Henryk Kubzdela

GLOBALNE ROZPOZNAWANIE ELEMENTÓW ZAMKNIĘTEGO ZBIORU HASEŁ

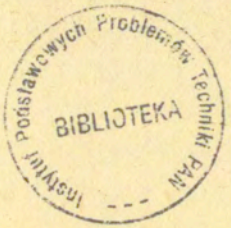
Przegląd problematyki

2/1985

WARSZAWA 1985

ISSN 0208-5658

Praca wpłynęła do Redakcji dnia 28 listopada 1984r.



Na prawach rękopisu

Instytut Podstawowych Problemów Techniki PAN

Nakład 140 egz. Ark.wyd. 2,1. Ark. druk. 3 .
Oddano do drukarni w styczniu 1985 r.
Nr zamówienia 19/85

Warszawska Drukarnia Naukowa, Warszawa,
ul.Śniadeckich 8

Henryk Kubzdela
Pracownia Fonetyki Akustycznej
IPPT PAN

GLOBALNE ROZPOZNAWANIE ELEMENTÓW ZAMKNIĘTEGO ZBIORU HASEŁ (PRZEGLĄD PROBLEMATYKI)[1]

Streszczenie

Praca stanowi przegląd technik stosowanych w spotykanych na świecie metodach globalnego rozpoznawania wyrazów. Metody te łączy wspólny tryb proceduralny, według którego przebiega proces rozpoznawania. Różnice natomiast występują w sposobie realizacji poszczególnych jego etapów. Po przedstawieniu krótkiej historii początków badań nad automatycznym rozpoznawaniem mowy omówiono wpierw różne czynniki, które badania te ograniczają i determinują i wskazano na powody skłaniające do rozpoznawania wyrazów w sposób globalny. W każdej metodzie rozpoznawania mowy występuje analiza akustyczna sygnału mowy, podczas której utworzony zostaje reprezentatywny obraz akustyczny rozpoznawanej wypowiedzi. Podano, jakie rodzaje analizy akustycznej stosowane są w systemach rozpoznawania mowy, nieco obszerniej przedstawiając metodę predykcji liniowej oraz analizę cepstralną. Zdefiniowano pojęcie obrazu akustycznego i przedstawiono stosowane metody porównywania dwóch obrazów akustycznych, osobno omawiając sposoby oceny podobieństwa lokalnego pomiędzy odnośnymi fragmentami porównywanych obrazów oraz podobieństwa globalnego całych obrazów. W kolejnym rozdziale dokonano przeglądu metod adaptacji systemu szczególnie dla przypadków tworzenia wzorców wspólnych dla pewnej liczby głosów. Na zakończenie podano in-

[1] Praca wykonana w ramach problemu międzyresortowego MR.I-24

formacje o wynikach rozpoznawania wyrazów uzyskanych różnymi metodami przez różnych badaczy oraz przytoczono przykłady konkretnych zastosowań automatycznego rozpoznawania wyrazów.

1. Wstęp

W ciągu ostatnich kilku lat opracowano w Pracowni Fonetyki Akustycznej IPPT PAN metodę globalnego rozpoznawania wyrazów na podstawie binarnych parametrów widmowych. Kolejne etapy związanych z tym badań są zrelacjonowane w pracach autora. Metoda ta powstawała w skromnych warunkach laboratoryjnych, co wpłynęło w znacznym stopniu na jej obecny kształt. Samo przedsięwzięcie stworzenia specyficznej metody rozpoznawania wyrazów było alternatywą wobec niemożności włączenia się w ustalony na świecie nurt badań nad automatycznym rozpoznawaniem mowy i to przede wszystkim ze względu na dzielący Pracownię Fonetyki Akustycznej od innych ośrodków badawczych zajmujących się podobną problematyką znaczny dystans w zakresie wyposażenia w środki badawcze. W zacytowanych powyżej pracach przedstawiono próby globalnego rozpoznawania wyrazów w oparciu o proste procedury obliczeniowe, nie wymagające komputera o dużej mocy obliczeniowej. Niniejsza praca porusza różne aspekty automatycznego rozpoznawania mowy, lecz przede wszystkim stanowi przegląd głównych metod globalnego rozpoznawania wyrazów stosowanych w świecie. Zawarte w tej pracy informacje stanowić mogą odniesienie dla oceny wartości metody autora o jakiej mowa wyżej. Z taką też intencją praca niniejsza została napisana.

