2.23 — rozpoznawanie mowy

W. Jassem, H. Kubzdela, P. Domagała

SEGMENTACJA SYGNAŁU MOWY NA PODSTAWIE ZMIAN ROZKŁADU ENERGII W WIDMIE

13/1983

WARSZAWA 1983

ISSN 0208-5658

Praca wpłynęła do Redakcji dnia 22 listopada 1982 r.

Na prawach rękopisu

Instytut Podstawowych Problemów Techniki PAN
Nakład 190 egz. Ark.wyd. 1,7. Ark.druk. 1,5.
Oddano do drukarni w lutym 1983 r.
Nr zamówienia 248/0/83 M-13 .

Warszawska Drukarnia Naukowa, Warszawa,
ul.Śniadeckich 8

Wiktor Jassem
Henryk Kubzdela
Piotr Domagała
Pracownia Fonetyki Akustycznej
IPPT PAN

SEGMENTACJA SYGNAŁU MOWY NA PODSTAWIE ZMIAN ROZKŁADU ENERGII W WIDMIE[1)]

Streszczenie

Niektórzy fonetycy i językoznawcy utrzymują, iż niemożliwa jest obiektywna segmentacja akustycznego sygnału mowy, jakkolwiek istnieje szereg prac tak teoretycznych, jak i eksperymentalnych wskazujących różne efektywne metody segmentacji. Ich automatyzacja wymaga jednak znacznych mocy obliczeniowych. Proponuje się nadzwyczaj prosty algorytm automatycznej segmentacji fonetycznej akustycznego sygnału mowy zaprogramowany na minikomputer biurowy MERA 303. Algorytm sprawdzono na materiale składającym się z wyrazów polskich (wymówionych przez 2 osoby), które dobrano tak, iż znalazły się w nich szczególnie trudne z punktu widzenia segmentacji sekwencje głosek. Około 80 % granic pomiędzy segmentami głoskowymi zaproponowany algorytm wykrywa poprawnie, co stanowi wynik nie gorszy od uzyskanych znacznie bardziej złożonymi metodami. Istota algorytmu polega na badaniu znaku różnicy poziomu sygnału między kolejnymi okienkami czasowo-częstotliwościowymi w poszczególnych przedziałach czasowych.

1. Ogólne problemy segmentacji.

W obrębie problematyki automatycznego rozpoznawania mowy przez

[1)] Praca wykonana w ramach problemu węzłowego 06.9.

człowieka lub maszynę pojęcie "segmentu" stosuje się, ogólnie biorąc, dla określenia czasowego fragmentu jednostki (zazwyczaj lingwistycznej) będącej elementem w zbiorze znaków podlegających rozpoznaniu. Najczęściej elementami takiego zbioru są wyrazy. Inaczej segmentem może być głoska, sylaba, część sylaby, przebieg od środka jednej sylaby do środka następnej itd. W przypadku modelu rozpoznawania przez układ biologiczny segment nie może być dobierany dowolnie, lecz musi być zgodny z faktycznymi procesami fizjologicznymi, neurologicznymi i psychologicznymi zachodzącymi w percepcji. Przy rozpoznawaniu automatycznym nie jest konieczne odwzorowywanie tych procesów i dlatego segmentem może być fragment sygnału, który nie odpowiada ani elementowi percepcyjnemu ani lingwistycznemu, chyba że zakłada się, iż w ramach problematyki bionicznej konstruuje się celowo układy odwzorowujące funkcje biofizyczne. Niektóre systemy automatycznego rozpoznawania mowy - ARM - mają takie założenia (por. np. Derkacz i in. 1971).

Chociaż istnieje obecnie bardzo wiele modeli i teorii percepcji mowy (Łobacz 1981), często w znacznym zakresie sprzecznych, zachodzi dość daleko idąca zgodność poglądów co do tego, że w normalnym procesie rozpoznawania mowy przez człowieka zostają w obrębie takich większych jednostek lingwistycznych jak zdanie (wypowiedź), czy wyraz, wydzielone co najmniej niektóre segmenty o rozciągłości głoski. Nie jest natomiast jasne, na którym poziomie percepcji następuje segmentacja głoskowa : audytorycznym (sensorycznym), fonetycznym, fonematycznym, czy też ewentualnie jeszcze wyższym.

Jeżeli system ARM nie jest z założenia odwzorowaniem funkcji biofizycznych, to postuluje się segmenty, które muszą spełniać co najmniej następujące warunki :

(1) Segment jest elementem skończonego zbioru o liczebności znacznie mniejszej niż liczebność zbioru elementów wyższego rzędu (np. liczba wyróżnianych sylab jest znacznie mniejsza niż liczebność zbioru rozpoznawanych wyrazów w założonym słowniku).

(2) Klasyfikacja (identyfikacja, rozpoznawanie) segmentów musi następować na podstawie parametrów akustycznych ekstraho-

wanych przez dany system z sygnału mowy (takich jak gęstość przejść przez zero, częstotliwość formantów, przebieg zmian poziomu natężenia w wybranych pasmach częstotliwości itd).

(3) Wyróżnione klasy segmentów winny wykazywać minimalną zmienność wewnątrzklasową przy maksymalnej zmienności międzyklasowej (jest to warunek ogólny, w różnych ujęciach matematycznych występujący w problematyce rozpoznawania obiektów).

Z powyższymi warunkami wiąże się ściśle zagadnienie granic pomiędzy segmentami.

Niektóre (nieliczne) systemy ARM nie zakładają konieczności wykrywania granic międzysegmentalnych w dziedzinie czasu (np. Hill 1970, 1972). Można wyznaczać pewne rozłączne i następujące sekwencyjnie w czasie przedziały czasowe, niekoniecznie styczne, w których występują charakterystyczne wartości lub cechy parametrów fonetyczno - akustycznych pozwalające na klasyfikację (identyfikację, rozpoznawanie) segmentów, bez wyznaczania granic między nimi. Liczne systemy, zwłaszcza należące do typu SUS (speech Understanding Systems) usiłujące rozpoznawać jednostki o rozciągłości zdania i stosujące w szerokim zakresie procedurę zstępującą ("top-down") dopuszczają wieloznaczność granic międzysegmentalnych (por. np. Wolf i Woods 1980).

