Henryk Kubzdela

SYSTEM DO BADANIA CECH WIDMOWYCH ORAZ SELEKCJI I ODSŁUCHU SEGMENTÓW SYGNAŁU MOWY

18/1993

WARSZAWA 1993

ISSN 0208-5658

Praca wpłynęła do Redakcji dnia 15 grudnia 1992 r.

Na prawach rękopisu

Instytut Podstawowych Problemów Techniki PAN
Nakład 100 egz. Ark.wyd.1,25 Ark.druk.1,60
Oddano do drukarni w maju 1993 r.

Wydawnictwo Spółdzielcze sp. z o.o.
Warszawa, ul.Jasna 1

Henryk Kubzdela
Zakład Fonetyki Akustycznej
IPPT PAN

SYSTEM DO BADANIA CECH WIDMOWYCH ORAZ SELEKCJI I ODSŁUCHU SEGMENTÓW SYGNAŁU MOWY

S t r e s z c z e n i e

Zjawisko rozległej wariantywności akustycznej głosek w mowie przedstawia główną trudność w znalezieniu właściwych zasad ich automatycznej identyfikacji. Logicznie spójny i zamknięty system fonetyczny mowy nie znajduje prostego odpowiednika w dziedzinie akustycznej. Sygnał mowy i zasady jego percepcji nie są w wystarczającym stopniu poznane. Stworzenie modelu automatycznej identyfikacji dźwięków mowy przy uwzględnieniu ich szerokiej wariantywności wymaga narzędzi szybkiej analizy akustycznej, selekcji dowolnych segmentów mowy oraz prostego i kombinowanego odtwarzania w celach odsłuchowych wyizolowanych segmentów. Przedstawiono system, który wykonuje takie operacje i na przykładach zilustrowano jego możliwości. Szerzej omówiono funkcję selekcji i odtwarzania prostego lub kombinowanego wybranych segmentów. Opisano eksperyment percepcji cyklicznie powtarzanych tzw. mikrosegmentów izolowanych oraz segmentów dłuższych osobnych a także kombinowanych z ich cyklicznie powtarzanymi skrajnymi mikrosegmentami. Przedstawiono i omówiono uzyskane wyniki zwracając uwagę na niezgodności ocen percepcyjnych wybranych segmentów oraz pobranych z nich cyklicznie powtarzanych mikrosegmentów.

1. Wstęp.

Elementem fonetycznym mowy jest głoska. Z punktu widzenia nauki o języku każda fraza może być przedstawiona jako ciąg głosek. W pisowni danego języka poszczególnym głoskom przypisuje się znaki zgodnie z obowiązującymi w tym języku zasadami ortografii. Niektóre z symboli ortograficznych używane są do oznaczania dwóch odmiennych głosek. Np. symbol *d* w wyrazie *kod* wymówionym w izolacji reprezentuje inny dźwięk mowy niż w wyrazie *kody* . Niektóre z głosek oznaczane są dwoma symbolami ortograficznymi. Np. zapis *nie* reprezentuje dwie głoski chociaż składa się z symboli ortograficznych trzech głosek. Pisownia

ortograficzna nie opisuje dźwięków mowy w sposób jednoznaczny. Fonetyka dostarcza pisowni, która wolna jest od wielu niedoskonałości pisowni ortograficznej. Pisownia fonetyczna nie uwzględnia jednak odmian kontekstowych podstawowych głosek. Zakłada ona bowiem, że dana głoska pozostaje w swojej klasie bez względu na sąsiedztwo innych głosek. Ta zasada wynika z jednakowego sposobu artykulacji przynajmniej w centralnej fazie powstawania wszystkich dźwięków mowy oznaczonych tym samym samym znakiem fonetycznym. Komputerowa konwersja zapisu ortograficznego na fonetyczny lub odwrotna nie przedstawia trudności. Komputerowy konwerter ortograficzno - fonetyczny dla j. polskiego został opracowany przez I. Nowaka dla systemu TEXT to SPEECH na podstawie reguł podanych przez M. Steffen-Batogową.

Badacze zajmujący się problemem automatycznego rozpoznawania mowy ciągłej przyjmują słuszne założenie, że wynikiem rozpoznania na najniższym poziomie powinien być ciąg etykiet głosek składających się na rozpoznawaną frazę. Podstawę automatycznego rozpoznawania mowy stanowi akustyczny sygnał mowy. Jedną z głównych trudności, jaką sygnał ten przedstawia jest brak w nim akustycznej jednorodności w segmentach poszczególnych głosek. Segmenty mowy, które w zapisie fonetycznym są oznaczane tym samym znakiem, mają w akustycznym sygnale mowy rozmaite obrazy widmowe o niewyraźnych granicach czasowych. Ten stan rzeczy powodują różne czynniki a dominującymi wśród nich są : indywidualne cechy głosowe osoby mówiącej i kontekst artykulacyjny, czyli pozycja startowa i docelowa narządów artykulacyjnych w procesie wymawiania segmentu utożsamianego z głoską. Parametry widmowe sygnału mowy w obrębie takiego segmentu nie mają wartości stałej lecz są często nieciągłymi funkcjami czasu. Stopień identyczności odpowiadających sobie parametrów w segmentach zaliczanych do tej samej głoski jest zróżnicowany i wykazuje niekorzystny rozrzut. Z tych powodów nie udało się dotychczas wskazać niezawodnych cech akustycznych, które jednoznacznie charakteryzowałyby

poszczególne głoski języka polskiego dla szerokiej populacji głosów i były wspólne dla danej głoski bez względu na jej różne konteksty. Dla akustycznego sygnału mowy trudno jest sformułować proste warunki tożsamości segmentów czasowych mowy uważanych z fonetycznego punktu widzenia za identyczne. Poszukiwanie takich warunków pozostaje podstawowym zadaniem dla badaczy pracujących nad problemem automatycznego rozpoznawania mowy ciągłej. Pomocą w tym powinna być znajomość roli poszczególnych podsegmentów głoskowych w percepcji całej głoski. Uzyskanie tej wiedzy jest obecnie możliwe dzięki technice komputerowej. Pozwala ona na szeroką i wielostronną analizę akustyczną sygnału mowy połączoną z oceną postrzegalności przez człowieka wyselekcjonowanych z niego segmentów. Komputer mający takiej roli sprostać musi być wyposażony w karty wejścia i wyjścia akustycznego oraz programy analizy akustycznej, grafiki jej wyników oraz selekcji, powielania i łączenia odpowiednich segmentów czasowych sygnału mowy. Programy komercyjne służące przetwarzaniu sygnału mowy spełniają jedynie fragmentarycznie wymienione wymagania. Przewidziane są one dla szerokiego grona zainteresowanych mową i z tego powodu nie zaspokajają wszystkich indywidualnych i nietypowych potrzeb badawczych. Specyficzne zadania badawcze wymagają stworzenia we własnym zakresie oryginalnych narzędzi. Poniżej przedstawiony zostanie system powstały do badania cech rozpoznawczych segmentów fonetycznych mowy. Celem tych badań będzie wskazanie cech identyfikacyjnych segmentów mowy fonetycznie tożsamych lecz akustycznie zróżnicowanych z powodu odmiennego kontekstu i osobliwości głosowych mówiącego.

