Henryk Kubzdela

AUTOMATYCZNA EKSTRAKCJA WYBRANYCH CECH WIDMOWYCH MOWY

2/1994

WARSZAWA 1994

ISSN 0208-5658

Praca wpłynęła do Redakcji dnia 16 grudnia 1993 r.

Na prawach rękopisu

Instytut Podstawowych Problemów Techniki PAN
Nakład 100 egz. Ark.wyd. 1,0 Ark.druk.1,25
Oddano do drukarni w styczniu 1994 r.

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

Henryk Kubzdela
Zakład Fonetyki Akustycznej
IPPT PAN

AUTOMATYCZNA EKSTRAKCJA WYBRANYCH CECH WIDMOWYCH MOWY

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

W modelowaniu procesu dekodowania informacji zawartej w wypowiedzi operacja automatycznego rozsegmentowania sygnału mowy na podstawowe dźwięki fonetyczne okazuje się szczególnie trudną. Podstawę takiej segmentacji powinny stanowić cechy widmowe, które charakteryzują obrazy spektrograficzne różnych dźwięków mowy. Sygnał mowy posiada niestety właściwości, które utrudniają ekstrakcję tych cech. Utrudnienie stwarzają też niedoskonałości cyfrowej analizy akustycznej sygnału. Założono, że charakterystyczne cechy widmowe dźwięku zawierają się przede wszystkim w poziomie i częstotliwości najsilniejszej składowej oraz składowych przypadających w pobliżu punktów widma amplitudowego wskazanych przez położenia ekstremów funkcji (FMOW) modelującej kształt obwiedni widma. Dyskretny przebieg tej funkcji uzyskuje się poprzez dwustopniowe przekształcenie wygładzające widma amplitudowego. Na przykładach pokazano wyniki działania procedur wyznaczania przedziałów wypukłości i wklęsłości oraz położenia ekstremów FMOW. Zamieszczono przykłady wskazujące na segmentacyjne walory informacji zawartej w przebiegach czasowych poziomu widma w punktach wskazanych przez maksima FMOW.

1. Wprowadzenie.

W procesie percepcji informacji zakodowanej w mowie następuje szereg analiz. Każda z nich polega na identyfikacji pewnych elementów jednostkowych oraz układów, w jakich one występują. W analizie akustycznej elementami jednostkowymi są drgania proste występujące w sygnale mowy w nieskończenie wielu konfiguracjach amplitudowo-częstotliwościowych. Duża jest dynamika zmian tych konfiguracji. Ciągi niektórych z układów amplitudowo-częstotliwościowych drgań prostych są elementami wyższego rzędu odbieranymi przez słuchacza jako podstawowe dźwięki mowy. W ciągach tych elementów słuchacz identyfikuje jednostki lingwistyczne, którymi są słowa. Ciągi słów tworzą zdanie wymagające analizy gramatycznej dla zdekodowania elementu informacji wyższego rzędu. Podstawowym ogniwem wyliczonego tutaj łańcucha analiz jest operacja

rozsegmentowania sygnału mowy na ciąg elementów pozwalających poprawnie zidentyfikować podstawowe dźwięki mowy. Ten etap jest najtrudniejszy w całym procesie dekodowania informacji zawartej w wypowiedzi. Jego wejście stanowi ciąg wybranych cech akustycznego sygnału mowy a wyjście ciąg symboli zidentyfikowanych jednostek fonetyczno- akustycznych. Od wyboru właściwych cech akustycznych sygnału mowy uzależniona jest efektywność a także złożoność tego etapu. W poniższej pracy przedstawiono ekstrakcję cech widmowych, które dominują na obrazach spektrograficznych dźwięków mowy. Cechy te rozpatruje się z zamiarem wykorzystania ich w identyfikacji głosek w mowie ciągłej.

2. Właściwości sygnału mowy utrudniające ekstrakcję jego cech widmowych.

Dźwięk mowy będący zawsze drganiem złożonym, jeśli jest periodyczny, składa się z szeregu harmonicznych zmodulowanych w częstotliwości i amplitudzie przez funkcje wolnozmienne. Funkcje modulujące amplitudowo poszczególne harmoniczne są w złożony sposób zależne od częstotliwości tych składowych, które synchronicznie zmieniają się z częstotliwością podstawową, także modulowaną w trakcie mówienia, oscylującą wokół wartości różnej dla poszczególnych głosów. Zależność amplitudy składowych od częstotliwości jest czynnikiem pozwalającym poklasyfikować dźwięki mowy, zaś periodyczny przeważnie charakter sygnału mowy decyduje, że informacja o tych dźwiękach ma charakter dyskretny. Podstawę akustycznej oceny sygnału mowy stanowi porównanie przebiegów poziomu poszczególnych jego składowych w obrębie różnych dźwięków tworzących mowę ciągłą. Wartości odczytane z tych przebiegów dla jednego momentu czasu i przedstawione z podaniem częstotliwości składowych, których poszczególne przebiegi dotyczą, tworzą widmo chwilowe. Można je określić jako rozkład amplitud składowych dźwięku. W tym rozkładzie zawierają się cechy charakteryzujące dźwięk. Dzielą się one najogólniej na główne i drugorzędne. Te pierwsze decydują o przynależności dźwięku mowy do klasy fonematycznej drugie zaś wyrażają indywidualne zabarwienie dźwięków nie wpływające na przynależność do klasy. Dla identyfikacji fonemów w mowie ciągłej służą przede wszystkim cechy główne. Nie wszystkie one pozwalają się precyzyjnie zdefiniować