2. Początki badań nad automatycznym rozpoznawaniem mowy

W społeczności ludzkiej mowa jest podstawowym środkiem komunikowania się. Za pomocą mowy ludzie przekazują sobie ustawicznie ogromne ilości informacji o najprzeróżniejszej wartości, wadze czy doniosłości. 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 kontakto-

waniu 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 wpierw 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, 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 identyfikującego 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 mówcy 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 oraz 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 (Dersch, 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 kom-

puterowej ożywiać zaczęła się idea stworzenia tak zwanego wejścia fonicznego umożliwiającego kontaktowanie się człowieka z komputerem za pomocą głosu i ewentualnie 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.

3. 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 szczyci się 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 (Vinczuk, 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 czasowym 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 mówcy, 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 mówcy 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 wcześniej określonych 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.

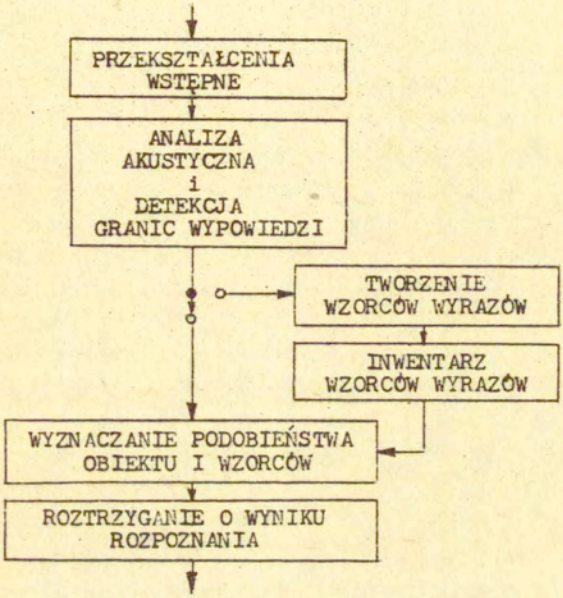
Fonem w mowie ciągłej należy zatem traktować jako zjawisko niestacjonarne z rozmytymi granicami. W takim też rozumieniu należałoby go klasyfikować.

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 przedstawił korzyści i niedogodności związane z posługiwaniem się w automatycznym rozpoznawaniu mowy różnego rodzaju jednostkami fonologicznymi, do których zaliczył alofony, fonemy, difony, sylaby i wyrazy. Z jego zestawienia wynika, że operowanie w automatycznym rozpoznawaniu mowy którąkolwiek z tych jednostek ma swoje wady i zalety. Schoup stwierdził też, że fonem identyfikuje się najtrudniej. Podobne oceny, lecz z nieco

innymi proporcjami wad i zalet przedstawił G. Mercier 1981 .

4. Uwagi ogólne o globalnym rozpoznawaniu mowy

Problemy jakie stwarza segmentacja mowy i klasyfikacja oraz 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 wyrazów 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.

5. 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 jeszcze 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 na przykład analizatory zapożyczone z praktyki wokoderowej, 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 1795 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)

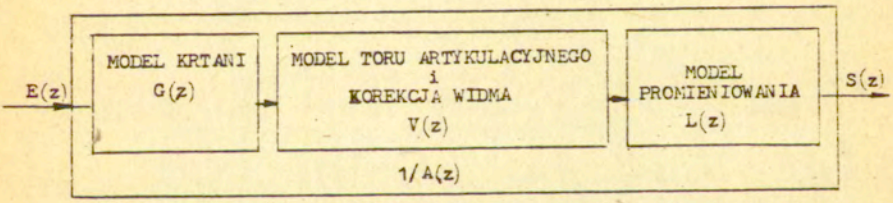
uważa się za pierwszych, którzy zastosowali metodę predykcji liniowej do analizy i syntezy mowy. Obszerne przedstawienie teorii i zastosowań predykcji liniowej w analizie, syntezie i rozpoznawaniu mowy zawierają prace Markela i Gray'a (1976) oraz Makhoula (1975 a i 1975 b).

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 :

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

posiada jedynie bieguny (nie ma zer).



Rys. 2. Model wytwarzania mowy przyjmowany w predykcji liniowej.