2. System WAWPSO

Prezentowany system służy do wpisu, analizy akustycznej, wybierania oraz powielania segmentów czasowych i odsłuchu sygnału mowy. Nazwę jego tworzą początkowe litery wyrazów określających jego podstawowe funkcje. Bazę

sprzętową systemu stanowią komputer PC-XT/386, oraz karty z konwerterami a/c i c/a. Na część softwar'ową systemu składają się procedura FFT oraz programy, których napisanie w językach C oraz turbo assembler stanowiło przedmiot relacjonowanego poniżej zadania badawczego. Ultraszybka procedura FFT oparta o algorytm Winograda jest autorstwa J. Ogórkiewicza z Politechniki Poznańskiej i została udostępniona przez samego autora. Także na Politechnice Poznańskiej powstała karta z przetwornikiem a/c. Poniżej opisane zostaną funkcje, jakie wykonuje system.

2.1. Funkcja kontrolowanego wpisu sygnału mowy do pamięci operacyjnej komputera.

Program tej funkcji napisano w języku turbo assembler. Było to konieczne ze względu na wysoką założoną częstotliwość próbkowania wynoszącą 16 kHz oraz zastosowanie kontroli granic wypowiedzi. Przyjęto następujące kryteria identyfikacji granic wypowiedzi :

Warunek początku wypowiedzi : Wystąpienie co najmniej trzech kolejnych fram niezerowych (pierwsza z nich uznana zostaje za początek wypowiedzi). Liczba ewentualnie po nich następujących fram zerowych nie może przekraczać 10.

Frama składa się z 16 próbek. Określa się ją jako niezerową, gdy zawiera przynajmniej jedną próbkę niezerową p_{nz}, czyli o wartości wykraczającej poza założony przedział : $(-p_r, +p_r)$, rozciągający się symetrycznie po obu stronach zera. Ze względu na poziom zakłóceń sygnału mowy p_r nie może być mniejsze od 2^5.

Warunek końca wypowiedzi : Wystąpienie bezpośrednio po framie niezerowej, 20-tu kolejnych fram zerowych.

Frama zerowa nie zawiera ani jednej próbki niezerowej. Maksymalna wartość bezwzględna próbki ma rozmiar liczby 11-bitowej. Maksymalną długość sygnału, jaki jednorazowo może przyjąć komputer ustalono na 2^{15}. Odpowiada to frazie długości 2.048 s. Wpis sygnału zostaje przerwany z chwilą

wystąpienia przesterowania lub przepełnienia zarezerwowanego na sygnał bufora pamięci operacyjnej. W pierwszym przypadku wysłane zostaje polecenie powtórzenia wypowiedzi. Poprawne zakończenie wpisu kończy się informacją o długości przyjętego sygnału.

2.2. Założenie i odczyt pliku wypowiedzi.

Wpisany do pamięci operacyjnej sygnał mowy może zostać zapamiętany na dysku jako zbiór oznaczony nazwą z rozszerzeniem wskazującym, iż zbiór jest zdigitalizowanym sygnałem mowy. Zbiór ten w razie potrzeby może zostać przesłany z dysku ponownie do pamięci operacyjnej w celu dokonania przekształceń, jakie umożliwia prezentowany system. Pobranie zbioru z dysku poprzedza wyświetlenie na ekranie monitora wykazu uporządkowanych alfabetycznie nazw zbiorów posiadających identyczne rozszerzenie, oznaczające np., że zbiory zawierają zdigitalizowany sygnał mowy z wypowiedzi różnych fraz. Nazwę zbioru, który ma być pobrany, należy wskazać kursorem przesuwanym we wzajemnie prostopadłych kierunkach. Obraz wykazu jest zwijalny, co pozwala zapamiętać pod wspólnym rozszerzeniem nazwy więcej zbiorów, niż liczba nazw, jaką jednorazowo może pomieścić okno ekranu. Tę dogodność zawdzięcza się procedurze, która może być dołączona do każdego napisanego w języku C programu, w którym zachodzi pobranie zbioru z dysku. Autorem tej procedury jest B. Szutowski.

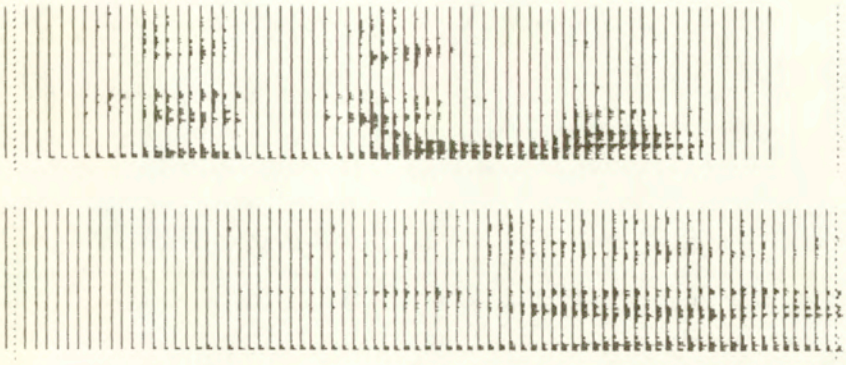
2.3. Odsłuch wypowiedzi.

Sygnał mowy umieszczony w pamięci operacyjnej wprost z przetwornika a/c lub wtórnie z dysku może być odsłuchany w całości synchronicznie z częstotliwością wcześniej dokonanej konwersji a/c. Poniżej opisana zostanie też możliwość odsłuchu w różnych wariantach wyselekcjonowanych z sygnału segmentów.