W większości systemów ASR segment ma rozciągłość g ł o s k i i skorelowany jest z f o n e m e m. Fonem jest określoną klasą segmentów lub ich ciągów. Wśród różnych definicji fonemu są również takie, które wiążą się ze stadium a k u s t y c z n y m procesu łączności za pomocą mowy (poza którym wyróżnia się stadia psychologiczne, neurologiczne i fizjologiczne - tak w wytwarzaniu, jak i odbiorze sygnału). Takie definicje uwzględniają zarówno systematyczną, jak i losową zmienność intrafonematyczną. Jednym z czynników intrafonematycznej zmienności systematycznej jest zjawisko k o a r t y k u l a c j i (lub transartykulacji). W stadium wytwarzania sygnału pojęcie koartykulacji związane jest z ciągłością i asynchronicznością wyodrębnionych czynności artykulacyjnych (np. zwierania podniebienia miękkiego ze ścianą gardła, kształtowania toru ustnego itd.), zaś w stadium akustycznym przez koartykulację rozumie się równoczesne występowanie w określonych przedziałach czasowych, akustycznych cech cha-

rakterystycznych dla dwóch ewentualnie trzech sąsiadujących na osi czasu realizacji fonemów. Zjawisko koartykulacji prowadziło od dawna do tworzenia teorii języka i modeli mowy, w których odrzuca się segmentację w wymienionych dwóch stadiach łączności : "Poza tym trzeba zwrócić uwagę na najpoważniejszą klęskę modeli tak fizykalistycznych, jak i interakcyjnych. Oba traktują fonemy i allofony jako jednostki dyskretne, jako segmenty. Jednakże w rzeczywistych fizycznych artykulacjach (ani, jeśli idzie o to, w sygnale akustycznym) nie podobna znaleźć dyskretnych elementów dających się skorelować z takimi segmentami. Segmentacja, która jest oczywiście rodzajem klasyfikacji, jest wyłącznie zjawiskiem psychicznym. Poza psychiką po prostu żadnych segmentów nie ma". Tak pisze całkiem niedawno fonetyk-językoznawca Hammarberg (1982). Jeśli taki głos pochodzi od specjalisty mniejszego kalibru, możnaby go pominąć, ale na pewno nie można zignorować zdania Studderta-Kennediego, uważanego za czołowego specjalistę w zakresie percepcji mowy w skali światowej. Pisał on, również zupełnie niedawno : "Fonetyczne segmenty nie istnieją ani w ruchach artykulacyjnych, ani (tym bardziej) w sygnale akustycznym"(Studdert-Kennedy 1981, str. 9). Takie zapatrywania, jak wyżej cytowane, są składową częścią spekulacyjnych i nader kontrowersyjnych teorii lingwistycznych oraz modeli percepcji, w szczególności metafizycznej (por. Lass 1976, str. 213-220) fonologii generatywnej i psychologistycznej motorycznej teorii percepcji mowy (por. Pisoni 1971). Na poparcie takiej postawy "antysegmentalnej" brak jakichkolwiek dowodów. Są one gołosłowne i ignorują świadomie lub nieświadomie łatwo dla ich autorów dostępne wyniki prac doświadczalnych i pomiarowych dających świadectwo obiektywnej segmentacji tak psychoakustycznej (sensorycznej), jak i fizyko-akustycznej.

Stwierdzenie granic międzysegmentalnych w akustycznym sygnale mowy było trudne z końcem ubiegłego i na początku bieżącego stulecia, kiedy przyrządy wizualizujące ten sygnał (np. kimograf) były obciążone znaczną bezwładnością mechaniczną. Ale już w latach trzydziestych ukazała się poważna i szeroko zakrojona praca wskazująca na istnienie granic międzysegmentalnych tak w sygnale, jak i prymarnym wrażeniu słuchowym (Menzerath i de Lacerda 1933).

Öhman (1962) wskazał na zgodność granic fizyko-akustycznych z sensorycznymi.

Jeśli sygnał mowy przedstawiony jest w postaci chwilowej wartości ciśnienia lub prędkości objętościowej w funkcji czasu, wówczas jego obraz jest z jednej strony silnie redundantny, a z drugiej niektóre właściwości sygnału, które mogą być przydatne dla jego segmentacji są mało widoczne (np. zmiany odpowiadające przesunięciom rozkładu mocy w widmie dynamicznym). Mimo to segmentacja fonetyczna da się na takim odwzorowaniu przeprowadzić, jak to ukazał Reddy (1967 a i 1967 b). Segmentacja taka umożliwia szczegółowe badania iloczasu głoskowego (np. Frąckowiak-Richter 1973). Jednakże znacznie bardziej wyraźnie występują granice międzysegmentalne w takim odwzorowaniu sygnału mowy, które przedstawia zmiany rozkładu energii w widmie dynamicznym. Ogólne założenia segmentacji takiego odwzorowania przedstawiono m.in. w pracach jednego z współautorów (Jassem 1970 i 1971), a proste ich sformułowanie matematyczne zaproponowano w jednej z dalszych jego prac (1977). Szczegółowe zasady opisowej segmentacji spektrogramów podali wcześniej Lehiste i Peterson (1960 str. 4-11) oraz Fant (1964). Fant zwrócił uwagę na to, że poszczególne głoski mogą być monosegmentalne lub polisegmentalne. Jak wynika ze szczegółowych nowszych opisów różnych języków, niektóre zasady określające liczbę segmentów składających się sekwencyjnie na głoskę są specyficzne dla poszczególnych języków (por. np. Jassem 1979).