Model wytwarzania mowy wyrażony wzorem (1) reprezentuje w domenie czasowej zależność :

$$x_n = - \sum_{i=1}^{M} a_i x_{n-1} + e_n \qquad (2)$$

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 :

$$\sum_{i=1}^{M} a_i c_{ij} = -c_{oj} \qquad (3)$$

gdzie M jest rzędem filtru biegunowego, a

$$c_{ij} = \sum_{n=n_o}^{n_1} x_{n-i} \cdot x_{n-j} \qquad (4)$$

dla j = 1,2,...,M. n_o i n_1 oznaczają granice przedziału, na przestrzeni którego dokonuje się minimalizacja błędu.
Istnieją dwie metody obliczenia 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 wyróżniające 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 :

$$X(k) = \sum_{n=0}^{N-1} x(n) \exp\left(-j \frac{2\pi}{N} \cdot kn\right) \qquad (5)$$

$$\hat{X}(k) = \mathrm{Log}\left(X(k)\right) \qquad (6)$$

$$\hat{x}(n) = \frac{1}{N} \sum_{k=1}^{N-1} \hat{X}(k) \exp\left(j \frac{2\pi}{N} kn\right) \qquad (7)$$

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 korespondującej z kryteriami percepcyjnymi. Gagnoulet i inni (1983) potwierdzili słuszność propozycji Davisa i Mermelsteina przytaczając korzystne wyniki globalnego rozpoznawania trudnych wyrazów francuskich z zastosowaniem reprezentacji cepstralnej i MEL-owskiej 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-owską skalę częstotliwości można wystarczająco reprezentatywnie dla potrzeb rozpoznawania przedstawić sygnał mowy za pomocą zaledwie 6 do 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 domenie czasowej stosowana jest w rozpoznawaniu mowy bardzo 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.

6. Pojęcie obrazu akustycznego

Wynikiem analizy akustycznej wypowiedzi w systemach globalnego rozpoznawania wyrazów jest zatem macierz wartości parametrów reprezentujących sygnał mowy w kolejnych przedziałach czasu. Nazywa się tę macierz w terminologii angielskiej przeważnie słowem 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".
W niektórych pracach angielskojęzycznych np. Rabinera (1978) spotyka się określenie TEMPLATE zamiast PATTERN. Słowo TEMPLATE nie ma w ogóle polskiego odpowiednika leksykalnego. W słownictwie polskim brakuje także odpowiednika wyrazu FRAME oznaczającego między innymi to, co powyżej określono przez PROBA. Słowo PROBA 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.

6.1. Porównywanie obrazów akustycznych

Po parametryzacji sygnału mowy i wyznaczeniu początku i 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 lub odległości odpowiadających sobie segmentów porównywanych obrazów. 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.

6.2. 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 globalnego 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 użyli jako miarę odległości porównywanych segmentów moduł różnicy wektorów cech A_x i B_y reprezentujących te segmenty :

$$d(x,y) = \|A_x - B_y\| \qquad (8)$$

Das przyjął jako miarę odległości sumę bezwzględnych różnic współrzędnych wektorów cech

$$d(x,y) = |A_x - B_y| = \sum_{i=1}^{n} |a_{xi} - b_{yi}| \qquad (9)$$

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. Wyraża się ona wzorem :

$$d(x,y) = \log \frac{\sum_{i=1}^{n} (a_{xi} b_{yi})}{\left(\sum_{i=1}^{n} a_{xi}^2\right)^{\frac{1}{2}} \left(\sum_{i=1}^{n} b_{yi}^2\right)^{\frac{1}{2}}} \qquad (10)$$

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 mowy 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 $\hat{a}$ 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 :

$$d(x/a) = \log\left(\frac{aVa^t}{\hat{a}V\hat{a}^t}\right) \qquad (11)$$

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

$$d(x/a) = c + \log\left[(br)/(\hat{a}r)\right] \qquad (12)$$

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 :

$$b(i) = 2\sum_{j=0}^{p-1} a(j)a(j+1)/(aa) \qquad (13)$$

r jest znormalizowanym wektorem korelacji r = (v(i)/v(o)), (i = o,...,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.