2.4. Analiza sygnału mowy.

Tym hasłem oznaczono szereg funkcji systemu, które przedstawione zostaną poniżej. Funkcje te odwołują się do widma dynamicznego wyznaczonego przez wyżej wspomnianą ultraszybką procedurę FFT. 120 - punktowe widmo chwilowe wyznaczone zostaje dla framy obejmującej tym razem 240 kolejnych próbek sygnału wydzielonych oknem Hamminga. Krok analizy spektrograficznej sygnału może być wybrany zależnie od potrzeb od wartości minimalnej równej okresowi próbkowania aż do długości jednej framy. W pierwszym etapie z krokiem długości jednej framy wykonana zostaje globalna analiza wypowiedzi. Jej wynikiem jest spektrogram składający się z ciągu widm chwilowych. Wartości amplitudowe widma w poszczególnych jego punktach wyrażone są na obrazie spektrograficznym 8-stopniową grubością kreski reprezentującej jedno widmo. Spektrogram zajmuje 1/3 szerokości ekranu i mieści się w jego górnej części. W tym obszarze można przesuwać spektrogram w kierunku poziomym przyporządkowując lewemu brzegowi ekranu dowolną framę sygnału. Na obraz spektrograficzny nałożone zostają dwa kolorowe kursory mające formę odcinków nieco wykraczających poza poziome brzegi spektrogramu. Zielonym kursorem można wskazać framę, od której ma być ukazany przesunięty spektrogram. Przesunięcie spektrogramu wywołuje się klawiszem *k* a jego powrót do pozycji początkowej klawiszem *p*. Wspomniane kursory służą jeszcze innym celom, które omówione zostaną poniżej.

Możliwe jest uzyskanie spektrogramu wydłużonego wielokrotnie dzięki skróceniu kroku analizy widmowej. Krotność wydłużenia spektrogramu jest proporcjonalna do względnej długości zakresu pokrywania się sąsiednich fram. Teoretycznie maksymalne zwielokrotnienie długości równe jest długości framy. W praktyce sensowne jest jedynie kilkakrotne wydłużenie. Na rys.1 zamieszczono normalny spektrogram całej wypowiedzi wyrazu *jemioła* oraz wydłużony spektrogram jej początku.



Rys.1. Spektrogram normalny całej wypowiedzi wyrazu 'jemioła' oraz wydłużony spektrogram jej początku.

2.4.1. Obrazy wybranego widma chwilowego.

Przy pomocy klawisza w wywołuje się szczegółowy obraz widma chwilowego widocznego w uproszczeniu na spektrogramie w miejscu wskazanym przez zielony kursor oraz wersję wygładzoną tegoż widma. Oba obrazy ukazują się poniżej spektrogramu wypowiedzi. Przykład takiej prezentacji zamieszczono na rys.2. Zastosowano zasadę wygładzania widma podaną między innymi w pracy [1]. Na obrazie widma wygładzonego kropkami ponad liniami widmowymi zaznaczone są punkty widma, w których obwiednia spełnia kryterium wypukłości. Na obraz ten nałożony jest też wykres funkcji określanej jako waga widma. Jej przebieg zależy od energii akustycznej sygnału w wybranych pasmach częstotliwości o szerokości zależnej od usytuowania pasma. Charakter zależności tej funkcji od energii sygnału dobiera się pod kątem uwypuklenia cech widmowych różnicujących fonetyczno - akustyczne segmenty mowy. Funkcja wagi widma zawiera też składową stałą proporcjonalną do poziomu najsilniejszej składowej. Niektóre wersje funkcji wagi widma zdefiniowano w

pracach [1],[2]. Poszukuje się nowych jej odmian. Prezentowany system ma służyć między innymi temu celowi.

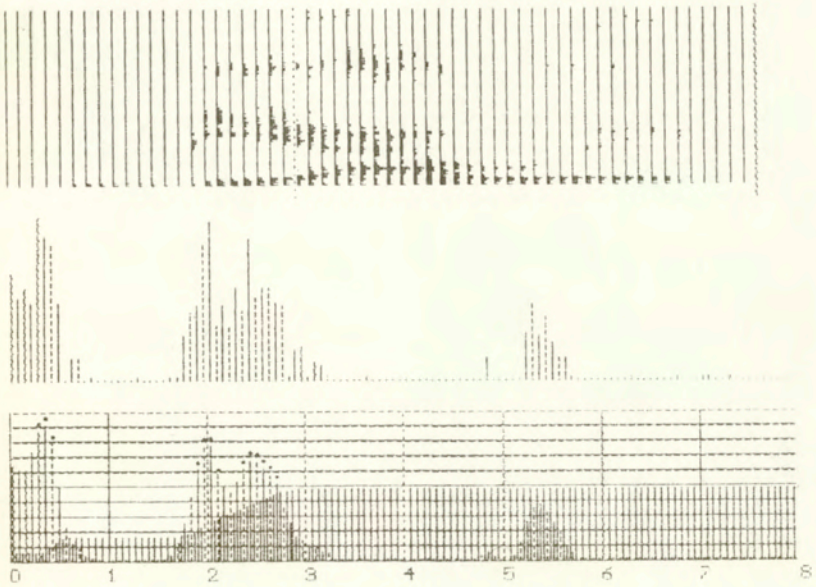
2.4.2. Obraz wyników pośrednich w wyznaczaniu i klasyfikacji segmentów wypukłych widma.

Wyniki wcześniejszych badań [3] wskazują, że rozkład segmentów widma o wypukłej obwiedni może służyć do identyfikacji segmentów mowy. Poszczególne widma sygnału mowy różnią się między innymi pod względem liczby, położenia i szerokości tych segmentów. Dla wykorzystania tej właściwości do identyfikacji segmentów fonetycznych w sygnale mowy konieczna jest dokładna znajomość tych różnic. Niektóre z segmentów widma o wypukłej obwiedni są ważniejsze od pozostałych i im automatycznie należy przypisać wyższą rangę. Umożliwić to ma między innymi odpowiednio dobrana funkcja wagi widma. Do weryfikacji różnych modeli tej funkcji służyć ma posiadana przez system możliwość prezentacji wyników pośrednich uzyskiwanych w trakcie wyznaczania rangi segmentów widma o wypukłej obwiedni. Możliwość tę ilustruje przykład zamieszczony na rys.3. Poszczególne fragmenty tej ilustracji w kierunku na prawo przedstawiają kolejno : widmo chwilowe oryginalne - identyczne z zamieszczonym wyżej w układzie poziomym, to samo widmo po wygładzeniu, segmenty o wypukłej obwiedni, segmenty wypukłe po skorygowaniu widma funkcją wagi, rozkład segmentów wypukłych uwzględniający ich zróżnicowaną rangę wynikającą z użycia określonego rodzaju funkcji ważącej składowe widma.

