNOWICZ, R. : Metoda przejść przez zero w analizie sygnału mowy i automatycznym rozpoznawaniu ograniczonego zbioru wyrazów, Prace IPPT 37/1974, Warszawa, 1974.

[35] GUBRYNOWICZ, R. : Automatyczne rozpoznawanie mowy w komunikacji człowiek - elektroniczna maszyna cyfrowa, Prace IPPT 71/1974, Warszawa, 1974.

[36] HATON, J.P., LOMOTTE, I.M. : Un algorithme de comparaison dynamique pour la reconnaissance automatique de la parole et son application pratique, Automatisme, t. XIX, 284--289, 1974.

[37] HERSCHER, M.B., COX, R.B. : An Adaptive Isolated Word Speech Recognition System, Conf. Speech Commun. Process., AD--742236, 89-92, 1972.

[38] HILL, D.R. : An ESOTerIc Approach to some Problems in Automatic Speech Recognition,International Journal of Man-Machine Studies, Vol. 1, 101, 1969.

[39] HUGHES, G.W. : The Recognition of Speech by Machine, Technical Report 395, Research Laboratory of Electronics,M.I.T. Cambridge, MA., 1961.

[40] ITAKURA, F. : Minimum Prediction Residual Principle Applied to Speech Recognition, IEEE Trans. on ASSP, Vol. ASSP--23, No. 1, 67-72, 1975.

[41] JASCHUL, J. : An Approach to Speaker Normalization for Automatic Speech Recognition, ICASSP'79, Washington DC., 235--238, Apr. 1979.

[42] JASCHUL, J. : Estimation of Speaker-Specific Adaptation Parameters, The Fourth F.A.S.E. Symposium, Venezia, 259--262, Apr. 1981.

[43] JASCHUL, J. : Speaker Adaptation by A Linear Transformation with Optimised Parameters, Proc. IEEE ICASSP'82, Paris, Vol. 2, 1657-1660, 1982.

[44] JASSEM, W. : The Distinctive Features of Polish Phonemes, STL QPSR, 1/62, 7-14, Stockholm, 1962.

[45] JASSEM, W. : The Formant Patterns of Fricative Consonants, STL QPSR, 3/62, 6-15, Stockholm, 1962.

[46] JASSEM, W. : Acoustical Description of Voiceless Fricatives in Terms of Spectral Parameters, Speech Analysis and Synthesis /ed. : Jassem, W./, Vol. 1, 189-206, 1968.

[47] JASSEM, W. : Fonetyczno-akustyczne założenia automatycznego rozpoznawania fonemów, Prace IPPT 17/70, Warszawa, 1970.

[48] JASSEM, W., ŁOBACZ, P. : Analiza fonotaktyczna tekstu polskiego, Prace IPPT, 63/71, Warszawa, 1971.

[49] JASSEM, W., KRZYŚKO, A., DYCZKOWSKI, A. : Klasyfikacja i identyfikacja samogłosek polskich na podstawie częstotliwości formantów, Prace IPPT 64/72, Warszawa 1972.

[50] JASSEM, W., KRZYŚKO, M., DYCZKOWSKI, A. : Identyfikacja izolowanych samogłosek polskich, Archiwum Akustyki, t.IX, z. III, 261-287, Warszawa, 1974.

[51] JASSEM, W., KRZYŚKO, M., DYCZKOWSKI, A. : Sekwencyjna identyfikacja samogłosek, Prace IPPT 6/74, Warszawa, 1974.

[52] JASSEM, W., SZYBISTA, D., DYCZKOWSKI, A. : Rozpoznawanie samogłosek polskich w typowych zdaniach, Prace IPPT, 43/75, Warszawa, 1976.

[53] JASSEM, W., SZYBISTA, D., KRZYŚKO, M., STOLARSKI, P., DYCZKOWSKI, A. : Rozpoznawanie polskich spółgłosek trących na

podstawie cech widmowych, Prace IPPT 46/76, Warszawa, 1976.

[54] JASSEM, W.: Podstawy fonetyki akustycznej, PWN, Warszawa, 1977.

[55] JASSEM, W. : Założenia ogólnego modelu rozpoznawania mowy, Prace IPPT 68/77, Warszawa, 1977.

[56] JASSEM, W. : Prace nad automatycznym rozpoznawaniem mowy w Polsce, Prace IPPT 67/72, Warszawa, 1972.

[57] JASSEM, W., GEMBIAK, D., DYCZKOWSKI, A. : Wspomagane przez komputer rozpoznawanie samogłosek polskich w mowie ciągłej, Archiwum Akustyki, t. XIV, z. I, 41-52, Warszawa, 1979.

[58] JASSEM, W., KUBZDELA, H., DOMAGAŁA, P. : Segmentacja sygnału mowy na podstawie zmian rozkładu energii w widmie, Prace IPPT 13/83, Warszawa, 1983.

[59] JASSEM, W., KUBZDELA, H., DOMAGAŁA, P. : Automatic acoustic-phonetic Segmentation of the Speech Signal, Acta Universitatis Umensis, From Sound to Words, Umea, 1983.