Funkcję determinującą parametry amplitudowo-fazowe składowych sygnału mowy aproksymuje się za pomocą ilorazu wielomianów. Moduł tego ilorazu jest funkcją częstotliwości określającą amplitudy poszczególnych składowych. Istnieją różne metody charakteryzowania tej funkcji na użytek klasyfikacji dźwięków mowy. Charakteryzacja może być ogólna lub dokładna, może być przeprowadzona wprost lub pośrednio. Na szczególną uwagę zasługują te metody charakteryzacji funkcji widma amplitudowego, które odwołują się do wiedzy o roli różnych szczegółów widmowych dla identyfikacji poszczególnych dźwięków mowy przez człowieka. Tylko nieliczne z istniejącej wielości tych szczegółów są ważne dla zmysłu słuchu w procesie percepcji mowy. Przy charakteryzowaniu funkcji widma amplitudowego one właśnie powinny zostać uwzględnione na pierwszym miejscu. Wiedza nasza o wadze różnych cech widmowych w percepcji dźwięków mowy jest niepełna. Jej uzupełnienie a może nawet głębsze zweryfikowanie wymaga stosownych badań. Wytłumaczenia wymaga na przykład istnienie bliskiego podobieństwa obrazów widmowych dźwięków mowy należących do różnych nieraz wzajemnie odległych klas jak również znaczna odmienność obrazów widmowych różnych realizacji jednego dźwięku mowy.

3. Widmo cyfrowe sygnału mowy i jego cechy będące przedmiotem ekstrakcji.

Poniższą pracę poświęcono ekstrakcji cech widmowych, które charakteryzują obrazy spektrograficzne różnych dźwięków mowy. Widmo amplitudowe dźwięków mowy cechują lokalne uniesienia i zahamowania spadku lub wzrostu poziomu. Cechy te określa się przede wszystkim przez podanie miejsc, w których one występują. Na ich podstawie nie zawsze możliwe jest odróżnienie dwóch dźwięków należących do różnych klas. Konieczne jest uzupełnienie tej charakterystyki przez porównanie poziomu widma w tych miejscach.
Do ekstrakcji oraz wizualizacji rozpatrywanych cech widmowych opracowano w języku C kilka nowych procedur, które dołączono do wcześniej zbudowanego [3], a ostatnio zmodyfikowanego, podstawowego systemu spektrografii oraz programowej filtracji czasowej sygnału mowy.

Cechy widmowe ekstrahowano na podstawie danych uzyskanych w wyniku 240-punktowego dyskretnego przekształcenia Fouriera

według algorytmu Winograda [1], [2], [3], z krokiem wynoszącym 120 próbek. Analizowany sygnał mowy był spróbkowany z częstotliwością 8 kHz. Rozdzielczość częstotliwości przy takich parametrach wynosi 33 Hz, czyli około trzykrotnie mniej niż wysokość niskiego głosu. Dyskretna transformacja Fouriera sygnału akustycznego jest mniej precyzyjna od wąskopasmowej analizy analogowej. O ile w tej ostatniej mierzona jest amplituda składowej przypadającej w danym paśmie o tyle w pierwszej następuje obliczenie amplitudy hipotetycznej składowej o częstotliwości odpowiadającej rozpatrywanemu punktowi widma. Obie te analizy spektralne nie dostarczają dokładnych danych o częstotliwości drgań składowych sygnału. Wielkość błędu, jakim obarczone jest widmo cyfrowe zależy też od charakteru sygnału oraz pewnych czynników losowych. W przypadku sygnału szumowego wynik cyfrowej transformacji Fouriera powinien być dokładniejszy niż w przypadku drgania periodycznego. Losowym czynnikiem wpływającym na dokładność wyniku cyfrowej transformacji Fouriera jest przypadkowość wzajemnego położenia okresu składowej podstawowej sygnału oraz okna selekcjonującego próbki do analizy.

Gdy częstotliwość podstawowa periodycznego drgania złożonego jest większa od rozdzielczości, z jaką wykonana została dyskretna analiza Fouriera tego drgania, wówczas oprócz linii widmowych w punktach odpowiadających pasmom, w których przypadają składowe harmoniczne pojawią się dodatkowe linie widmowe z obwiednią o znacznej nieregularności. Brak takich błędów zdarza się tylko w przypadku idealnego nałożenia się prostokątnego okna selekcjonującego próbki do obliczenia widma i okresu podstawowego analizowanego drgania. Ponieważ zgodność taka zachodzi niezwykle rzadko wynik analizy jest przeważnie obarczony błędem. Poszczególne linie widma cyfrowego są więc jedynie przybliżonym obrazem faktycznego składu zanalizowanego sygnału. Z tego faktu należy zdawać sobie sprawę przy charakteryzowaniu dźwięków mowy na podstawie cyfrowego widma z analizy Fouriera. Charakteryzacja taka powinna polegać na ocenie widma w punktach, w których informacja nie jest zdominowana przez artefakty.