Spośród różnych form wizualnego odwzorowania sygnału mowy klasyczne spektrogramy typu Visible Speech (np. otrzymywane z analizatora spektrograficznego Sona-Graph) wykazują różnego rodzaju granice międzysegmentalne w sposób najbardziej naoczny, co było jedną z przyczyn, dla których takie odwzorowanie stosowano w latach czterdziestych w Stanach Zjednoczonych dla celów rewalidacji osób głuchych w zakresie odbioru sygnału mowy poprzez zastąpienie wrażeń słuchowych przez wzrokowe (Potter, Kopp i Green 1947). Jednakże algorytmizacja segmentacji na podstawie analizy spektrograficznej tradycyjnego typu w sformułowaniu matematycznym jest niezwykle trudna i wymagałaby znacznych mocy obliczeniowych przy jej implementacji, jako że

ilość informacji zawarta w takim spektrogramie jest bardzo znaczna (szacunkowo rzędu kilkudziesięciu tysięcy bitów na sekundę).

W systemach automatycznego rozpoznawania mowy, szczególnie w systemach SUS, ekstrahuje się różne parametry akustyczne, redukujące ilość informacji, które służą tak do segmentacji, jak i rozpoznawania elementów o rozciągłości głoskowej. Do takich parametrów (Vaissière 1981) należą : częstotliwości formantowe, rodzaje ugięć formantowych, częstotliwość podstawowa, kształt obwiedni widma, gęstość przejść przez zero, współczynniki predykcji liniowej, obwiednia czasowa poziomu natężenia w różnych zakresach częstotliwości i kilka dalszych, mniej typowych. Jeden z szczególnie efektywnych systemów segmentacji wykorzystujący takie właśnie parametry przedstawili Gresser i Mercier (1975).

2. Układ analogowo-cyfrowy i jego wykorzystanie.

Część pomiarową wykonano przy pomocy układu analogowo-cyfrowego, w którego skład wchodzą : 63-kanałowy analizator widma, interface łączący analogowe źródło sygnału z minikomputerem, minikomputer MERA 303. Analogowy analizator widma posiada 43 pasma analizy o stałej szerokości wynoszącej 80 Hz pokrywające zakres częstotliwości od 120 do 3560 Hz oraz 20 pasm o szerokości zależnej od częstotliwości środkowej w sposób liniowy i pokrywających zakres od 3560 do 8310 Hz. Wyjścia poszczególnych kanałów analizatora są cyklicznie załączane na wspólny tor wyjściowy. Czas T_k, który dzieli kolejne momenty załączenia tego samego kanału wynosi ok. 23 ms i może być regulowany. Każdy kanał pozostaje w stanie połączenia ze wspólnym wyjściem przez czas Δt, wynoszący ok. 100 us. Czas Δt wynika z długości trwania zautomatyzowanej operacji odczytu danej wyjściowej z analizatora, przekształcenia jej do postaci cyfrowej i umieszczenia pod zadanym adresem w pamięci operacyjnej minikomputera. Czas T podyktowany jest rozmiarem pamięci minikomputera, w której winien zmieścić się ciąg próbek widmowych sygnału mowy o długości trwania ok. 1 sekundy. Pamięć operacyjna minikomputera MERA 303 wynosi zaledwie 8 kbajtów. Podczas konwersji analogowo-cyfrowej próbki widmowe są logarytmowane

i przyjmują wartości w skali decybelowej. Na etapie wpisu do minikomputera zachodzi również wygładzanie widma, które polega na uśrednianiu w obrębie każdych czterech kolejnych danych widmowych zgodnie z wyrażeniem

$$\bar{a}_i = \frac{1}{4} \sum_{j=0}^{3} a_{i+j} \cdot \qquad (1)$$

Wskutek uśrednienia liczba parametrów widmowych zostaje zmniejszona o 3. W pracy posługiwano się systemem operacyjnym, który umożliwia wykonywanie następujących działań na danych widmowych zapisanych w pamięci operacyjnej minikomputera :

1. Wydruk spektrogramu cyfrowego, czyli ciągu kolumn liczbowych reprezentujących wygładzone widma.
2. Wydruk różnicowego spektrogramu cyfrowego RSC, czyli ciągu kolumn liczb i znaków wyrażających różnice wartości tych samych składowych dwóch bezpośrednio po sobie następujących widm wygładzonych. Znak np. w wierszu j i kolumnie i oznacza, że parametr j w widmie odnoszącym się do momentu t_i ma wartość większą niż w widmie pochodzącym z momentu t_{i-1}.
3. Wydruk różnicowego spektrogramu znakowego RSZ, który stanowi uproszczoną wersję RSC, gdyż zawiera jedynie znaki bez liczb.
4. Wyznaczenie spektrogramu binarnego według zasad podanych w pracach Kubzdela 1980 oraz 1981 .
5. Wydruk spektrogramu binarnego.
6. Przepisanie spektrogramu cyfrowego z pamięci operacyjnej minikomputera do pamięci zewnętrznej na dysku elastycznym i odwrotnie.

3. Algorytm automatycznej segmentacji

Ryc. 1 przedstawia spektrogram cyfrowy wyrazu "płaszczyzna" (/pwaʃtʃaźna/) wymówionego przez głos męski (WJ). Oś czasu (pozioma) wyskalowana jest w jednostkach kwantyzacji, tj. w odstępach Δt = 23 ms. Był to najkrótszy skok kwantyzacji czasowej możliwy do uzyskania w danych warunkach technicznych. Na osi częstotliwości wyprowadza się z maszyny numer pasma analizy (pierwsza kolumna z lewej). Z prawej strony wydruku dla orien-

tacji wpisano wartości częstotliwości środkowych niektórych pasm analizy. Liczby na spektrogramie podają uśrednioną za skok Δt wartość poziomu sygnału w umownych jednostkach powyżej przyjętego doświadczalnie poziomu odniesienia. Wartość jednostki umownej wynosi ok. 0.6 dB. Na ryc. 2 widnieje tzw. spektrogram różnicowy. Zawiera on wartości różnicy poziomów w każdym paśmie częstotliwościowym pomiędzy kolejnymi wartościami. Różnice podane są w wartościach dodatnich lub ujemnych, w zależności od tego, czy następuje wzrost czy spadek poziomu. Podstawą przyjętego w niniejszej pracy algorytmu automatycznej segmentacji jest jedynie z n a k różnicy występujący w poszczególnych kolumnach (widmach chwilowych). Znaki te wykazują określone regularności. Na przykład w kolumnach 000 i 030 wszystkie znaki są dodatnie, natomiast w kolumnach 024 i 080 występują wyłącznie znaki ujemne. Ponadto, w niektórych kolumnach pojawiają się w kolejnych niższych pasmach częstotliwości wyłącznie minusy, a w wyższych wyłącznie plusy lub na odwrót, np. w kol. 032 najniższe pasmo ma plus, a pozostałe wyłącznie minus. Kolejność znaków w kolumnie 048 jest przeciwna.