[60] JARVIS, R.A., PATRICK, E.A. : Clustering Using a Similarity Measure Based on Shared Near Neighbors, IEEE Trans. on Comput., Vol. C-22, 1025-1034, 1973.

[61] KACPROWSKI, J. : Teoretyczne podstawy syntezy samogłosek polskich w rezonansowych układach formantowych, Rozprawy Elektroniczne, t. VIII, 1962.

[62] KACPROWSKI, J. : Synteza polskich spółgłosek nosowych w rezonansowych układach formantowych, Rozprawy Elektroniczne t. IX, nr 3, 439-465, 1963.

[63] KACPROWSKI, J., GUBRYNOWICZ, R. : Automatyczne rozpoznawanie samogłosek polskich metodą segmentacji widma, Prace IPPT 22/67, Warszawa, 1967.

[64] KACPROWSKI, J. : Teoretyczne podstawy procesu automatycznego rozpoznawania mowy, Archiwum Akustyki, 2, 123-151, 1967.

[65] KACPROWSKI, J. : Teoretyczne podstawy metody automatycznego rozpoznawania samogłosek, Archiwum Akustyki, 2, 227-254 1967.

[66] KACPROWSKI, J. : Akustyczne aspekty problemu komunikacji człowiek-komputer w języku naturalnym, Archiwum Akustyki 7, 201-212, 1972.

[67] KACPROWSKI, J. : Akustyczny sygnał w komunikacji człowiek-komputer, Prace IPPT, 3/72, Warszawa, 1972.

[68] KACPROWSKI, J. : Sygnał akustyczny w procesach sterowania i diagnostyki, Archiwum Akustyki, 9, 375-388, 1974.

[69] KACPROWSKI, J. : Model symulacyjny kanału głosowego z uwzględnieniem zjawiska nazalizacji, Archiwum Akustyki, t. XII, z. 4, 281-302, 1977.

[70] KACPROWSKI, J. : Fizyczne modele źródła krtaniowego, Archiwum Akustyki, t. XII, z. 1, 47-70, 1977.

[71] KACPROWSKI, J. : Obiektywne metody akustyczne w diagnostyce narządu głosu, Archiwum Akustyki, t. XIV, z. 4, 287--304, 1979.

[72] KELLEY, T.P., MARTIN, J.T., BARGER, J.R. : Voice Controller for Astronaut Manuevering Unit, Technical Report AFAL-TR-68-308, Air Force Avionics Laboratory, Wright Patterson Air Force Base, OH, 1968.

[73] KOT, L. : Proste metody rozpoznawania obrazów w zastosowaniu do rozpoznawania elementów mowy polskiej, Materiały V Symp. MPN WEAiE AGH, 72-76, Kraków, 1979.

[74] KOT, L. : Ocena przydatności analizy pasmowej do rozpoznawania prostych elementów mowy polskiej, Praca doktorska na AGH, Instytut Informatyki i Automatyki, Kraków 1980.

[75] KUBZDELA, H. : An Automatic Formant Frequency Extractor, Speech Analysis and Synthesis /ed. Jassem, W./, Vol. 2, 209-220, PWN, Warszawa, 1970.

[76] KUBZDELA, H. : Automatyczna ekstrakcja częstotliwości tonu podstawowego oraz pierwszych trzech formantów sygnału mowy, Prace IPPT 51/73, Warszawa, 1973.

[77] KUBZDELA, H. : Analogowe ekstraktory częstotliwości formantów EXFOR-2 i EXFOR-3, Materiały XX OSA, 1, 150-154, Mierzyn, 1973.

[78] KUBZDELA, H. : Techniczna realizacja formantowej metody rozpoznawania samogłosek polskich, Prace IPPT 90/75, Warszawa, 1975.

[79] KUBZDELA, H. : Exfor-2 - An Analogue Tracker of the First three Formants, Speech Analysis and Synthesis /ed.Jassem W./, Vol. 4, 281-292, PWN, Warszawa, 1976.

[80] KUBZDELA, H. : Model analogowo-cyfrowego układu rozpoznawania samogłosek polskich w uproszczonych ciągach fonemowych, Materiały XXIII OSA, 1, 217-218, Wisła, 1976.

[81] KUBZDELA, H. : Wyznaczanie charakterystycznego fragmentu samogłoskowego i pomiar częstotliwości formantów dla automatycznej klasyfikacji i identyfikacji samogłosek, Prace IPPT 41/79, Warszawa, 1979.

[82] KUBZDELA, H.: Metoda automatycznego rozpoznawania wyrazów w oparciu o spektrogramy binarne, Prace IPPT 14/80, Warszawa, 1980.

[83] KUBZDELA, H. : Automatyczne rozpoznawanie wyrazów na podstawie spektrogramów binarnych, Prace IPPT 15/81, Warszawa 1981.

[84] KUBZDELA, H. : Weryfikacja i optymalizacja metody rozpoznawania wyrazów w skończonych zbiorach hasłowych w oparciu o spektrogramy binarne, Prace IPPT 10/82, Warszawa, 1982.