Jako pierwszą cechę widma amplitudowego rozpatrywanego dźwięku mowy uznano poziom oraz częstotliwość najsilniejszej składowej. Poziom tej składowej przyjęto jednocześnie jako

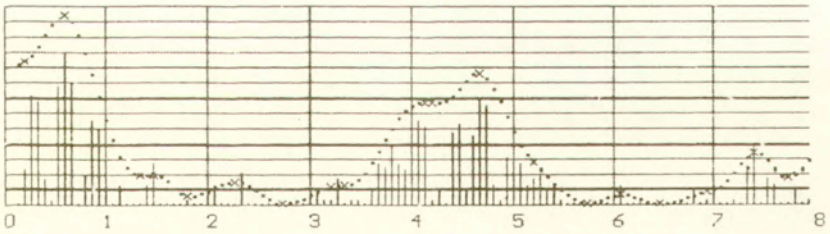
odniesienie dla oceny poziomu składowych przypadających w osobliwych punktach widma. Osobliwe punkty widma amplitudowego wskazane zostają przez położenia ekstremów funkcji modelującej kształt jego obwiedni. Funkcja ta oznaczana będzie niżej skrótem FMOW. Jej przebieg uzyskuje się poprzez dwustopniowe przekształcenie wygładzające widma amplitudowego. Przekształcenie pierwszego stopnia polega na zweryfikowaniu wartości w_i widma z dyskretnej analizy Fouriera w poszczególnych jego punktach zgodnie z zasadą :

$$w_{Zi} = MSWOPI \ , \quad \text{jeśli} \quad w_{Zi} = MSWOPI,$$
$$w_{Zi} = w_i \quad \text{w przeciwnym razie,}$$

gdzie $MSWOPI = \max(k_1 w_{i-1}, k_1 w_{i+1}, k_2 w_{i-2}, k_2 w_{i+2})$,
k_1, k_2 współczynniki normujące echo składowej widma przypadającej w punkcie i.
Wynik tego przekształcenia poddany zostaje następnemu, które polega na wyliczeniu nowych wartości w_{pi} w poszczególnych punktach widma zgodnie ze wzorem :

$$w_{pi} = \sum_{j=-3}^{+3} g_j w_{Z(i-j)} \ ,$$

gdzie g_j jest współczynnikiem określonym przez wartości dyskretnej funkcji okna, będącej zmodyfikowaną wersją funkcji zastosowanej przez autora między innymi w pracy [4].



Rys.1. Widmo amplitudowe z dyskretnej transformacji Fouriera według algorytmu Winograda (ciąg równoległych odcinków pionowych nazywanych liniami widmowymi) oraz przebieg wyznaczonej dla tego widma funkcji modelującej jego obwiednię (ciąg punktów).

Na rys.1 przedstawiono widmo amplitudowe uzyskane w wyniku dyskretnej transformacji Fouriera według algorytmu Winograda (ciąg równoległych odcinków pionowych nazywanych liniami widmowymi) oraz wyznaczon dla tego widma przebieg funkcji modelującej jego obwiednię (ciąg punktów).

Dla przebiegu tej funkcji znalezione zostają miejsca, w których występują jej ekstrema. Dla punktów widma, w których ekstremum funkcji modulującej obwiednię ma charakter maksimum określa się, w jakim przedziale poziomów mieści się odstęp wartości widma od wartości p_{max} najsilniejszej składowej w widmie. Na podstawie wyników pracy [5] przyjęto wstępnie trzy następujące przedziały poziomów :

I. 0 dB - -15 dB ,

II. -16 dB - -30 dB ,

III. -31 dB - $-(p_{max} - 5)$ dB .

Dla dźwięków mowy o słabszym natężeniu p_{max} może być mniejsze od 30 dB lub nawet od 15 dB. W pierwszym przypadku trzeci przedział poziomów staje się nieaktualny a jego dolną granicę przejmuje drugi przedział. W drugim przypadku nieaktualne stają się przedziały drugi i trzeci a pierwszy przejmuje dolną granicę od trzeciego.

Charakter funkcji modelującej w przybliżeniu obwiednię widma amplitudowego oraz wartości tego widma w osobliwych punktach przebiegu tej funkcji rozpatruje się w pięciu pasmach częstotliwości. Pasma te ustalono eksperymentalnie na podstawie warunku wykluczającego obecność w jednym paśmie dwóch formantów samogłosek polskich w ich stadium ustalonym. Mogą zachodzić przypadki, w których warunek ten nie będzie spełniony. Zdarza się bowiem, że w widmie danego dźwięku oprócz spodziewanych formantów głównych istnieją formanty dodatkowe nie istniejące na klasycznym obrazie widmowym tego dźwięku mowy. Zakresy wspomnianych pasm wyrażone numerami punktów widma pochodzącego ze 240-punktowej transformacji Fouriera są następujące :

I. 3 - 14,

II. 15 - 30,

III. 31 - 58,

IV. 59 - 92,

V. 93 - 119.

Jednemu punktowi widma odpowiada w tym przypadku pasmo o szerokości 33 Hz.