Na ryc. 3, przedstawiającej spektrogram binarny tej samej wypowiedzi, w tych samych kolumnach, w bezpośrednio sąsiednich lub co najwyżej z przesunięciem o 2 skoki czasowe występują bardzo wyraźne zmiany obrazu widma binarnego w postaci zamiany znacznej części kolejnych (pionowo) znaków z zera na jedynkę lub odwrotnie, jak na przykład (odpowiednio do wymienionych wyżej) w kolumnach 000, 022, 032 oraz 078. Na podstawie takich wyraźnych zmian widma binarnego oraz opisanych w literaturze akustyczno-fonetycznej cech akustycznych głosek różnego typu wyznaczyć można granice międzysegmentalne wizualnie, przy czym należy pamiętać, że niektóre głoski są polisegmentalne, tj. składają się z 2 lub większej liczby segmentów. W górnej części ryc. 3 oznaczono przeprowadzoną wizualnie segmentację oraz wpisano między granicami, które nazywać będziemy c e z u r a m i, oznaczenia fonetyczne segmentów na podstawie znajomości wymówionego wyrazu. Segmentacja wizualna na ogół nie pozwala

na rozgraniczenie, w każdym razie na spektrogramach binarnych, (ale często też na spektrogramach analogowych) elementów dyftongów - w tym przypadku /wa/. Nagłosowa głoska /p/ jest na spektrogramie niewidoczna, gdyż jej plozja jest za słaba.

U dołu spektrogramu binarnego widoczna jest segmentacja automatyczna w postaci liter A, B, D lub G, przy czym kolejne 2 litery B oznaczają jedną cezurę. Automatyczna segmentacja dzieli wypowiedź zilustrowaną na ryc. 3 na podobną liczbę fragmentów, co segmentacja wizualna, a cezury wyznaczone automatycznie są zgodne czasowo z wizualnymi z dopuszczeniem przesunięcia co najwyżej o 2 skoki czasowe. Zastosowany algorytm zawiódł w wyznaczaniu granicy cezury /a-ʃ/. Na ryc. 4 przedstawiono spektrogram binarny wypowiedzi "chciałbym" /'xtɕawbɨm/ (głos WJ). Podobnie jak na ryc. 3, u góry zaznaczono segmenty wizualne. Kwazi-dyftong /aw/ (por. wyż.) nie został rozsegmentowany.

Przez "sekwencję" będziemy rozumieli ciąg takich samych znaków kolejnych w kolumnie. Oznaczamy :

$c(k)$ - liczba ciągów jednoznakowych w k-tej kolumnie

$z(k)$ - znak ostatniego ciągu obejmującego kanały o najwyższych numerach

$$z(k) = \begin{cases} 0 \text{ dla ciągu dodatniego} \\ 1 \text{ dla ciągu ujemnego} \end{cases}$$

Def. cezury A

Cezura A występuje w k-tej kolumnie, jeśli :

$$c(k) = 1 \wedge \left[c(k+1) \neq 1 \vee \left[c(k+1) = 1 \wedge z(k+1) \neq z(k) \right. \right.$$

Def. cezury B

Cezura B występuje w k-tej i i w k+1-ej kolumnie, jeśli :

$$c(k) = c(k+1) = 1 \wedge z(k) = z(k+1) \wedge \left\{ c(k+2) \neq 1 \vee \left[c(k+2) = 1 \wedge z(k+2) \neq z(k) \right] \right\}$$

Def. cezury C

Cezura C występuje w k-tej i k+n-1-ej kolumnie, jeśli :

$c(k) = c(k+i) \wedge z(k) = z(k+1) \wedge [c(k+n) \neq 1 \vee [c(k+n) = 1 \wedge z(k) \neq z(k+n)]]$

gdzie $i=1,2,\ldots,n-1$ oraz $n \geqslant 3$

Def. cezury D

Cezura D występuje w k-tej kolumnie, jeśli :

$c(k) = 2 \wedge c(k-1) \neq 1 \wedge c(k+1) \neq 1 \wedge c(k+2) \neq 2$

Def. cezury F

Cezura F występuje w k-tej kolumnie, jeśli :

$c(k) = 2 \wedge c(k-1) \neq 1 \wedge c(k+1) = 2 \wedge c(k+i) = 2 \wedge z(k) = z(k+i) \wedge z(k) \neq z(k+n) \wedge c(k+n) = 2,$

gdzie $i=0,1,2,\ldots,n-1$ oraz $n \geqslant 1$

Def. cezury G

Cezura G występuje w k-tej kolumnie, jeśli :

$c(k) = 2 \wedge c(k-1) \neq 1 \wedge c(k+1) = 2 \wedge c(k+i) = 2 \wedge z(k) = z(k+i) \wedge c(k+n) \neq 1 \wedge c(k+n) \neq 2,$

gdzie $i=0,1,\ldots,n-1$ oraz $n \geqslant 2$

Oznaczenie BB określa jedną cezurę rozmytą na dwa kolejne przedziały czasowe. Oznaczenie C...C określa dwie cezury : na początku i na końcu ciągu kolejnych kolumn.

Algorytm segmentacji automatycznej ukazany na ryc. 5 napisano w języku wewnętrznym minikomputera MERA 303 i zrealizowano na tymże komputerze. Realizację poszczególnych typów cezur przedstawia dodatkowo Tablica 1.