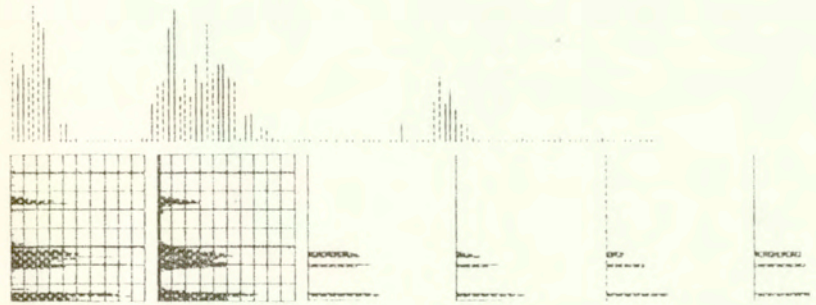
2.4.3. Prezentacja zależności wyniku analizy widmowej od usytuowania okna.

Nieregularności występujące niekiedy na obrazach widmowych mogą być wynikiem niedoskonałości analizy dyskretnej. Taką niedoskonałością jest między innymi zależność wyniku analizy widmowej od usytuowania okna względem okresu podstawowego sygnału. Wykrycie i ocenę

rozmiarów tej przyczyny umożliwiają : prezentacja całego widma dla wybranego usytuowania okna oraz przebieg wartości widma w dowolnym jego punkcie w funkcji pozycji okna.

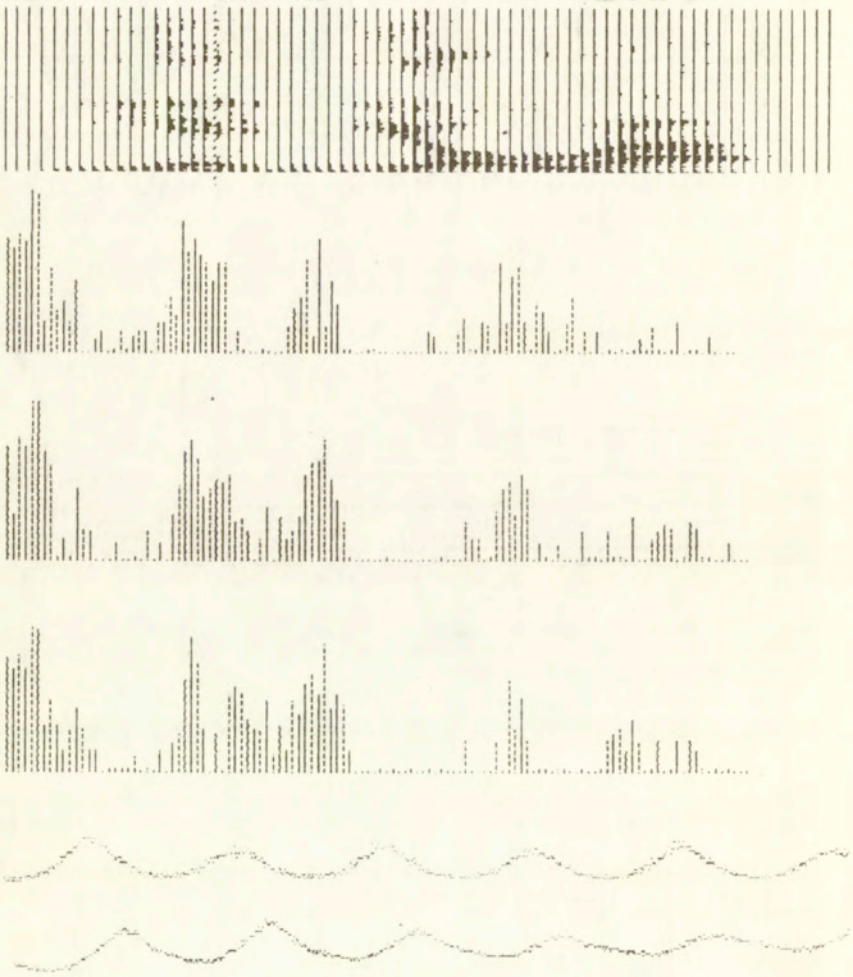


Rys.2. Obrazy widma oryginalnego i wygładzonego z zaznaczonymi przedziałami wypukłości obwiedni oraz nałożonym wykresem funkcji wagi widma.



Rys.3. Spektrogram, wybrane widmo chwilowe oraz ilustracja wyników pośrednich przy wyznaczaniu rangi segmentów wypukłych tego widma.

Na rys.4 przedstawiono 3 widma wyliczone dla 3-ech pozycji okna odległych od siebie odpowiednio o 1/4 i 1/2 długości



Rys.4. Widma chwilowe dla trzech wzajemnie bliskich pozycji okna oraz przebiegi wartości widma w drugim i czwartym punkcie widma w funkcji pozycji okna.

framy oraz przebiegi wartości widma w punktach 2 i 4 w funkcji pozycji okna kroczącej o okres próbkowania analizowanego sygnału. Przebiegi te mogą być wykorzystane do wyznaczenia częstotliwości podstawowej w segmentach dźwięcznych sygnału mowy pozbawionego pierwszej harmonicznej. Przytoczone przykłady ilustrują, iż widma uzyskane dla kilku wzajemnie bliskich pozycji okna są niejednakowe.