[85] KUBZDELA, H. : Badania nad udoskonaleniem spektrogramów binarnych, Prace IPPT 24/83, Warszawa, 1983.

[86] KUBZDELA, H. : Worterkennung auf Grund der binären Sonagramme, 11th, ICA Abstracts, Revue d'acoustique, 4, 207-210,

Paris, 1983.

[87] KUBZDELA, H. : Próby automatycznego rozpoznawania wyrazów wymawianych przez różne głosy w oparciu o grupowe zbiory wzorcowych spektrogramów binarnych, Prace IPPT 47/83, Warszawa 1983.

[88] KUBZDELA, H. : Rozpoznawanie wyrazów w oparciu o uproszczoną reprezentację binarną sygnału mowy, Materiały XXXII OSA, 65-68, Kraków, 1985.

[89] LAMEL, L.F., ZUE, V.W. : Performance Improvement in a Dynamic-Programming-Based Isolated Word Recognition System for the Alfa-Digit Task, Proc. IEEE ICASSP'82, Paris, Vol. 1, 558-561, 1982.

[90] LEA, W.A. : Speech Recognition: Past, Present and Future, Trends in Speech Recognition, Ed. : Lea, W.A., Englewood Cliffs, NJ, Prentice-Hall, 39-98, 1980.

[91] LEVINSON, S.E., ROSENBERG, A.E., FLANAGAN, J.L.: Evaluation of a Word Recognition System Using Syntax Analysis, Proc. IEEE Int. Conf. on ASSP, 483-486, May 1977.

[92] LEVINSON, S.E., RABINER, L.R., ROSENBERG, A.E., WILPON, J.G. : Interactive Clustering Techniques for Selecting Speaker-Independent Reference Templates for Isolated Word Recognition, IEEE Trans. on ASSP, Vol. ASSP-27, 134-141, 1979.

[93] MAKHOUL, J.I. : Linear Prediction in Automatic Speech Recognition, In Reddy, 183-220, 1975 a.

[94] MAKHOUL, J.I. : Linear Prediction: A Tutorial Review, Proc. IEEE Special Issue on Digital Signal Processing, Vol. 63, 561-580, 1975 b.

[95] MARKEL, J.D., GRAY, A.H. : Linear Prediction of Speech, Springer-Verlag, Berlin, Heidelberg, New York, 1976.

[96] MARTIN, T.B., NELSON, A.L., ZADELL, A.J. : Speech Recognition by Feature Abstraction Techniques, Wright-Paterson AFB Avionics Laboratories Report, Dayton, Ohio, 1964.

[97] MARTIN, T.B., ZADELL, H., GRUNZA, E., HERSCHER, M.: Numeric Speech Translating Machine, Automatic Pattern Recognition, Washington, D.C. : National Security Industrial Association, 113-141, 1969.

[98] MARTIN, T.B. : Practical Applications of Voice Input Machines, Proc. IEEE 64, 487-501, 1976.

[99] MARTIN, T.B., WELCH, J.R. : Practical Speech Recognizers and Some Performance Effectiveness Parameters, Trends in Speech Recognition, Ed.: Lea, W.A., Englewood Cliffs, NJ : Prentice-Hall, 24-38, 1980.

[100] MERCIER, G. : Acoustic-Phonetic Decoding and Adaptation in Continuous Speech Recognition, Automatic Speech Analysis and Recognition, Ed.: Haton, J.P., Dordrecht, Holland, 69--99, 1982.

[101] MOTYLEWSKI, J., KACPROWSKI, J. : Automatyczne rozpoznawanie wyrazów metodą segmentacji widma sygnału mowy, Prace IPPT 29/69, Warszawa, 1969.

[102] MYERS, C., RABINER, L.R., ROSENBERG, A.E. : Performance Tradeoffs in Dynamic Time Warping Algorithms for Isolated Word Recognition : IEEE Trans. on ASSP, Vol. ASSP-28, No. 6, 623-635, 1980.

[103] MYTKOWSKI, K. : Kanał funkcji analogowych typ KF-01 do wprowadzania i wyprowadzania informacji w systemie „ON-LINE" do/z pamięci minikomputera MOMIK 8B/100, Prace IPPT 39/76,1976.

[104] NARA, Y., IWATA, K., KIJIMA, Y., KOBAYASHI, A., KIMURA, S., SASAKI, S., TANAHASHI, J. : Large-Vocabulary Spoken Word Recognition Using Simplified Time-Warping - Patterns , Proc. IEEE ICASSP'82, Paris, Vol. 2, 1266-1269, 1982.

[105] NIEDERJOHN, R.J. : A Mathematical Formulation and Comparison of Zero Crossing Technics which have been Applied to Automatic Speech Recognition, IEEE Trans. on ASSP, Vol.ASSP -23, No. 4, 1975.

[106] OKOCHI, M., SAKAI, T. : Trapezoidal DP Matching with Time Reversibility, Proc. IEEE ICASSP'82, Paris, Vol. 2, 1239--1242, 1982.

[107] OPPENHEIM, A.V., SCHAFER, R.M. : Homomorfic Analysis of Speech, IEEE Trans. on Audio and Electroacoustics, Au -16, No. 2, 27-31, 1968.