4. Test algorytmu automatycznej segmentacji

Algorytm automatycznej segmentacji został sprawdzony na liście kilkunastu wyrazów wymówionych naturalnie, bez przesadnej staranności, przez jeden głos męski (WJ) i jeden kobiecy (LFR), zapisanych na taśmie magnetofonowej w studio bezpogłosowym. Lista obejmowała następujące wyrazy : wiedźma, płaszczyzna, chciałbym, właściwy, pstrągi, sklepy, sprawca, barszczyk, kształty, listwa, płotkarz, chrząszcz, gęsty, częściej, pluskwa, błędny oraz sąsiedzki (głos WJ) i sąsiedzi (głos LFR). Wyrazy dobrano tak, aby znalazły się w nich liczne sekwencje, które według dotychczasowych danych w literaturze akustyczno-fonetycznej są szczególnie trudne do rozsegmentowania, a mianowicie dyftongi (np. /wa, aw, je/ itp.), sekwencje spółgłosek płynnych z następującą lub poprzedzającą samogłoską (np. /lu, ar/ itp.) oraz zbitki spółgłoskowe (np. /pstr, ʃtʃ , skf/ itp.). Z tego

punktu widzenia test algorytmu był szczególnie ostry.

Wyniki testu ukazuje następujące ogólne zestawienie :

głos WJ

zgod.	brak	nad.	A	B	C	D	F	G
129	25	13	70	23	12	30	3	4

głos LFR

zgod.	brak	nad.	A	B	C	D	F	G
122	23	9	75	21	8	19	4	4

W powyższym zestawieniu w kolumnie "zgod." podano łączną liczbę cezur zgodnych dla segmentacji wizualnej oraz automatycznej, w kolumnie "brak" liczbę cezur wizualnych, dla których zastosowany algorytm zawiódł, a w kolumnie "nad." liczbę artefaktów segmentacji automatycznej polegających na oznaczeniu dodatkowej cezury, której nie odpowiada ani cezura wizualna, ani teoretyczno-fonetyczna (tzn. oczekiwana na podstawie znajomości wyrazu i jego struktury fonetyczno-segmentalnej). W kolumnach oznaczonych literami A B C D F G podano liczbę odpowiednich rodzajów cezur według definicji na str.12 i algorytmu na ryc. 5. W głosie WJ dyftong /wo/ w "płotkarz" został rozsegmentowany tak wizualnie, jak i automatycznie, podobnie jak dyftong [õw̃] w"chrząszcz", natomiast w 8 przypadkach dyftongów i sekwencji spółgłoska płynna-plus-samogłoska (np. /lu/) spektrogramy nie wykazywały cezury wizualnej ani automatycznej. W głosie LFR rozsegmentowano oboma sposobami /wa/ w "właściwy", ale również w 8 przypadkach dyftongi i sekwencje "płynna-plus--samogłoska" nie zostały rozsegmentowane ani wizualnie ani automatycznie. Ponadto w 3 przypadkach spektrogramy nie wykazywały spółgłoski nagłosowej w głosie WJ (zawsze /p/). Odpowiednia liczba w głosie LFR wynosiła 5. Są to głoski, w których poziom sygnału jest zbyt niski dla zastosowanego systemu analizy. Zawodzą w takich okolicznościach wszystkie opisane w literaturze systemy automatycznego rozpoznawania mowy.

Efektywność zastosowanego algorytmu można obliczać w różny sposób. Najbardziej liberalnie można ją wyrazić porównując liczbę cezur wizualnych i automatycznych z pominięciem cezur nadmiarowych jako że takie - jeśli ich liczba jest względnie

niska - nie stanowią przeszkody w rozpoznawaniu automatycznym. W takim ujęciu 84 % cezur zaproponowany algorytm wykrył poprawnie w obu głosach. Przy ostrzejszym kryterium można tak braki, jak i cezury nadmiarowe uznać za błędy. Wówczas efektywność algorytmu wynosi w teście 77 % dla WJ oraz 79 % dla LFR. Przy najostrzejszym kryterium wymagającym wykrycia automatycznego wszystkich cezur tj. również tych, które nie są wizualnie uchwytne na spektrogramach binarnych, efektywność algorytmu wynosiłaby 70 % dla WJ i 75 % dla LFR. Byłyby to wyniki liczbowo niemal dokładnie takie same, jak otrzymane za granicą przy zastosowaniu nieporównywalnie większych mocy obliczeniowych (por. Vaissière 1981). Jednakże należy w rozpatrywanym tutaj przypadku uwzględnić fakt, iż materiał, na którym sprawdzano algorytm był celowo bardzo znacznie utrudniony, jako że względny udział dyftongów i trudnych fonetycznie zbitek spółgłoskowych był znacznie wyższy niż w normalnym tekście. Stąd wszystkie podane powyżej wartości procentowe należy uznać za wyraźnie niższe, niż te, których można się spodziewać w materiale reprezentatywnym.

Uwzględniając zatem niezwykle skromne warunki techniczne należy uzyskane wyniki uważać za bardzo zadowalające. Dodać trzeba, iż skrajna prostota algorytmu pozwala na jego realizację w trywialnym cyfrowym układzie wyspecjalizowanym.

Prowadzone bieżąco prace wskazują, iż uzyskawszy możliwość gęstszej kwantyzacji czasowej oraz stosując drobne zmiany w analizie częstotliwościowej lub w samym algorytmie uzyskane wyniki będzie można wyraźnie poprawić.

BIBLIOGRAFIA

[1] DERKACZ, M., GUMIECKI, R., MISZYN, L., OWIERSZENKO, M., CZABAN, M. : Wosprijatije rieczi w raspoznajuszczich modeliach, izd. Lwowskogo Uniw., Lwow, 1971.

[2] FANT, G. : Phonetics and Speech Research, w : Research Potentials in Voice Physiology, New York, 173-277, 1964.

[3] FRACKOWIAK-RICHTER, L. : The Duration of Polish Vowels, w : Speech Analysis and Synthesis, W.Jassem, ed. , vol. 3, 87-115, Warszawa, 1973.