2.5. Selekcja segmentów sygnału mowy.

Program umożliwia precyzyjne wskazanie granic segmentu czasowego sygnału mowy przewidzianego do wyselekcjonowania oraz utworzenie tablicy ze współrzędnymi takich granic dla większej liczby segmentów. Długość segmentu może wynosić od wartości rzędu długości jednego okresu podstawowego sygnału mowy aż do wartości równej rozciągłości czasowej całej badanej wypowiedzi. Wyboru granic segmentu dokonuje się dwuetapowo. Zgrubne wskazanie przeprowadza się za pomocą dwóch kursorów przesuwanych po spektrogramie. Kursory czerwony i zielony służą odpowiednio do wskazania fram początkowej i końcowej segmentu. Precyzyjnego ustalenia granic segmentu dokonuje się na dwóch oscylogramach. Pierwszy z nich odnosi się do framy, na której postawiono lewą granicę segmentu na spektrogramie a drugi do framy wybranej na spektrogramie jako prawa granica segmentu. Za pomocą dwóch kursorów przesuwanych po tych oscylogramach wskazać można precyzyjnie granice wybranego na spektrogramie segmentu. W trakcie przesuwania kursory te zatrzymują się jedynie w miejscach, w których oscylogram wskazuje na wartość chwilową sygnału zbliżoną do zera. Spośród wielu takich miejsc wybrać należy jedynie te, które mogą wskazywać na początek okresu podstawowego. Współrzędne granic wybranego segmentu zostają wpisane na kolejną pozycję lub początek listy przeznaczonej na współrzędne granic wielu segmentów. Listę tę można zapamiętać na dysku pod dowolną 8-znakową nazwą z rozszerzeniem *ta*.

2.6. Odsłuch segmentów.

Na podstawie listy współrzędnych granic pobranej z dysku lub zakładanej na bieżąco mogą być odtwarzane segmenty sygnału mowy również przepisanego z dysku lub pochodzącego ze świeżej wypowiedzi. System aktualnie umożliwia 3 warianty odtwarzania lecz ich liczbę i rodzaj zależnie od potrzeb użytkownika da się łatwo zwiększyć.

Odtwarzanie według pierwszego wariantu można określić jako proste. Na podstawie danych zawartych na jednej z pozycji tablicy współrzędnych granic odtworzony zostaje odpowiedni segment sygnału mowy. Odtworzenie może zostać wielokrotnie powtórzone. Numer pozycji tablicy, z której pochodzą współrzędne granic segmentu może być wybrany według kolejności lub losowo. Po skończeniu odtwarzania danego segmentu możliwe jest przerwanie bieżącego cyklu odtwarzań serii segmentów lub przejście do odtwarzania innego segmentu wskazanego losowo lub segmentu kolejnego. Wybór segmentów na podstawie tablicy może być jednak w całej serii odtwarzań albo losowy albo chronologiczny.

Drugi wariant odtwarzania przewidziany jest dla segmentów o rozciągłości jednego okresu podstawowego. Polega on na cyklicznym powtarzaniu segmentu bez przerwy między kolejnymi powtórzeniami. Przyjęto liczbę powtórzeń równą 20. Wybór pomiędzy obiema wariantami odtwarzania następuje automatycznie na podstawie oceny różnicy wartości współrzędnych granic segmentu na spektrogramie. Dla segmentów o rozciągłości okresu podstawowego różnica ta wynosi zero.

Trzeci wariant odtwarzania zawiera cechy obu poprzednich. Opiera się on o założenie, że współrzędne odczytane z tablicy dotyczą segmentu o rozciągłości okresu podstawowego i że na następnych pozycjach tej tablicy znajdują się dane o granicach podobnego segmentu oddalonego wstecz o n fram. Przy spełnieniu tego warunku możliwy jest odsłuch segmentu, którego środkowa część ma rozciągłość n stycznych fram o

odpowiednio wielokrotnymi powtórzeniami segmentów 1-okresowych pochodzących odpowiednio z pierwszej i ostatniej z tych fram.
W trakcie projekcji akustycznej każdego segmentu na ekranie wyżej opisane kursory wskazują jego granice - zgrubnie na spektrogramie a dokładnie na oscylogramach. W przypadku 3-go wariantu odtwarzania granice segmentu wyświetlane są niejednocześnie. Wpierw krótkotrwale ukazana zostaje lewa granica, a następnie prawa, pozostając na ekranie do czasu projekcji akustycznej albo ponownie dźwięku ostatnio usłyszanego albo dźwięku nowego.

3. Testy odsłuchowe wyizolowanych segmentów mowy.

Automatyczna identyfikacja granic międzygłoskowych nastręcza wiele trudności szczególnie w przypadku dyftongów. Opisany wyżej system umożliwił przeprowadzenie badania, którego celem było uzyskanie na razie uproszczonego poglądu o relacji pomiędzy wrażeniami słuchowymi wywołanymi przez wielokrotnie powielane segmenty mowy o rozciągłości okresu podstawowego zwane mikrosegmentami i przez segmenty dłuższe utworzone z ciągu kilku kolejnych fram o długości 240 próbek sygnału każda.

3.1. Wyznaczenie granic segmentów.

Zdigitalizowany sygnał mowy pochodzący z wypowiedzi wyrazu *biały* głosem męskim począwszy od głoski *i* przedstawiony został jako ciąg stycznych fram o długości jak wyżej. W obrębie każdej takiej framy wybrano jeden okres podstawowy i wyznaczono współrzędne jego granic. Uzyskane stąd dane zapisano we wspomnianej wyżej tablicy. Tablicę tę zapamiętano następnie na dysku. W tym samym ciągu próbek sygnału wyznaczono też granice segmentów o rozciągłości 6-iu kolejnych fram. Sąsiednie takie segmenty oddalone były od siebie o długość jednej framy. Współrzędne ich granic zapisano również w tablicy a następnie zapamiętano na dysku.

3.2. Materiał do oceny słuchowej.

Na podstawie zapisanych na dysku danych o granicach segmentów tworzono trzy serie bodźców słuchowych przeznaczonych do indywidualnej oceny przez kilku słuchaczy. Bodźce pierwszej serii tworzyły wyżej zdefiniowane cyklicznie powtarzane mikrosegmenty. Bodźcami drugiej serii były segmenty 6-cio framowe. Bodźce trzeciej serii tworzyły także segmenty 6-framowe lecz poszerzone o cyklicznie powtarzane inicjujące i kończące je mikrosegmenty. W zakresie każdej serii bodźce podawane były poszczególnym słuchaczom losowo. Wybór tego rodzaju bodźców podyktowany był potrzebą uzyskania odpowiedzi na następujące pytania: Z jakim odbiorem słuchaczy spotykają się segmenty 6-cio framowe w porównaniu z odbiorem wchodzących w ich skład pojedynczych mikrosegmentów ? Czy wszystkie mikrosegmenty pochodzące z wypowiedzi wyrazu *biały* są identyfikowane wyłącznie z głoskami, jakie występują w tym wyrazie, czy też oprócz tego z głoskami, których w tym wyrazie brak lub z dźwiękami, które są nieznane w mowie polskiej ? Czy wszystkie dłuższe segmenty wypowiedzi tego wyrazu słyszane są wyłącznie jako izolowane głoski lub połączenia głosek występujących w tym wyrazie ? Odpowiedzi na te pytania ewentualnie potwierdzone lub skorygowane później wynikami dalszych badań posłużą do tworzenia algorytmu identyfikacji głosek w mowie ciągłej.