6.2. Wyznaczanie odległości globalnej porównywanych obrazów akustycznych

Oddzielny problem w porównywaniu obrazów akustycznych przedstawia 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 inny wymiar czasowy. 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. Vinczuka (1969), Veliczkę 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 obrazów akustycznych A i B wyznacza się odległość globalną między nimi stosując następujący wzór :

$$D(A,B) = \min_{F} \left[\frac{\sum_{k=1}^{K} d(c(k)) \cdot w(k))}{\sum_{k=1}^{K} w(k)} \right] \qquad (14)$$

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 funkcja 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. Ostatnie zdanie wyrażane jest następującym zapisem : $c(k) = (i(k), j(k))$. w(k)we wzorze (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 (1978):

1. Warunek monotoniczności, który wymaga, aby

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

2. Warunek ciągłości wyrażony nierównościami :

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

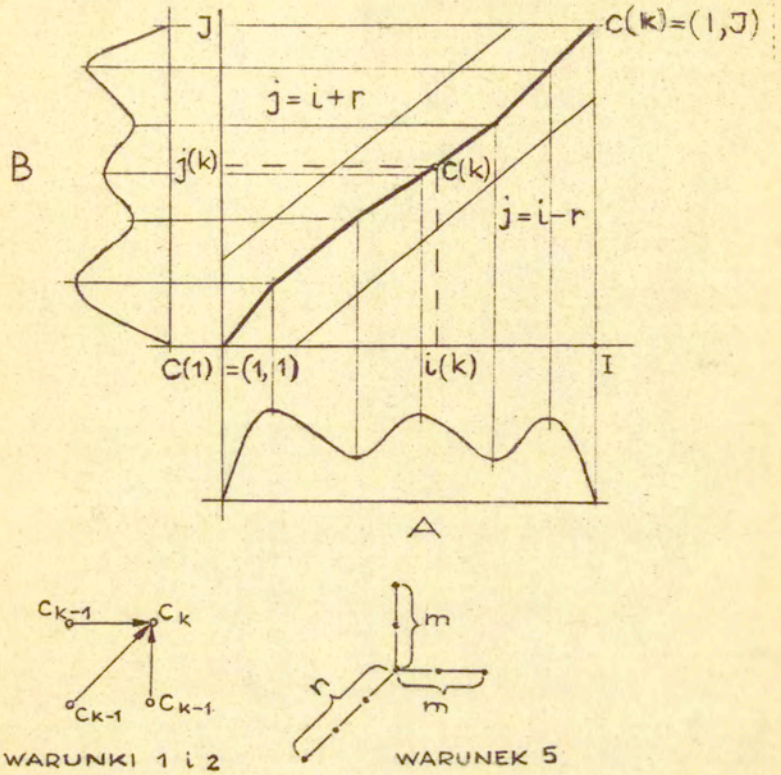
Z warunków 1 i 2 wynika, iż punkt c(k-1) bezpośrednio poprzedzający punkt c(k) odnosić się może do jednej z trzech par fragmentów porównywanych wyrazów, co zapisać można następująco wyrażając te fragmenty ich indeksami :

$$c(k-1) = \left[(i(k), j(k)-1) \text{ v } (i(k)-1, j(k)-1) \text{ v } (i(k)-1, j(k))\right].$$

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

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

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



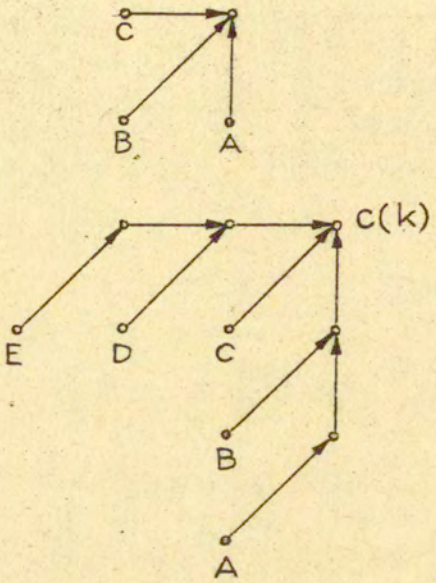
Rys. 3. Ilustracja zasady dynamicznej nieliniowej normalizacji czasowej.

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

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

5. Ostatni piąty warunek określa dopuszczalne nachylenie funkcji NC określone 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 (1975), 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.