[4] GRESSER, J.Y., MERCIER, G. : Automatic segmentation of speech into syllabic and phonemic units : application to French words and utterances, w : Auditory Analysis and Perception of Speech red. G.Fant, M.A.A.Tatham , Academic Press, London 1975.

[5] HAMMERBERG, R. : On redefining co-articulation, Journal of Phonetics, vol. 10, 123-137, 1982.

[6] HILL, D.R., WACKER, E.B. : ESOTerIC II - an approach to practical voice control : Progress Report 69, w : Machine Intelligence B.Meller and D.Michin, eds., Edinburgh Univ. Press, 463-493, 1970.

[7] HILL, R.H. : A basis for model building and learning in acoustic speech pattern discrimination, w : Machine Perception of Patterns and Pictures, Proc. Inst. Physics /Inst. El. Engineering/ Nat. Physical Lab. Conference NPL , Inst. of Physics, London, 151-160, 1972.

[8] JASSEM, W. : Fonetyczno-akustyczne założenia automatycznego rozpoznawania fonemów, Prace IPPT 14/1970, Warszawa, 1970.

[9] JASSEM, W. : Phonological segmental units in the speech signal, w : Form and Substance L.L.Hammerich, R.Jakobson and E.Zwirner, eds. , Akademisk Forlag, Copenhagen, 181-192, 1971.

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

[11] JASSEM, W. : Podręcznik wymowy angielskiej, wyd. 7, PWN, Warszawa, 1979.

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

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

[14] LEHISTE, I., PETERSON, G.E. : Studies of Syllabic Nuclei 2, Univ. of Michigan Speech Research Laboratory, Ann Arbor Report No. 4, 1960.

[15] LASS, R. : English Phonology and Phonological Theory, Cambridge Univ. Press, Cambridge, 1976.

[16] ŁOBACZ, P. : Processing and Decoding the Signal in Speech Perception, Prace IPPT 5/1981, Warszawa, 1981.

[17] MENZERATH, P., LACERDA, A. : Koartikulation, Steuerung und Lautabgrenzung, I.Dummler, Berlin, 1933.

[18] OHMAN, S. : On the Perception of Swedish Consonants in Intervocalic Position, The Royal Institute of Technology, Stockholm, 1962.

[19] PISONI, D.B. : On the Nature of Categorical Perception of Human Speech Sounds, Haskins Labs., Suppl. to Status Reports on Speech Research, Nov. 1971.

[20] POTTER, R.K., KOPP, G.A., GREEN, H.C. : Visible Speech, New York, 1947.

[21] REDDY, D.R. : Computer recognition of continuous speech, Journ. Acoust. Soc. Am. 42/2, 329-347, 1967 a.

[22] REDDY, D.R. : Phoneme grouping for speech recognition, Journ. Acoust. Soc. Am. 41/5, 1295-1300, 1967 b.

[23] STUDDERT-KENNEDY, M. : Perceiving phonetic segments, in : The Cognitive Representation of Speech, T.Myers, J.Laver and J.Anderson, eds., North-Holland Publ. Co., Amsterdam, 3-10, 1981.

[24] VAISSIERE, J. : Speech recognition programs as models of speech perception, w : The Cognitive Representation of Speech, /red. T.Myers, J.Laver, J.Anderson/, North-Holland, Amsterdam, 1981.

[25] WOLF, J.J., WOODS, W.A. : The WHIM Speech Understanding System, in : Trends in Speech Recognition, (W.A.Lea, ed.), Prentice Hall, Englewood Cliffs, New York, 316-339, 1980.

Tablica 1.

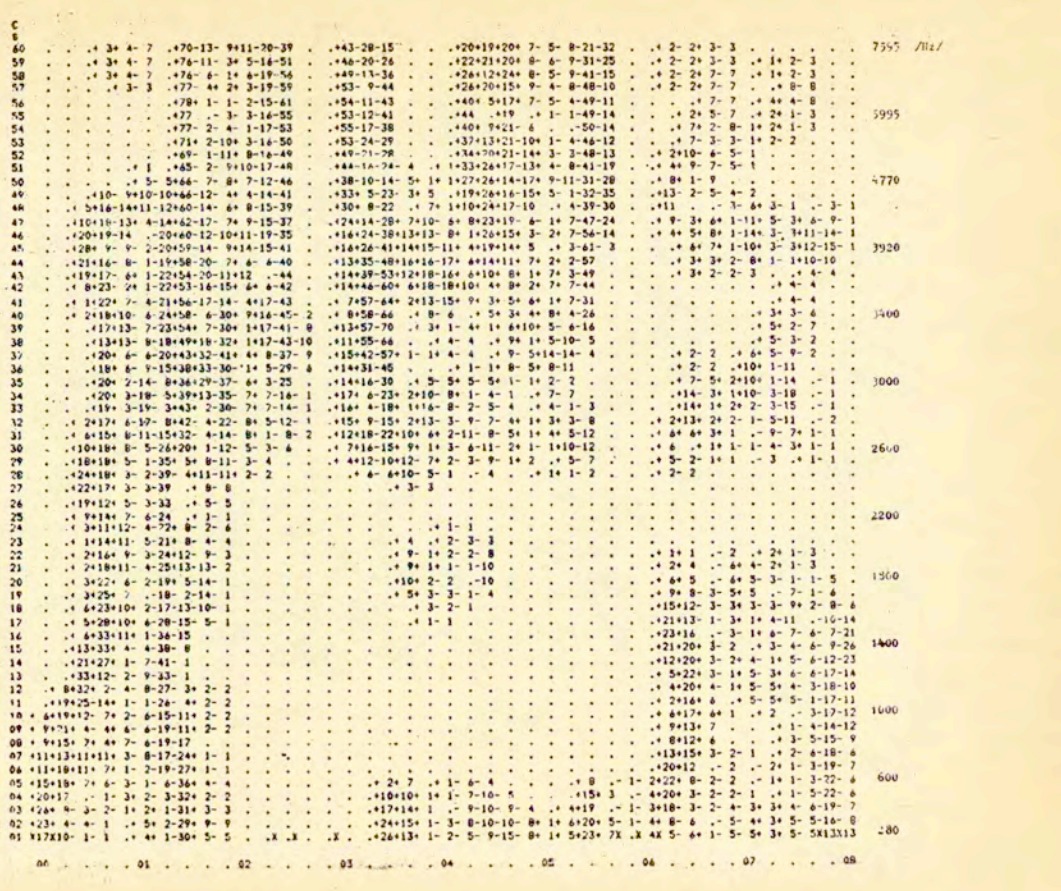
C/K/ Nr.kan	2	1	3		2	1	1	3		2	1	1	2		2	1	1	1	1	1	2
03	+	+	-		+	+	-	+		+	+	+	+		+	+	+	+	+	+	+
02	+	+	+		-	+	-	-		-	+	+	-		-	+	+	+	+	+	-
01	-	+	-		-	+	-	+		-	+	+	+		-	+	+	+	+	+	-
CEZ		A				A	A				B	B				C				C	