3.3 Wyniki oceny słuchowej.

Wyniki odsłuchów zamieszczono w dwóch układach tabelarycznych w tablicach nr 1 oraz nr. 2a...2d. W pierwszym oceny kolejnych mikrosegmentów wypowiedzi dostarczone przez poszczególnych słuchaczy umieszczono w oddzielnych kolumnach. Początek kolumny odpowiada zawsze miejscu wypowiedzi blisko jej początku a pozycje kolejnych wierszy kolumny odnoszą się do kolejno następujących fragmentów wypowiedzi. W tablicy 1 zaznaczono, które wiersze dotyczą ocen mikrosegmentów pochodzących z przedziałów

dotyczą ocen mikrosegmentów pochodzących z przedziałów ustalonych głosek.

Tablica nr 1. Wyniki oceny słuchowej cyklicznie powtarzanych mikrosegmentów z wypowiedzi wyrazu *biały*.

i	1	i	i	i	i	i
	2	i	i	i	i	i
	3	y	y	i	i	i
	4	y	x	y	y	i
	5	y	y	y	y	y
	6	y	y	y	y	y
	7	e	e	e	y	e
	8	e	e	e	e	e
	9	e	e	e	e	e
	10	e	e	e	e	e
a	11	x	x	x	a	e
	12	x	a	e	a	a
	13	x	a	a	a	a
	14	x	a	a	a	a
	15	a	a	a	a	a
	16	a	a	a	a	a
	17	a	a	a	a	a
	18	a	a	a	a	a
	19	o	a	o	o	o
	20	o	x	o	a	a
w	21	u	x	u	x	u
	22	u	uu	u	u	u
	23	u	u	u	y	u
	24	u	u	u	x	u
	25	u	x	u	x	u
	26	u	y	u	y	u
	27	x	x	x	u	y
	28	x	y	y	y	y
	29	y	y	y	x	y
t	30	y	y	y	y	y
	31	y	y	y	y	y
	32	y	y	y	y	y
	33	y	y	y	y	y
	34	y	y	y	y	y
	35	y	y	y	i	i
	36	x	x	i	i	i
	37	x	x	y	i	i

Tablica nr 2. Wyniki oceny słuchowej cyklicznie powtarzanych mikrosegmentów, segmentów 6-framowych oryginalnych i rozszerzonych (o cyklicznie powtarzane mikrosegmenty z fram skrajnych) z wypowiedzi wyrazu *biały*. Część *a* tablicy nr 2.

1	i	i	y	y	y	y								ije	ije
1	i	i	y	x	y	y								ije	ijy
1	i	i	i	y	y	y								xj	iy
1	i	i	i	y	y	y								je	ije
1	i	i	i	i	y	y								je	ije
2		i	y	y	y	y	e							ije	ije
2		i	y	x	y	y	e							je	ijx
2		i	i	y	y	y	e							je	ije
2		i	i	y	y	y	y							je	ije
2		i	i	i	y	y	e							je	ije
3			y	y	y	y	e	e						yje	yje
3			y	x	y	y	e	e						je	ije
3			i	y	y	y	e	e						jex	ije
3			i	y	y	y	y	e						je	ije
3			i	i	y	y	e	e						je	ije
4				y	y	y	e	e	e					yja	yje
4				x	y	y	e	e	e					je	ije
4				y	y	y	e	e	e					bia	ije
4				y	y	y	y	e	e					je	ije
4				i	y	y	e	e	e					je	ije
5					y	y	e	e	e	e				yja	ye
5					y	y	e	e	e	e				ja	yje
5					y	y	e	e	e	e				ja	yje
5					y	y	y	e	e	e				je	yje
5					y	y	e	e	e	e				ja	yje
6						y	e	e	e	e	x			ea	ya
6						y	e	e	e	e	x			ja	yja
6						y	e	e	e	e	x			ja	yja
6						y	y	e	e	e	a			ja	yja
6						y	e	e	e	e	e			ja	yja
7							e	e	e	e	x	x		xa	ea
7							e	e	e	e	x	a		ja	yxa
7							e	e	e	e	x	e		ja	yxa
7							y	e	e	e	a	a		ja	ya
7							e	e	e	e	e	a		ja	yja
8								e	e	e	x	x	x	ea	ea
8								e	e	e	x	a	a	xa	exa
8								e	e	e	x	e	a	xa	exa
8								e	e	e	a	a	a	ja	ea
8								e	e	e	e	a	a	ea	ea

Część b tablicy nr 2.

```
9    e  e  x  x  x  x     ea     ea
9    e  e  x  a  a  a     xa    exa
9    e  e  x  e  a  a     xa    exa
9    e  e  a  a  a  a     xa     ea
9    e  e  e  a  a  a     ea     xa

10      e  x  x  x  x  a     aa     ex
10      e  x  a  a  a  a      a     ea
10      e  x  e  a  a  a      x    exa
10      e  a  a  a  a  a     xa     ea
10      e  e  a  a  a  a      x     ea

11         x  x  x  x  a  a     ax     ea
11         x  a  a  a  a  a      a    exo
11         x  e  a  a  a  a      a     ex
11         a  a  a  a  a  a      a     ea
11         e  a  a  a  a  a      x     ea

12            x  x  x  a  a  a     all     ax
12            a  a  a  a  a  a      a     ao
12            e  a  a  a  a  a      a    axx
12            a  a  a  a  a  a      a      a
12            a  a  a  a  a  a      a     aa

13               x  x  a  a  a  a     au     ao
13               a  a  a  a  a  a      a     ao
13               a  a  a  a  a  a      a   axll
13               a  a  a  a  a  a     ax     ax
13               a  a  a  a  a  a     ax     ax

14                  x  a  a  a  a  o     all     au
14                  a  a  a  a  a  a      au    aox
14                  a  a  a  a  a  o     all    all
14                  a  a  a  a  a  o     all     au
14                  a  a  a  a  a  o      au    all

15                     a  a  a  a  o  o     au     au
15                     a  a  a  a  a  x     au     au
15                     a  a  a  a  o  o     all    all
15                     a  a  a  a  o  a     all    all
15                     a  a  a  a  o  a     au     au

16                        a  a  a  o  o  u    allu     au
16                        a  a  a  a  x  x     all     au
16                        a  a  a  o  o  u     all    all
16                        a  a  a  o  a  x     all    all
16                        a  a  a  o  a  u      au    all
```