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 :

$$g_k(c(k)) = \min_{c(k-1)} \left[g_{k-1}(c(k-1)) + d(c(k) \cdot w(k) \right] , \qquad (15)$$

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. $w(k) = (i(k) - i(k-1) + (j(k) - j(k-1)$ (16)

dla tzw. formy symetrycznej,

2. $w(k) = (i(k) - i(k-1))$ (17)

dla formy niesymetrycznej.

Mianownik we wzorze (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 :

$$g(i,j) = \min \begin{bmatrix} g(i,\ j-1) + d(i,j) \\ g(i-1,\ j-1) + 2d(i,j) \\ g(i-1,\ j) + d(i,j) \end{bmatrix} \qquad (18)$$

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 :

$$g(i,j) = \min \begin{bmatrix} g(i-1,j-3) + 2d(i,j-2) + d(i,j-1) + d(i,j) \\ g(i-1,j-2) + 2d(i,j-1) + d(i,j) \\ g(i-1,j-1) + 2d(i,j) \\ g(i-2,j-1) + 2d(i-1,j) + d(i,j) \\ g(i-3,j-1) + 2d(i-2,j) + d(i-1,j) + d(i,j) \end{bmatrix} \qquad (19)$$

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.

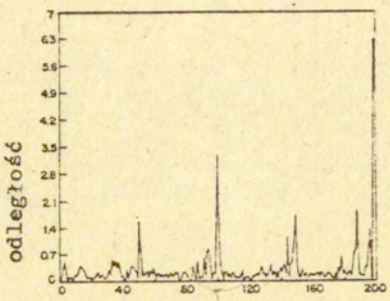
7. Metody uczenia adaptacji .

Rozpoznanie wypowiedzianego wyrazu następuje w wyniku porównania jego obrazu akustycznego z wzorcowymi obrazami akustycznymi wyrazów do rozpoznawania 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 pojedyńczo 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 indywidaulnych 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. Wymiar czasowy przyszłego wzorca wyrazu 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 pojedyńczego wyrazu utworzonych może być kilka wzorców na podstawie wypowiedzi wielu głosów. Levinson i inni (1979) podają 4 procedury klasteryzacji. Pierwsza, zastosowana przez Patricka (1972) nazwana 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 innych.

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

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, iż 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



Rys. 5. Wykres odległości sąsiednich obrazów w szeregu łańcuchowym.

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 :

$$x_j \in \omega_i \text{ , jeżeli } \delta(x_j, x_p^{(i)}) \leq \delta(x_j, x_p^{(k)}), \quad 1 \leq k \leq M \qquad (20)$$

Obraz x_j zaliczony zostaje do klasy ω_i, jeżeli jego odległość do najbliższego obrazu $x_p^{(i)}$ tej klasy jest najmniejszą ze 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, $\min(\delta(x_j^{(i)}, x_k^{(i)}))$. 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 każdego 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 segmentó.. 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.

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 rozpoznawania wyniosła 88.6 %. Itakura zacytował prace Velicki 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 danych 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 każdy z 39 wyrazów słownika. Głosy te nie brały udziału w etapie uczenia. 86 % - ową 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 zaproponо-

wali prostą metodę treningu dla systemu rozpoznawania izolowanych wyrazów, posługując się tym samym co poprzednio słownikiem uzyskali poprawność rozpoznawania w granicach 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, bo liczącym aż ponad 1000 wyrazów słowniku, z udziałem 5 głosów męskich z indywidualnymi zbiorami wzorców uzyskując 95.8 %-ową 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. Tradycyjne dynamiczna normalizacja czasowa zastąpiona została porównywaniem ciągów uprzednio 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 zależności 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 jego 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.