C/K/ Nr.kan	3	2	3		3	2	2	2	2	3		3	2	2	2	2	2	2	3	
03	+	+	-		+	+	+	+	+	-		-	+	+	+	-	-	-	+	
02	-	+	+		-	+	-	+	-	+		+	-	+	-	+	-	+	-	
01	+	-	-		+	-	-	-	-	-		-	-	-	-	+	+	+	+	
CEZ		D				G							F			G				

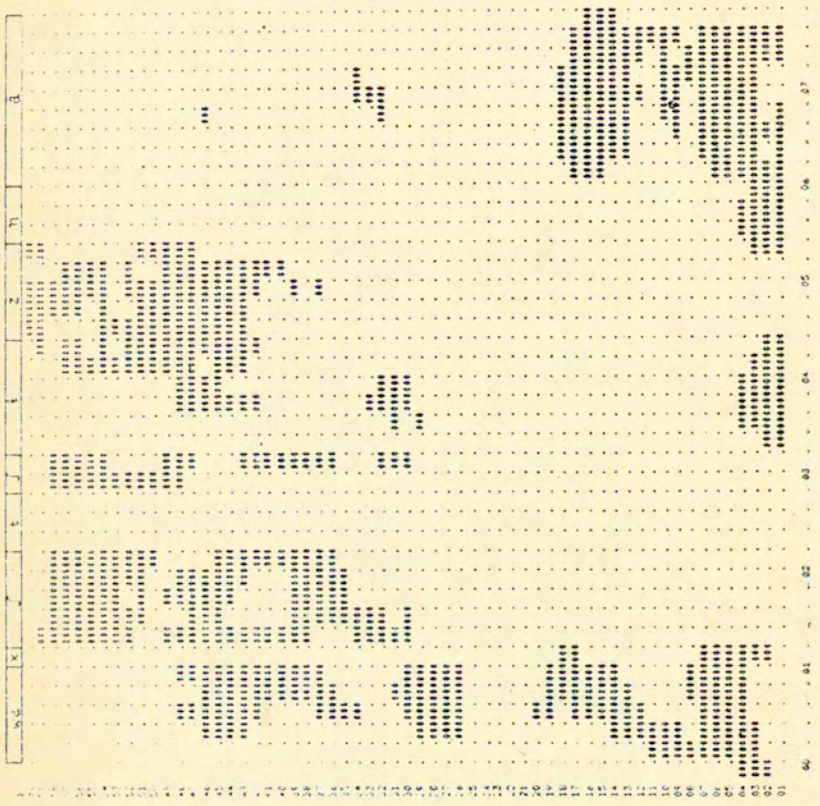
C/K/ Nr.kan	3	2	2	2	2	2	2	1	3		3	2	2	2	2	2	2	3	
03	+	+	+	+	-	-	-	-	+		+	+	-	+	-	+	-	-	
02	-	-	+	-	-	+	+	-	-		-	-	-	+	-	-	-	+	
01	+	-	-	-	+	+	+	-	+		+	-	+	-	+	-	+	-	
CEZ		F						A				F	F	F	F	F	D		

Typy cezur / przykłady oparte na trzech kanałach/.

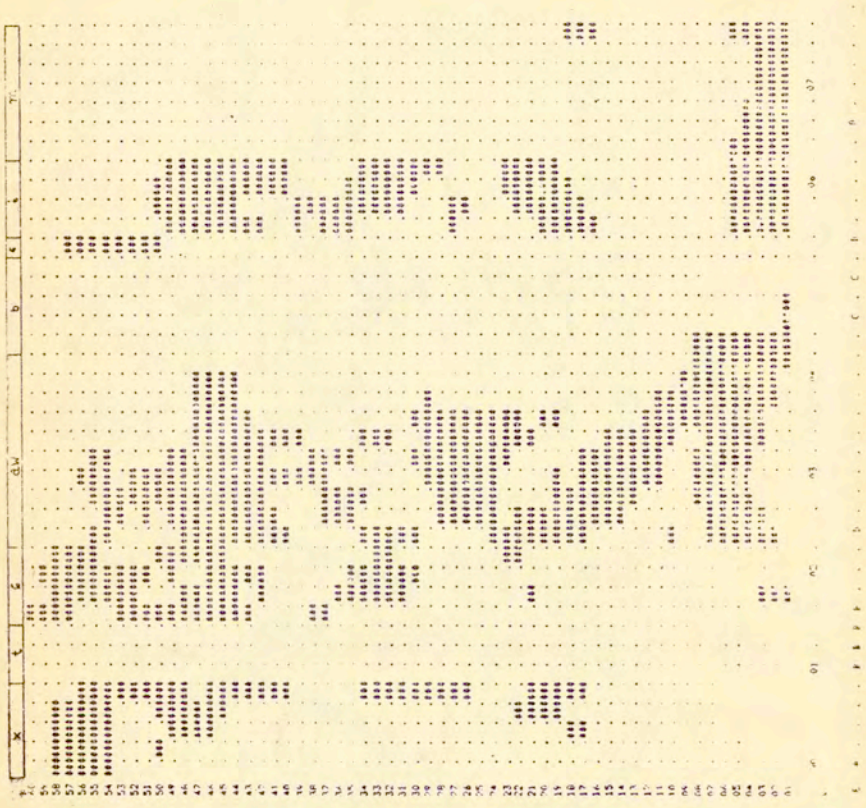
Ryc. 1. Spektrogram cyfrowy wyrazu "prawdziwy" /praf'dʒivɨ/