[108] PATRICK, E.A. : Fundamentals of Pattern Recognition, Englewood Cliffs, NJ : Prentice-Hall,1972.

[109] RABINER, L.R. : On Creating Reference Templates for Speaker Independent Recognition of Isolated Words, IEEE Trans. on ASSP, Vol. ASSP-26, No. 1, 34-42, 1978.

[110] RABINER, L.R., WILPON, J.G. : Considerations in Applying Clustering Techniques to Speaker Independent Word Recognition, JASA, Vol. 66, 663-673, 1979.

[111] RABINER, L.R., WILPON, J.G. : Applications of Clustering Techniques to Speaker-Trained Isolated Word Recognition, Bell Syst. Tech. J., Vol. 58, 2217-2233, 1979.

[112] RABINER, L.R., WILPON, J.G. : A Simplified Robust Training Procedure for Speaker Trained, Isolated Word Recognition Systems, JASA, Vol. 68, 1271-1276, 1980.

[113] RABINER, L.R., WILPON, J.G. : Isolated Word Recognition Using a Two-Pass Pattern Recognition Approach, Proc. Int. Conf. ASSP, Atlanta, Georgia, 724-727, 1981.

[114] REDDY, D.R. : Computer Recognition of Connected Speech, JASA, Vol. 42, No. 2, 329-347, 1967.

[115] REDDY, D.R. : Segment Synchronization Problem in Speech Recognition, JASA, Vol. 46, No. 1, p. 89, 1969.

[116] ROSENBERG, A.E., ITAKURA, F. : Evaluation of an Automatic Word Recognition System over Dialed-up Telephone Lines, JASA, Supl. 1 60, S. 12 A, 1976.

[117] ROSS, P.W. : A Limited-Vocabulary Adaptive Speech

Recognition System, Journal of the Audio Engineering Society, Vol. 15, 414–418, 1967.

[118] RUSKE, G. : An Efficient binary representation of Sonograms, Acustica, Vol. 34, No. 4, 234–239, Stuttgart, 1976.

[119] RUSKE, G., SCHOTOLA, T. : The Efficiency of Demi-syllable Segmentation in the Recognition of Spoken Words, Automatic Speech Analysis and Recognition, Ed. : Haton, J.P., Dordrecht, Holland, 153–163, 1982.

[120] SAITO, S., ITAKURA, F. : The Theoretical Consideration of Statistically Optimum Methods for Speech Spectral Density, Report No. 3107, Electrical Communication Laboratory, N.T.T., Tokyo, 1966, /w języku japońskim/.

[121] SAKOE, H., CHIBA, S. : A Dynamic Programming Approach to Continuous Speech Recognition, Proc. Intern. Congr. Acoust. Budapest, Hungary, Rep. 20-C-13, 1971.

[122] SAKOE, H., CHIBA, S. : Dynamic Programming Algorithm Optimization for Spoken Word Recognition, IEEE Trans. on ASSP, Vol. ASSP-26, No. 1, 43–49, 1978.

[123] SAKOE, H. : Two-Level DP Matching-A Dynamic Programming Based Pattern Matching Algorithm for Connected Word Recognition, IEEE Trans. Acoustics, Speech and Signal Processing Vol. ASSP-27, No. 6, 588–595, Dec. 1979.

[124] SAMBUR, M.R., RABINER, L.R. : A Statistical Decision Approach to the Recognition of Connected Digits, IEEE Trans. on ASSP, Vol. ASSP-24, 550–558, 1976.

[125] SHOUP, J.E. : Phonological Aspects of Speech Recognition, Trends in Speech Recognition, Ed.: Lea, W., Englewood Cliffs, NJ, Prentice-Hall, 125–138, 1980.

[126] SORENSON, H.W. : Least-Squares Estimation : From Gauss to Kalman, IEEE Spect. 7, 63–68, 1970.

[127] TADEUSIEWICZ, R. : Wybrane zagadnienia rozpoznawania obrazów dźwiękowych, Praca doktorska, Kraków, 1974.

[128] TADEUSIEWICZ, R. : Głosowa łączność człowieka z maszyną cyfrową, Zesz. Nauk. AGH, Automatyka, z. 22, Kraków, 1978.

[129] TEACHER, C.F., KELLETT, H.G., FOCHT, L.R. : Experimental Limited Vocabulary Speech Recognizer, IEEE Trans. on Audio and Electroacoustics, Vol. AU-15, 127-130, 1967.

[130] VELICHKO, V.M., ZAGORUJKO, N.G. : Automatic Recognition of 200 Words, Int. J. Man-Machine Stud., Vol. 2, p. 223, 1970.

[131] VINTSYUK, T.K. : Speech Recognition by Dynamic Programming Methods, Kibernetika, No. 1, 1968.

[132] VINTSYUK, T.K. : Element-weise Recognition of Continuous Speech Consisting of Words of a Given Vocabulary, Kibernetika, No. 2, 1971.

[133] WELCH, J.R. : Automatic Data Entry Analysis, RADC TR-77-306, Final Technical Report, 1977.