Część c tablicy nr 2.

```
17   a  a  o  o  u  u      au      au
17   a  a  a  x  x uu      au    al1u
17   a  a  o  o  u  u     al1    al1x
17   a  a  o  a  x  u     al1     al1
17   a  a  o  a  u  u      au      au

18      a  o  o  u  u  u       ax      au
18      a  a  x  x uu  u     al1    ol1u
18      a  o  o  u  u  u     al1    al1u
18      a  o  a  x  u  y     ol1      au
18      a  o  a  u  u  u      au    al1u

19         o  o  u  u  u  u       ul      ou
19         a  x  x uu  u  u     xl1    ol1u
19         o  o  u  u  u  u       x     ol1
19         o  a  x  u  y  x     ol1     ol1
19         o  a  u  u  u  u      xu      au

20            o  u  u  u  u  u       ux      ou
20            x  x uu  u  u  x        u    ol1u
20            o  u  u  u  u  u        u      ou
20            a  x  u  y  x  x        u     ul1
20            a  u  u  u  u  u        u      ax

21               u  u  u  u  u  u       ul      ux
21               x uu  u  u  x  y        u     ul1
21               u  u  u  u  u  u        u    ol1y
21               x  u  y  x  x  y        u      ux
21               u  u  u  u  u  u       l1      ul

22                  u  u  u  u  u  x      uly      ux
22                 uu  u  u  x  y  x        u     ul1
22                  u  u  u  u  u  x       ux    ul1y
22                  u  y  x  x  y  u        u     ul1
22                  u  u  u  u  u  y       l1      ux

23                     u  u  u  u  x  x      yly      uy
23                     u  u  x  y  x  y        x    ul1x
23                     u  u  u  u  x  y      l1y    ul1y
23                     y  x  x  y  u  y        u    ul1y
23                     u  u  u  u  y  y       l1      ul

24                        u  u  u  x  x  y      yly      uy
24                        u  x  y  x  y  y     xl1y    ul1y
24                        u  u  u  x  y  y      l1y    xl1y
24                        x  x  y  u  y  x       l1    ul1y
24                        u  u  u  y  y  y       uy     uly
```

Część *d* tablicy nr 2.

25	u	u	x	x	y	y	uy	uy
25	x	y	x	y	y	y	ulli	uxx
25	u	u	x	y	y	y	lly	lly
25	x	y	u	y	x	y	lly	uliy
25	u	u	y	y	y	y	uy	yly
26	u	x	x	y	y	y	uji	ully
26	y	x	y	y	y	y	lli	uxy
26	u	x	y	y	y	y	lly	lly
26	y	u	y	x	y	y	xy	xlly
26	u	y	y	y	y	y	uy	yly
27	x	x	y	y	y	y	yji	yuy
27	x	y	y	y	y	y	ulli	xx
27	x	y	y	y	y	y	lly	lly
27	u	y	x	y	y	y	y	xx
27	y	y	y	y	y	y	uy	uy
28	x	y	y	y	y	y	yji	xxy
28	y	y	y	y	y	y	ui	uxx
28	y	y	y	y	y	y	lly	uxj
28	y	x	y	y	y	y	y	xj
28	y	y	y	y	y	y	uyj	uji
29	y	y	y	y	y	y	yji	yx
29	y	y	y	y	y	y	xy	yx
29	y	y	y	y	y	y	y	yxi
29	x	y	y	y	y	y	yx	xji
29	y	y	y	y	y	y	yj	yji
30	y	y	y	y	y	y	yi	yy
30	y	y	y	y	y	y	i	yxx
30	y	y	y	y	y	y	yx	yji
30	y	y	y	y	y	i	y	yji
30	y	y	y	y	y	i	yj	yji
31	y	y	y	y	y	x	yji	yj
31	y	y	y	y	y	x	y	yxi
31	y	y	y	y	y	i	y	yxi
31	y	y	y	y	i	i	y	yj
31	y	y	y	y	i	i	yx	yji
32	y	y	y	y	x	x	yx	yx
32	y	y	y	y	x	x	y	yxi
32	y	y	y	y	i	y	y	yx
32	y	y	y	i	i	i	x	yj
32	y	y	y	i	i	i	yx	yi

W drugim układzie tabelarycznym dane zawarte w kolejnych pięciu wierszach z jednakowym marginesem odnoszą się do jednej grupy 6-ciu fram. W początkowych kolumnach każdego z pięciu wierszy figurują dostarczone przez różnych słuchaczy oceny kolejnych cyklicznie powtarzanych mikrosegmentów pochodzących z tych fram. W dwóch ostatnich kolumnach mieszczą się oceny segmentów będących ciągiem tych fram. Oceny zawarte w przedostatniej kolumnie dotyczą ściśle takich segmentów, natomiast oceny w ostatniej kolumnie odnoszą się do tychże samych segmentów 6-framowych poszerzonych o wielokrotne powtórzenie pierwszego i ostatniego mikrosegmentu w ciągu. Dane zawarte w następnych 5-ciu wierszach dotyczą kolejnej grupy sześciu fram o jedną framę dalszej od poprzedniej grupy. W tablicach literą x oznaczono oceny nieokreślone a znakami 11 oceny wskazujące na głoskę ł.

3.4. Wnioski z ocen słuchowych mikrosegmentów i segmentów.

Wpierw omówione zostaną oceny mikrosegmentów pochodzących z przedziałów ustalonych poszczególnych głosek. W przypadku samogłoski *i* oceny były właściwe w 70 procentach. Reszta ocen prócz jednej nieokreślonej wskazywała na głoskę *ł*. Mikrosegmenty z części ustalonej samogłoski *a* rozpoznano jako *a* w 80 %-ach. Reszta ocen za wyjątkiem dwóch wskazujących na *e* była nieokreślona. Mikrosegmenty z przedziału ustalonego spółgłoski *w* rozpoznano w blisko 80 procentach jako samogłoskę *u* raz jako *ł* a w reszcie przypadków oceny były nieokreślone. Mikrosegmenty z części ustalonej samogłoski *ł*, będącej ostatnią głoską badanej wypowiedzi zostały ocenione w 100% -ach właściwie.