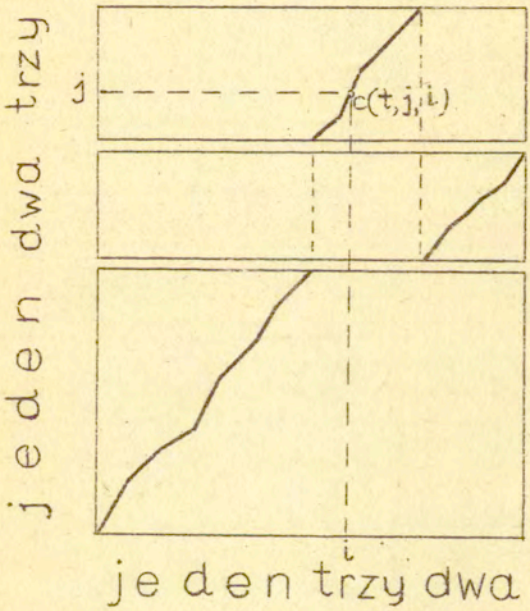
Na zakończenie tego przeglądu wybranych przykładów globalnego rozpoznawania wyrazów wspomnieć warto o wynikach, jakie osiągnęli Baudry i Dupeyrat (1982) stosując metodę opartą o parametry czasowe. Metodę tę scharakteryzowano już pokrótce w niniejszej pracy w rozdziale poświęconym omówieniu metod parametryzacji sygnału mowy w systemach rozpoznawania mowy. W próbie testowej uzyskano 98-procentową poprawność rozpoznawania dla 10-krotnej wypowiedzi każdego z 32 słów 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 mówcy, dla którego utworzony został zbiór wzorców.

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

$$N = k! + \sum_{n=1}^{k-1} \binom{k}{n} (k-n)! \qquad (21)$$

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 Winczuka (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. Mowę jak wiadomo interpretować można jako ciąg różnego rodzaju elementów segmentalnych. Ciąg wyrazów stanowi jedną z kilku przyjmowanych interpretacji. Ponieważ segmentacja przedstawia jeden z zasadniczych problemów automatycznego rozpoznawania mowy pożądane są takie metody rozpoznawania elementów segmentalnych w mowie, które nie wymagają znajomości położenia granic segmentalnych. Metoda przedstawiona powyżej wydaje się być wartą rozpatrzenia propozycją rozwiązania problemu rozpoznawania ciągów elementów segmentalnych innych niż wyrazy, np. difonów.

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 i 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 po-

mocnik. 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 uciążliwość. 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). Wprowadzanie 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 głoso-wejścia. Dla przykładu kilka z nich :

Kontrola jakości ściany przedniej kineskopu telewizji kolorowej składa się z 54 niezależnych pomiarów. Kiedyś wymagała taka kontrola 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 głosowejścia w automatycznych sortowniach już w roku 1973.

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 głosem podaje odnośne dane o głębokości. W podobnej roli znalazło fono-wejście 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.

Ta stosunkowo młoda dziedzina, jaką jest automatyczne rozpoznawanie mowy rozwija się w ostatnim czasie niezwykle dynamicznie. Żywić należy nadzieję, iż znajdzie ona także właściwe uznanie w Polsce poprzez stworzenie jej odpowiednich warunków rozwoju.

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

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

6 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 .

7 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 .

8 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 .

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

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

11 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 .

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

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

14 DREYFUS-GRAF, J. : Sonagraph and Sound Mechanics, JASA, Vol. 22, 731-739, 1949 .

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

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

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

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

19 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 .

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

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

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

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

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

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

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

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

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

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

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

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

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

33 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 .

34 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 .

35 KUBZDELA, H. : Automatyczne rozpoznawanie wyrazów na podstawie spektrogramów binarnych, Prace IPPT 15/1981, Warszawa, 1981 .

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

37 KUBZDELA, H. : Badania nad udoskonaleniem spektrogramów binarnych, Prace IPPT 24/1983, Warszawa, 1983 .

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

39 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 .

40 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 .

41 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 .

42 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 .

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

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

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

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

47 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 .

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

49 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 .

50 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 .

51 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 .

52 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 .

53 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 .

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

55 OPPENHEIM, A.V., SCHAFER, R.W. : Homomorfic Analysis of Speech, IEEE Trans. on Audio and Electroacoustics, AU-16, No. 2, 27-31, 1968 .

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

57 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 .

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

59 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 .

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

61 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 .

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

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

64 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 .

65 ROSS, P.W. : A Limited-Vocabulary Adaptive Speech Recognition System, Journal of the Audio Engineering Society, Vol. 15, 414-418, 1967 .

66 RUSKE, G., SCHOTOLA, T. : The Efficiency of Demisyllable Segmentation in the Recognition of Spoken Words, Automatic Speech Analysis and Recognition, Ed. : Haton J.-P., Dordrecht, Holland, 153-163, 1982 .

67 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 j. japońskim .

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

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

70 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 .

71 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 .

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

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

74 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 .

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

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

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

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

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

80 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 .

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