Zarówno zakresy wymienionych pasm jak i granice wcześniej omówionych przedziałów poziomów będą weryfikowane aż do uzyskania najkorzystniejszego rozróżniania segmentów fonetycznych na podstawie cech widmowych rozpatrywanych w tej pracy.

Widmo amplitudowe dźwięku mowy zostaje w pierwszym kroku scharaktryzowane przez podanie, w których pasmach zostały stwierdzone maksima przebiegu funkcji modelującej w przybliżeniu kształt jego obwiedni oraz jaki poziom posiadają najsilniejsze linie widmowe położone najbliżej punktów maksimum. W każdym paśmie mogą mieć miejsce następujące przypadki :

1. Brak maksimum FMOW.
2. Istnieje jedno maksimum przebiegu FMOW. Poziom najsilniejszej składowej widma w punkcie tego ekstremum lub jego pobliżu odniesiony jest do poziomu najsilniejszej składowej w całym widmie.
3. Istnieje kilka maksimów przebiegu FMOW. Różnice poziomów w punktach widma wskazanych przez położenia tych maksimów nie przekraczają pewnej wartości progowej a wklęsłości między tymi maksimami są płytkie.
4. Występują cechy inne niż podane w punktach 1, 2, 3.

Cechami uzupełniającymi charakteryzację widma amplitudowego na podstawie usytuowania maksimów funkcji modelującej jego obwiednię i poziomu składowych położonych najbliżej miejsc tych maksimów mogą być także zakresy wydatnej wypukłości i wklęsłości oraz położenia minimów wspomnianej funkcji. Duża asymetria zakresu wypukłości względem osi poprowadzonej przez punkt maksimum wskazuje na ślad słabego formantu zgubionego na zboczu bliskiego formantu silniejszego. Blisko siebie położone wpierw maksimum a potem minimum świadczą, iż maksimum jest prawostronnie słabo wypukłe a minimum lewostronnie słabo wklęsłe. Podobnie w przypadku blisko siebie położonych ekstremów w kolejności odwrotnej. Pierwsze ekstremum jest prawostronnie płytkim minimum a drugie lewostronnie słabo wydatnym maksimum przebiegu FMOW.

Poziom widma amplitudowego w punkcie minimum stanowi również jego cechę przydatną przy identyfikacji dźwięku, którego

rozpatrywane widmo dotyczy. Dwa bezpośrednio po sobie następujące wzajemnie odległe minima w punktach, w których widmo ma niski poziom, świadczą o braku lub bardzo niskim poziomie składowych w paśmie wskazanym przez te punkty.

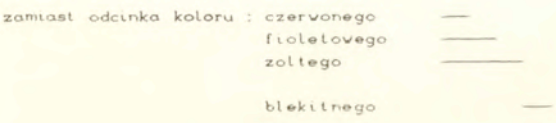
Omówione cechy pozwalają scharakteryzować dźwięk mowy. Identyfikuje się je przez analizę kształtu przebiegu funkcji modelującej obwiednię widma. Nie jest w tym celu potrzebna formuła matematyczna tej funkcji. Wystarcza znajomość ciągu jej wartości w punktach określonych przez dyskretną transformatę Fouriera.

Na podstawie zbioru opisanych wyżej cech widmowych można utworzyć wektor cech. Poszczególne grupy zmiennych tego wektora charakteryzowałyby widmo amplitudowe w rozpatrywanych pasmach. Pierwszymi zmiennymi w grupie byłyby numer kolejny oraz poziom najdłuższej linii widmowej usytuowanej w pobliżu punktu maksimum FMOW. Poziom mierzony jest względem poziomu najsilniejszej składowej w widmie. Nie uwzględnia się składowych słabszych od pewnego założonego progu. Kolejne zmienne są zarezerwowane na dane związane z ewentualnymi innymi maksimami w paśmie. Przewiduje się dodatkowe zmienne na informację o poziomie widma w punkcie minimum FMOW pomiędzy dwoma kolejnymi maksimami nie koniecznie przypadającymi w jednym paśmie.

4. Przykłady ekstrakcji cech widmowych.

Test ekstrakcji wyżej zdefiniowanych cech widmowych przeprowadzono na sygnale mowy pochodzącym z wypowiedzi logatomu *bimu* wymówionego trzykrotnie przez 3 głosy męskie. Użyto tego logatomu ze względu na cel testu, którym było zbadanie na ile różnią się pod względem rozpatrywanych w tej pracy cech dźwięki mowy podobne pod względem rozkładu energii akustycznej. Głoski występujące w logatomie *bimu* mają podobny rozkład energii w dolnym zakresie częstotliwości i są za wyjątkiem *i* oraz częściowo *b* dźwiękami niskimi. Użyto kilku form prezentacji wizualnej wyników ekstrakcji rozpatrywanych cech widmowych. Pierwszą podstawową formę stanowił obraz, który umownie nazwano konvex-kavgramem. Do utworzenia tej nazwy użyto międzynarodowych terminów matematycznych konvex i konkav oznaczających wypukłość i wklęsłość. Elementami tego obrazu są dwa krótkie odcinki o różnej długości i kolorze oraz szary