Ocena nieokreślona może być uważana za neutralną. Zatem uzyskane oceny mikrosegmentów z części ustalonych rozpatrywanych głosek uznać należy za poprawne.

Oceny mikrosegmentów pochodzących z fragmentów nieustalonych wskazywały na inne głoski niż te, które stanowiły kontekst

obustronny fragmentu. I tak np. mikrosegmenty z początkowej 1/3 fragmentu wypowiedzi pomiędzy częściami ustalonymi samogłosek *i* oraz *a* oceniono jako *ɨ* a mikrosegmenty z pozostałej części tego fragmentu rozpoznano z jednym wyjątkiem jako *e*. Wyjątek ten wskazywał na *ɨ* i pojawił się w ciągu ocen wskazujących na *e*. Tak więc mikrosegmenty położone bliżej *i* oceniono jako *ɨ* a położone bliżej *a* jako *e*. Granica zmiany ocen nie przypadła w środku przedziału. Bardziej urozmaicony rozkład ocen uzyskano dla mikrosegmentów z przedziału obejmującego przejście od głoski *a* do *w*. 60% ocen wskazywało na *o* , 30% na *a* a jedna była nieokreślona. Jednakowe oceny nie tworzyły w pełni zwartych ciągów, lecz były przeplatane innymi ocenami. Podobnie przedstawia się układ ocen mikrosegmentów z przedziału pomiędzy częściami ustalonymi głosek *w* oraz *ɨ*. Prócz pięciu ocen nieokreślonych nie ma ocen, wskazujących na głoskę inną niż te, które przedział ten łączy. Oceny wypadły w tym przypadku nie równolicznie. Ponad 50% wskazywało na samogłoskę *ɨ* a jedynie 20% na *u*. Wynik ten dowodzi braku objawów dodatkowej głoski w przedziale stanowiącym przejście z głoski *w* do *ɨ*, chociaż na obrazie spektrograficznym widoczna jest długotrwała ewolucja głoski *w* w *ɨ*.

Oceny segmentów 6-io framowych prostych (bez rozszerzenia o cykliczne powtórzenie mikrosegmentu ze skrajnych fram ciągu) wskazywały na połączenia głosek, których nie zidentyfikowano w mikrosegmentach z poszczególnych fram tworzących segment. Np. w pierwszym ciągu fram postrzegano *j* oraz *e* , których nie zauważono w żadnym z mikrosegmentów pochodzących z fram tego ciągu. Podobna uwaga odnosi się do *j* oraz *a* odebranych w 5-tym i 6-tym ciągu fram, do *u* oraz *w* w ciągu 14-tym i 15-tym, a także częściowo do *i* oraz *j* usłyszanych w 26-tym, 27-ym, 28-ym i 29-tym ciągu fram. W niektórych ciągach fram nie spostrzeżono natomiast głosek, które zidentyfikowano oddzielnie w należących do tych fram mikrosegmentach. Np. poza sporadycznymi wyjątkami nie spostrzeżono głoski *ɨ* w ciągach od pierwszego do 6-tego *e* w ciągach : 7, 10, *o* w

ciągach : 15, 16, 17.

Oceny segmentów 6-framowych rozszerzonych o wielokrotne powtórzenie pierwszego i ostatniego mikrosegmentu ze skrajnych fram ciągu wskazują w wielu przypadkach na ich odbiór inny niż tych samych segmentów podawanych bez rozszerzenia a także na inny odbiór w nich skrajnych mikrosegmentów w porównaniu z odbiorem tychże w izolacji.
W wierszach o numerze 5 w tablicy 2a są oceny segmentów 6-framowych obejmujących dokładnie przedział pomiędzy częściami ustalonymi *i* oraz *ɑ* w badanej wypowiedzi. W przypadku segmentu bez rozszerzenia słuchacze niemal zgodnie odebrali go jako *jɑ*. Oceny te nie są jednak zgodne z ocenami cyklicznie powtarzanych mikrosegmentów pochodzących z fram tego segmentu. Żadna z tych ocen nie wskazywała na *j* lub *ɑ*. W segmencie rozszerzonym słuchacze właściwie rozpoznali cyklicznie powtarzane skrajne mikrosegmenty. Segmenty przejściowe pomiędzy *ɑ* oraz *u* i *u* oraz *ɨ* są krótsze niż 6 fram. W wierszach 23 i 26 tablicy nr 2 są odpowiednio podane oceny segmentu 6-framowego obejmującego fragment części ustalonej *u* oraz cały przedział przejściowy pomiędzy *u* oraz *ɨ* i oceny takiegoż segmentu obejmującego cały ten przedział przejściowy oraz kawałek części ustalonej *ɨ*. Oceny segmentów 6-framowych bez rozszerzenia są w tych przypadkach zróżnicowane natomiast oceny segmentów z rozszerzeniem wskazują, iż słuchacze konsekwentnie identyfikowali cyklicznie powtarzane skrajne mikrosegmenty i tym samym poprawnie rozpoznawali połączenie.
Zebrane oceny są wycinkowym przykładem złożoności problematyki segmentacji głoskowej sygnału mowy. Pokazują one nieostrość granicy między głoskami występującymi w badanej wypowiedzi oraz trudność w trafnym zaliczeniu kolejnych mikrosegmentów wypowiedzi do właściwej głoski.

B i b l i o g r a f i a

[1]. KUBZDELA, H., *Metoda globalnego rozpoznawania wyrazów na podstawie spektrogramów binarnych.* Prace IPPT 28/1986, Warszawa, 1986.

[2]. KUBZDELA, H., *Udoskonalenie reprezentacji sygnału mowy w formie obrazów binarnych*, Prace IPPT 24/1987, Warszawa, 1987.

[3]. KUBZDELA, H., *Badanie cech segmentów fonetyczno-akustycznych w uproszczonych reprezentacjach wyznaczonych z widm.* Prace IPPT 41/1991, Warszawa, 1991.

[4] NOWAK, I., *Automatyczna transkrypcja polszczyzny nieregionalnej (Odmiana północno-wschodnia i południowo - zachodnia)*, Prace IPPT 31/1991, Warszawa, 1991.

[5] STEFFEN-BATOGOWA, M., *Automatyzacja transkrypcji fonematycznej tekstów polskich*, PWN, Warszawa, 1975.