[134] WIENER, N. : Extrapolation, Interpolation and Smoothing of Stationary Time Series, M.I.T. Press Cambridge, Mass. 1966.

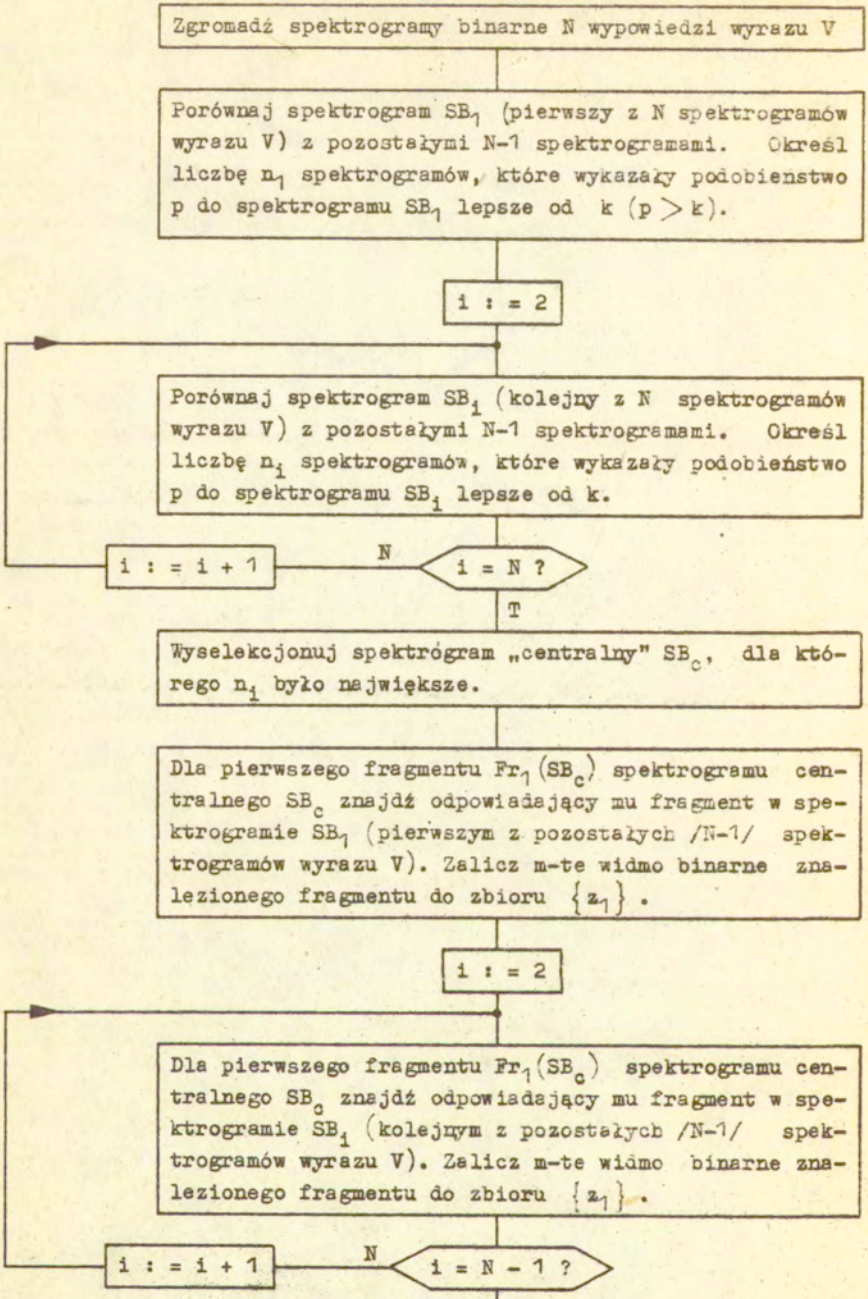
[135] WŁODARCZYK, H. : Podstawowe parametry fizyczne mowy w zastosowaniu do automatycznego rozpoznawania głosek i fonemów, Prace IA PAN, nr 94, Warszawa, 1971.