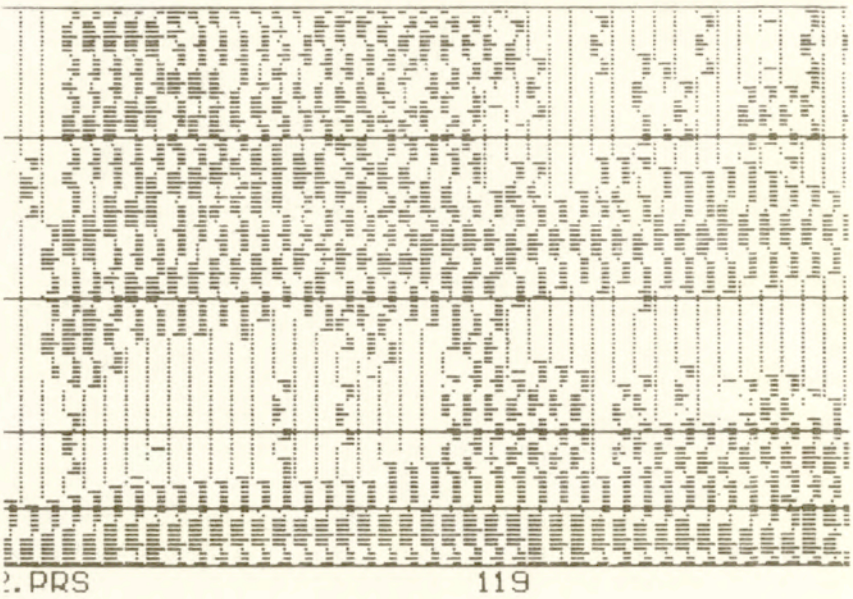
punkt. Dłuższy z tych odcinków wskazuje punkty widma, w których stwierdzone zostało ekstremum przebiegu FMOW. Punktowi odnoszącemu się do maksimum przypisany jest odcinek zabarwiony jednym z trzech kolorów zależnie od zakresu, w jakim przypada różnica między poziomem widma amplitudowego w tym punkcie a poziomem najsilniejszej składowej w widmie. Wyżej zdefiniowanym zakresom I, II, III przypisano odpowiednio kolory żółty fioletowy i czerwony. Punkt widma wskazany przez minimum przebiegu FMOW zaznaczony jest dłuższym odcinkiem koloru błękitnego. Krótszymi odcinkami zaznaczone są punkty widma pozaekstremalne. Zależnie od odstępu poziomu widma w danym punkcie od poziomu maksymalnego w całym widmie odcinek krótszy przyjmuje kolor żółty, fioletowy lub czerwony, jeśli FMOW jest w tym punkcie wypukła. Punkty, w których funkcja modelująca obwiednię widma jest wklęsła, oznaczono krótkimi odcinkami błękitnymi. Punkty widma, w których funkcja ta wykazuje przegięcie, zaznaczone są odcinkami szarymi lub białymi. Biel i szarość jasna oraz ciemna odnoszą się odpowiednio do trzech wyżej zdefiniowanych zakresów odstępu poziomów. Szarym punktem oznaczone są punkty widma, w których jego poziom jest pomijalnie niski. Zamieszczenie w tej pracy kolorowego konvex-kavgramu było ze względów wydawniczo-drukarskich niemożliwe. Zastąpiono go obrazem czarno-białym, na którym kolory zastąpiono przez długość odcinka a wklęsłość przebiegu FMOW zaznaczono przez odwrócenie początku, od którego liczy się wzrost długości odcinka. Zasadę takiej zamiany zilustrowano na rys.2.



Rys. 2. Ilustracja zasady zastąpienia koloru przez długość odcinka na konvex-kavgramie czarnobiałym.

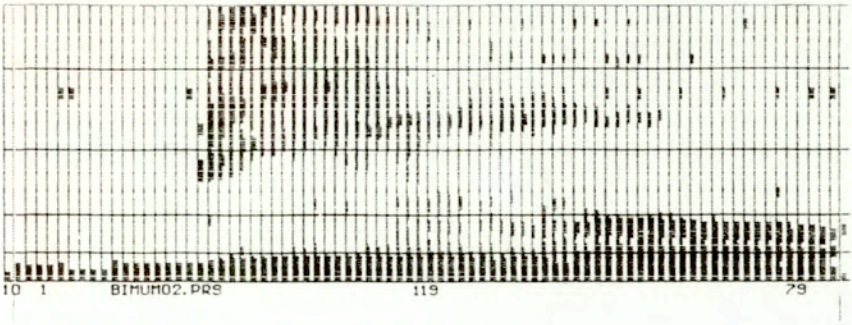
Względny poziom widma w zakresie przegięcia przebiegu FMOW wyrażony jest przez odległość kropki od wspólnej linii, na której oparte są odcinki znaczące zakresy wypukłości tego przebiegu.

Powiększony konvex-kavgram w wersji czarnobiałej fragmentu wypowiedzi logatomu *bimu* pokazano na rys. 3.