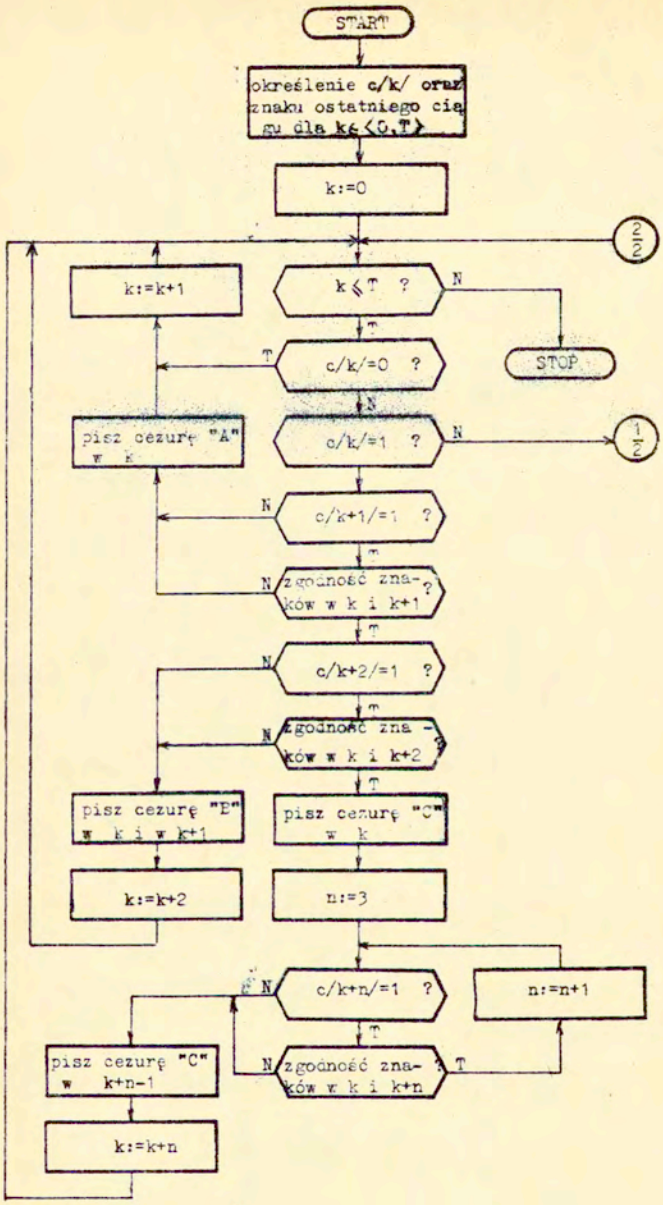
Rys. 2. Spektrogram różnicowy wyrazu "pijaszczyzna" /pwaʃ'tʃazna/.



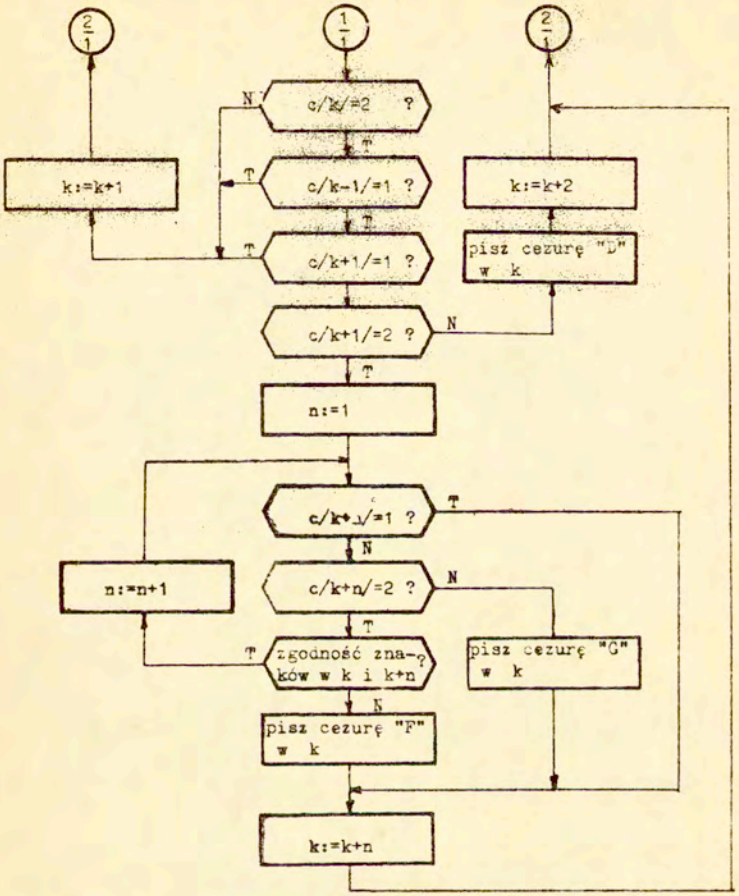
Ryc. 3. Spektrogram binarny wyrazu "piłeczczyzna" z segmentacją wizualną i automatyczną.



Ryc. 4. Spektrogram binarny wyrazu "chciałbym" /'xtɕɔmbim/ z segmentacją wizualną i automatyczną.



Ryc. 5. Cz. I. Algorytm automatycznej segmenta



k - numer kolumny w spektrogramie cyfrowym
T -dyskretny czas analizowanej frazy
c/k/ -liczba ciągów jednoznakowych w k-tej kolumnie
n -indeks pomocniczy

Ryc. 5. Cz. II. Algorytm automatycznej segmentacji.