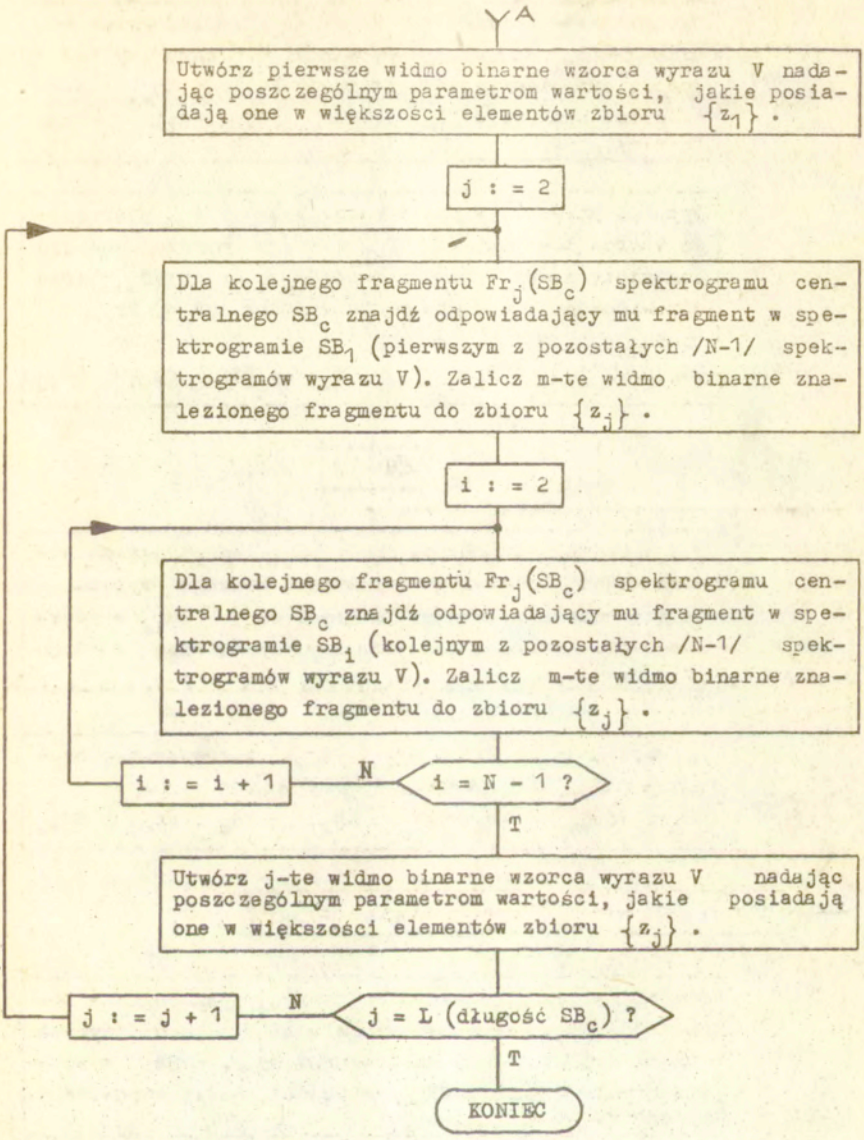
[136] ZUE, V.W., SCHWARTZ, R.M. : Acoustic Processing and Phonetic Analysis, Trends in Speech Recognition, Ed. : Lea, W. Englewood Cliffs, NJ, Prentice-Hall, 101-124, 1980.

[137] ZDANOWICZ, M. : Automatyczne rozpoznawania odpowiedzi w badaniach psychofizjologicznych, Archiwum Akustyki, t. XVI, z. 3, 357-366, 1981.

[138] ZWICKER, E., TERHARDT, E., PAULUS, E. : Automatic Speech Recognition Using Psychoacoustic Models, JASA, 65, 487--498, 1979.

SPIS TREŚCI





Rys. 23. Schemat blokowy algorytmu wyznaczania wzorcowego spektrogramu binarnego metodą A1

Dla pierwszego fragmentu $Fr_1 \in SB_{v1}$ spektrogramu binarnego SB_{v1} pierwszej wypowiedzi v-tego wyrazu znajdź odpowiadający mu fragment $Fr_{\sim 1} \in SB_{v2}$ spektrogramu binarnego SB_{v2} drugiej wypowiedzi v-tego wyrazu.

Wyznacz pierwsze widmo binarne $WB_1 \in WZP_{v1}$ pierwszego wzorca pośredniego WZP_{v1} v-tego wyrazu sumując logicznie m-te widmo binarne $WB_m \in Fr_1 \in SB_{v1}$ oraz odpowiadające mu m-te widmo binarne $WB_m \in Fr_{\sim 1} \in SB_{v2}$:

$$(WB_1 \in WZP_{v1}) = (WB_m \in Fr_1 \in SB_{v1}) \cup (WB_m \in Fr_{\sim 1} \in SB_{v2})$$

i := 2

Dla kolejnego fragmentu $Fr_1 \in SB_{v1}$ spektrogramu binarnego SB_{v1} pierwszej wypowiedzi v-tego wyrazu znajdź odpowiadający mu fragment $Fr_{\sim i} \in SB_{v2}$ w spektrogramie binarnym SB_{v2} drugiej wypowiedzi v-tego wyrazu.

Wyznacz kolejne widmo $WB_i \in WZP_{v1}$ pierwszego wzorca pośredniego WZP_{v1} v-tego wyrazu według reguły:

$$(WB_i \in WZP_{v1}) = (WB_m \in Fr_i \in SB_{v1}) \cup (WB_m \in Fr_{\sim i} \in SB_{v2})$$

i = L / długość SB_{v1} /?