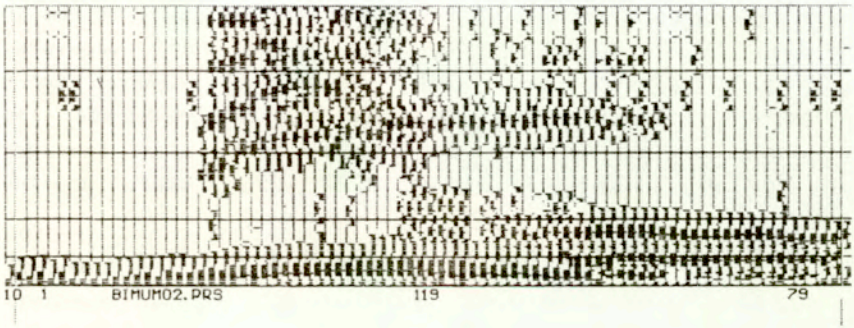


Rys. 3. Powiększony konvex-kavgram w wersji czarnobiałej dla fragmentu wypowiedzi logatomu *bimu*.

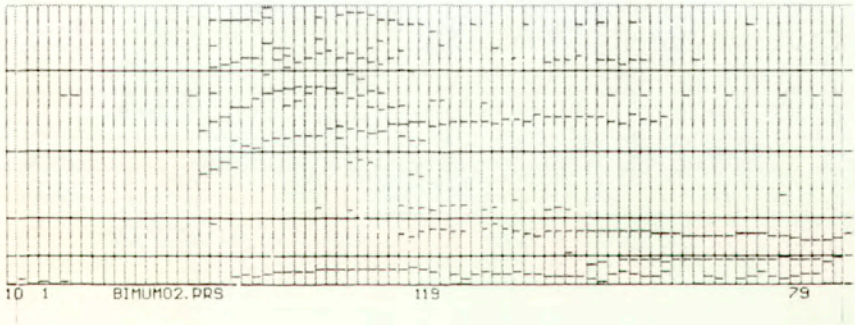
Z dwóch innych uproszczonych form wizualizacji cech widmowych pierwsza nazwana maksimogramem przedstawia przebiegi, w jakie układają się punkty widma odpowiadające miejscom wystąpienia maksimum przebiegu FMOW, druga jest trójpoziomowym spektrogramem ze skalą poziomu wyznaczoną przez 3 wyżej zdefiniowane zakresy poziomu. Na rys. 4 zamieszczono łącznie wszystkie trzy rodzaje obrazów pokazujących rozpatrywane cechy widmowe. Obraz z rysunku 4a ilustruje, do którego z trzech wyżej zdefiniowanych zakresów poziomu zaliczają się poszczególne punkty kolejnych widm badanej wypowiedzi logatomu *bimu*. Obraz na rys. 4b jest konvex-kavgramem uzyskanym dla tej samej wypowiedzi. Uwidacznia on zakresy wypukłości, wklęsłości oraz otoczenia punktów przegięcia przebiegu FMOW dla kolejnych



Rys.4a.



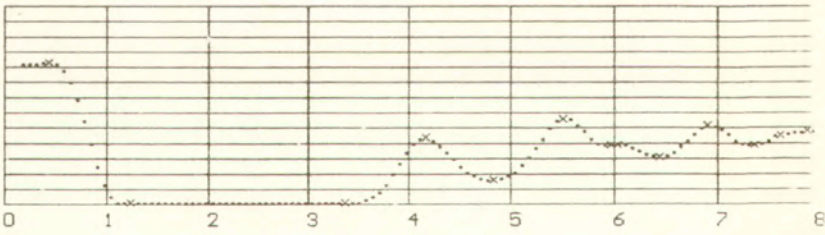
Rys.4b.



Rys.4c.

Rys.4. a) Uproszczony spektrogram z 3-stopniową skalą poziomu odniesioną do maksymalnego poziomu w poszczególnych widmach, b) konvex-kavgram, c) maksimogram. Wszystkie trzy obrazy dotyczą wypowiedzi logatomu *bimu* głosem męskim. Liniami poziomymi zaznaczone są zdefiniowane wyżej granice pasm.

widm. Punkty widma leżące poza zakresami wklęsłości przebiegu FMOW rozróżniane są dodatkowo zależnie od tego, do którego z trzech zakresów różnic poziomów zalicza się w nich poziom widma. Obraz na rysunku 4c przedstawia punkty widma, w których stwierdzone zostało maksimum przebiegu FMOW. Za pomocą długości odcinka wskazującego te punkty rozróżnia się je również według trzech zakresów różnic poziomów, z którymi konfrontuje się różnicę pomiędzy poziomem maksymalnym w widmie a poziomem w punktach wskazanych przez maksima przebiegu FMOW. Na omawianych rysunkach kierunek poziomy odpowiada osi czasu. Widoczne na nich sekcje pionowe odnoszą się do poszczególnych fram, dla których wyznaczano widma. Liczby 1 i 79 podają pozycje dwóch kursorów wskazujących framy. Umieszczona na początku liczba 10 oznacza, że dziesiątej framie sygnału analizowanej wypowiedzi nadano nr 1. Framę tę wskazuje lewy ze wspomnianych dwóch kursorów mających postać pionowego odcinka. Liczba 119 dotyczy położenia kursora wskazującego punkt widma. Kierunek przesuwu tego kursora jest pionowy i można nim wskazać dowolny punkt widma. Na rysunku jest to punkt ostatni w widmie.