N

i := i+1

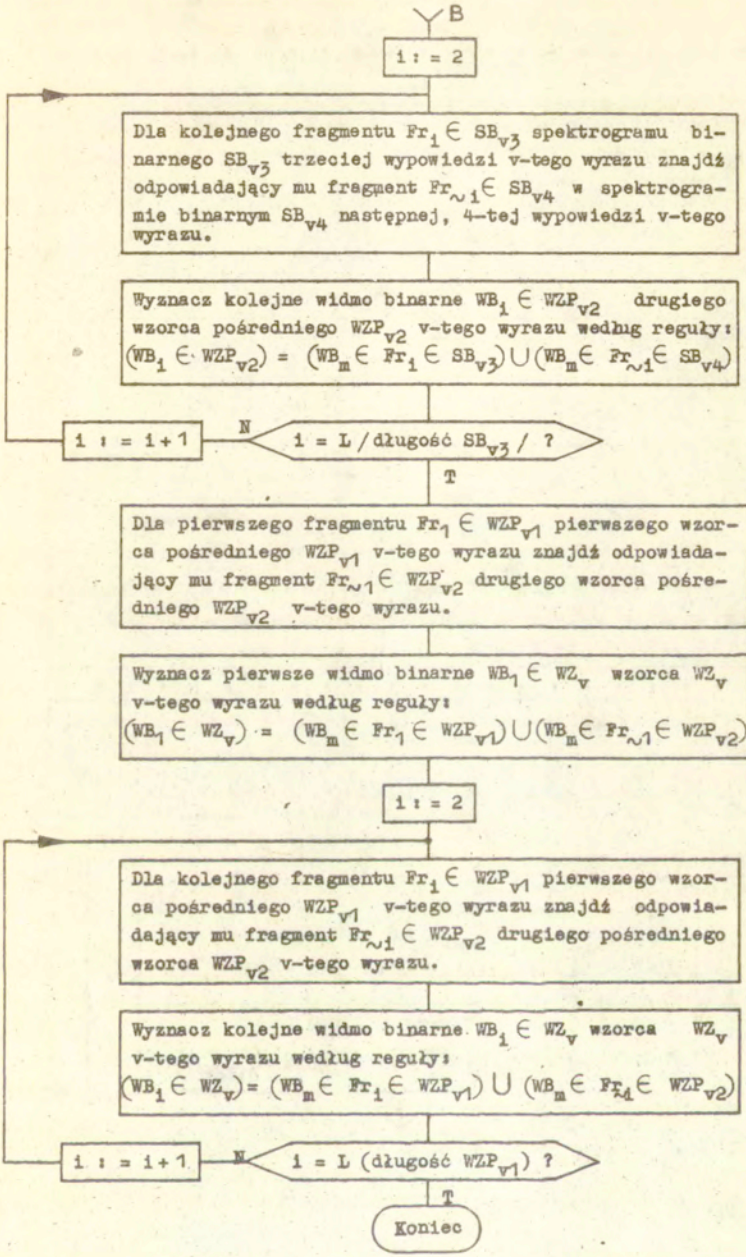
T

Dla pierwszego fragmentu $Fr_1 \in SB_{v3}$ spektrogramu binarnego SB_{v3} trzeciej wypowiedzi v-tego wyrazu znajdź odpowiadający mu fragment $Fr_{\sim 1} \in SB_{v4}$ w spektrogramie binarnym SB_{v4} następnej, 4-tej wypowiedzi v-tego wyrazu.

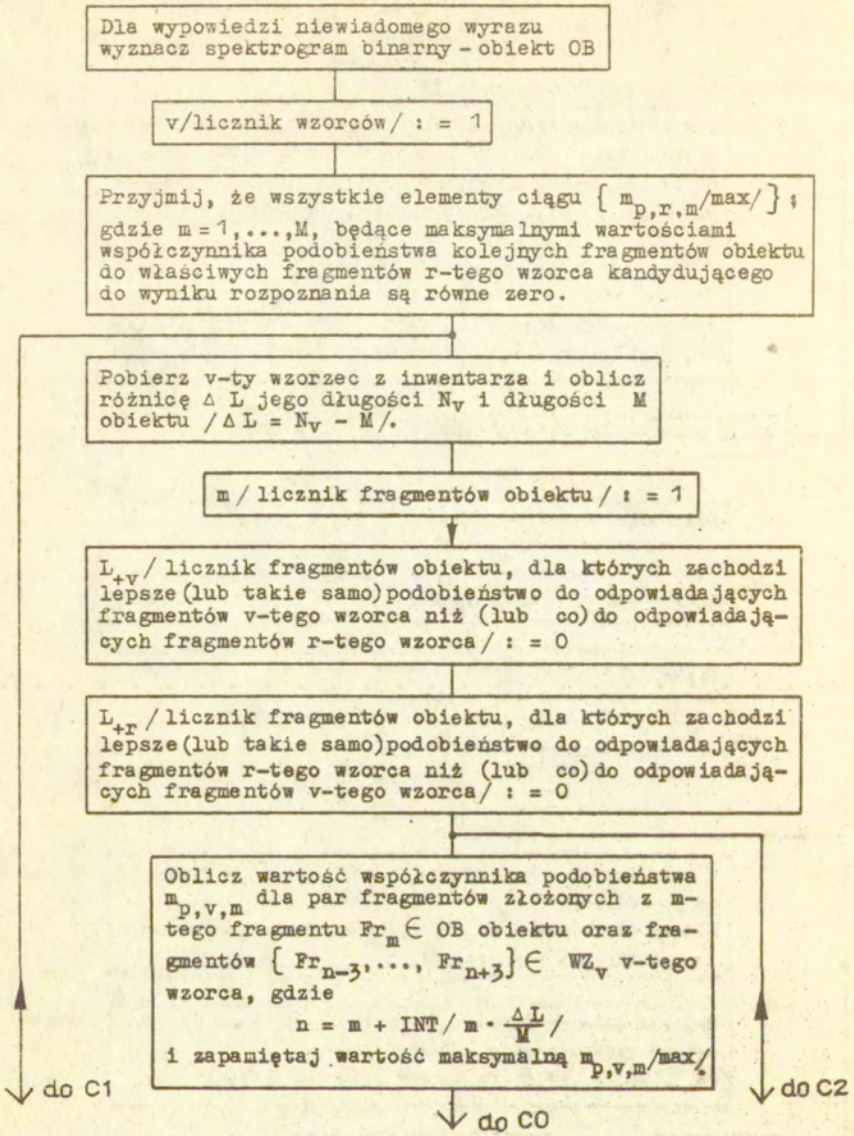
Wyznacz pierwsze widmo binarne $WB_1 \in WZP_{v2}$ drugiego wzorca pośredniego WZP_{v2} v-tego wyrazu według reguły:

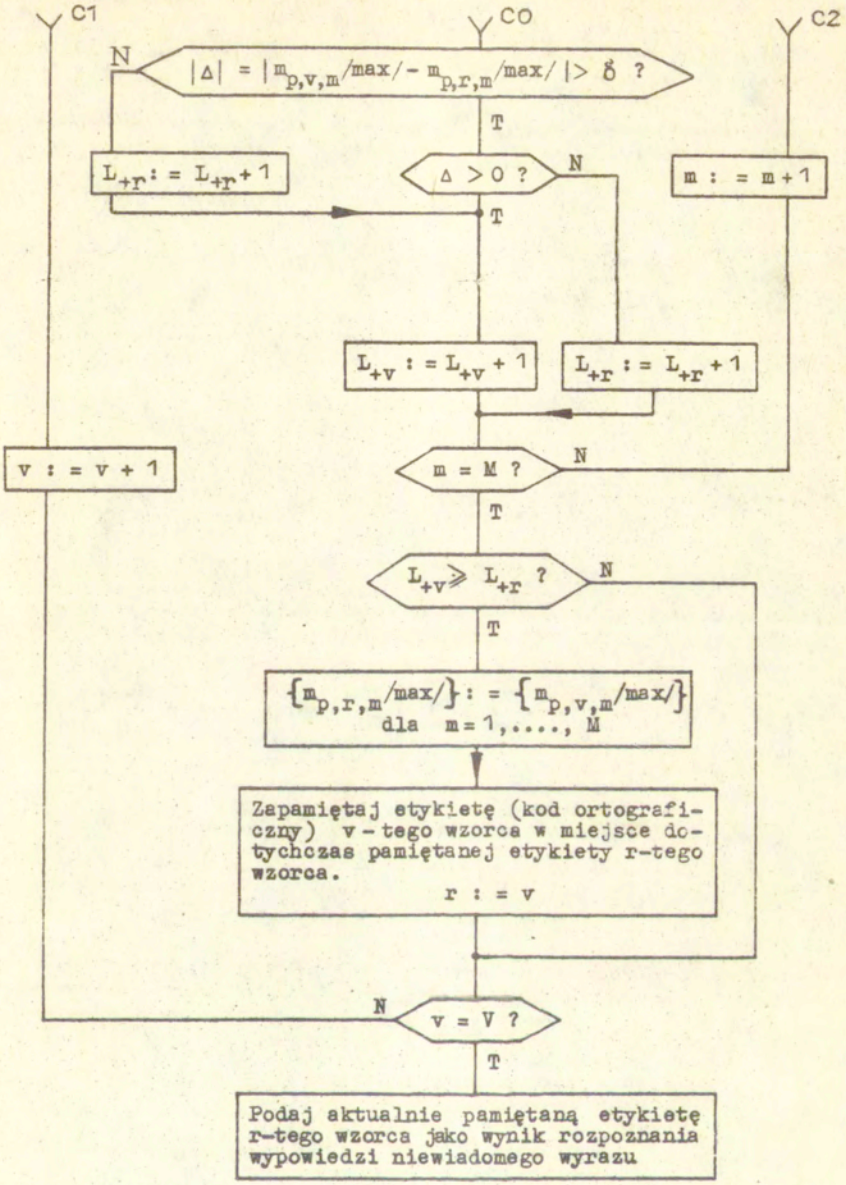
$$(WB_1 \in WZP_{v2}) = (WB_m \in Fr_1 \in SB_{v3}) \cup (WB_m \in Fr_{\sim 1} \in SB_{v4})$$

do B



Rys. 24. Schemat algorytmu wyznaczania wzorcowego spektrogramu binarnego wyrazu metodą A2





Rys. 29. Schemat algorytmu rozpoznawania wyrazu w modelu ROWBIR 1