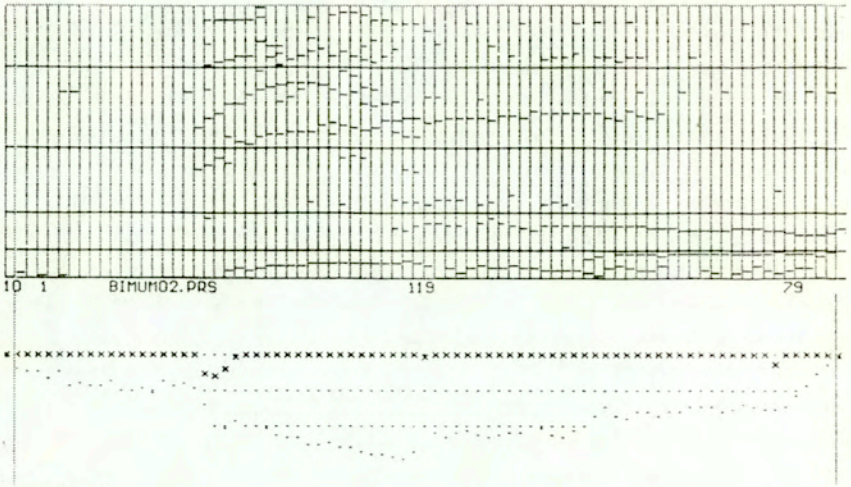


Rys.5. Przykład wykresu przebiegu funkcji modelującej obwiednię widma z zaznaczonymi za pomocą krzyżyków punktami ekstremalnymi.

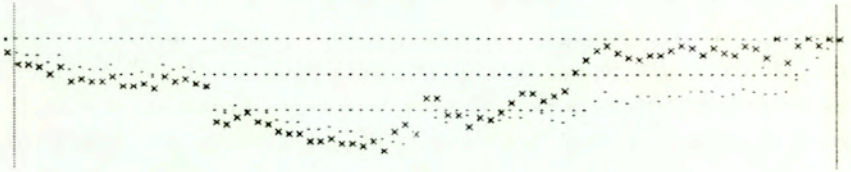
Wynik ekstrakcji rozpatrywanych cech dla jednego widma można skontrolować na obrazie przedstawiającym : a) widmo amplitudowe z transformacji Fouriera z zaznaczonymi przedziałami poziomów, stanowiącymi skalę, w jakiej określa się odstęp poziomu widma w poszczególnych punktach od poziomu najwyższego w widmie, b) wykres przebiegu FMOW z zaznaczonymi miejscami ekstremalnymi, oddzielnie lub łącznie z widmem amplitudowym z transformacji Fouriera.
Przykłady takich obrazów pokazano na rysunkach 1 i 5.

Uwagę budzą przebiegi czasowe różnic poziomu widma w punktach maksimum FMOW w pięciu wyżej zdefiniowanych pasmach względem poziomu najwyższego w widmie. Wykresy tych przebiegów oznaczonych skrótem PCRP ukazują się poniżej każdej z omówionych wyżej form wizualizacji wyników ekstrakcji cech widmowych. Kolor wykresu oznacza pasmo, do którego wykres się odnosi. Dzięki przypisaniu każdemu pasmu innego koloru możliwe jest wizualne porównanie poszczególnych PCRP. Na wykresy PCRP nałożone są linie oznaczające granice zakresów, w których rozpatruje się różnice poziomów. Jak wiadomo, niektóre z tych zakresów w pewnych przedziałach czasu ulegają skróceniu. Punkty znaczące koniec zakresu przestają wówczas układać się na linii prostej, lecz tworzą przebieg będący odwróceniem przebiegu wartości maksymalnej w kolejno rozpatrywanych widmach analizowanego dźwięku mowy.

Na rysunkach 6a(1-3) oraz 6b(1-3) pokazano wykresy PCRP dla pięciu pasm. Z tych samych względów co inne ilustracje zamieszczone w tej pracy rysunki 6a oraz 6b są jednobarwne. Dlatego poszczególne wykresy PCRP przedstawiono oddzielnie zamiast we wspólnym układzie współrzędnych. Dotyczą one jednej wypowiedzi logatomu *bimu* głosem męskim charakteryzującym się pod względem przebiegów czasowych rozpatrywanych różnic poziomów dla poszczególnych pasm mniejszym międzygłoskowym kontrastem niż u innych badanych głosów. Różnice pomiędzy czasowymi przebiegami różnic poziomów w poszczególnych pasmach odzwierciedlają segmentalny charakter mowy. Brak w danym paśmie maksimum przebiegu FMOW jest uwidoczniony przez położenie PCRP poniżej progu poziomu określonego przez : $-p_{max} + 5$ dB. Wartość tego progu należałoby w przyszłości bliżej sprecyzować osobno dla poszczególnych pasm oraz zależnie od tego, w którym paśmie przypada najsilniejsza linia widmowa.



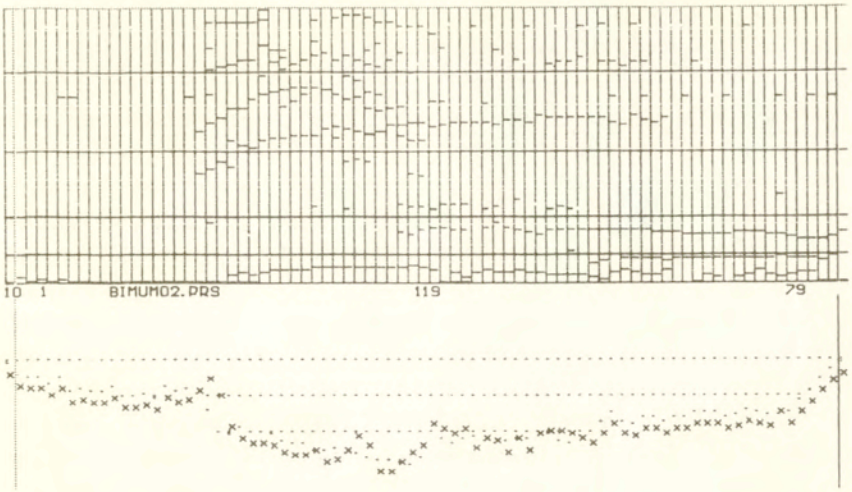
Rys.6a.1.



Rys.6a.2.

Rys.6a.3.

Rys.6a.1-3. Wykresy przebiegów czasowych różnic poziomu (PCRP) w punkcie maksimum FMOW i poziomu maksymalnego w widmie dla pięciu wstępnie przyjętych pasm i dla wypowiedzi logatomu *bimu* głosem męskim. Rysunek dotyczy pasm I,II,III licząc od góry. Powyżej wykresów przedstawiony jest maksimogram wskazujący w poszczególnych pasmach punkty widma, których dotyczą dane na wykresie. Granice pasm zaznaczone są poziomymi liniami ciągłymi. Przebiegi oznaczone są krzyżykami. Ciągi kropek oznaczają granice zakresów różnic poziomów.



Rys.6b.1.

Rys.6b.2.

Rys.6b.3.

Rys.6b.1-3. Wykresy przebiegów czasowych różnic poziomu (PCRP) w punkcie maksimum FMOW i poziomu maksymalnego w widmie dla pięciu wstępnie przyjętych pasm i dla wypowiedzi logatomu *bimu* głosem męskim. Rysunek dotyczy pasm III, IV, V licząc od góry. Powyżej wykresów przedstawiony jest maksimogram wskazujący w poszczególnych pasmach punkty widma, których dotyczą dane na wykresie. Granice pasm zaznaczone są poziomymi liniami ciągłymi. Przebiegi oznaczone są krzyżykami. Ciągi kropek oznaczają granice zakresów różnic poziomów.

5. Uwagi końcowe.

Wynikiem przedstawionego opracowania są procedury ekstrakcji cech widmowych dźwięków mowy. Umożliwią one poszerzenie bardzo aktualnych badań nad segmentacją mowy ciągłej. W pracy położono nacisk na parametr różnicy poziomów wyrażający pewne cechy widma. Zwykle w próbach rozpoznawania mowy większą uwagę zwraca się na parametry częstotliwościowe, w tym głównie częstotliwości formantów lub inne z nimi związane. Analityków dźwięków mowy frapują nie rzadkie przypadki niemal jednakowych częstotliwości formantów w dwóch różnych dźwiękach mowy. Przykładem mogą być głoski *l* oraz *i* w wypowiedzi logatomu *ili* lub i oraz *m* w rozpatrywanej tutaj wypowiedzi logatomu *bimu*. Cechę identyfikacyjną różnych dźwięków mowy posiadających zbliżone częstotliwości formantów stanowią zapewne różnice poziomów w charakterystycznych pasmach widma. To przekonanie stanowiło inspirację zajęcia się ekstrakcją cech widmowych wyrażonych przez różnice poziomów w charakterystycznych punktach widma świadomie nie utożsamianych z częstotliwościami formantów. W kolejnym zamierzonym przedsięwzięciu, którym będą próby identyfikacji segmentów mowy na podstawie między innymi parametrów różnic poziomów sugestie te poddane zostaną weryfikacji.

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

[1]. GROCHOLEWSKI, S., LUKASIK, E., *Układ do cyfrowego przetwarzania sygnałów mowy w czasie rzeczywistym*, Mat. II KK Przetwarzanie sygn. w telekom., sterow. i kontroli, BYDGOSZCZ 1986, s. 121 - 128.

[2]. GROCHOLEWSKI, S., OGÓRKIEWICZ, J., *Mikroprocesorowa realizacja WFTA do przetwarzania sygnału mowy w czasie rzeczywistym*, Mat. III KK Przetw. sygn. w telekom., sterow. i kontroli, Bydgoszcz 1988.

[3]. KUBZDELA, H., *System do badania cech widmowych oraz selekcji i odsłuchu segmentów sygnału mowy*, Prace IPPT 18/1993, Warszawa 1993.

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

[5]. KUBZDELA, H., OWSIANNY, M., *Wpływ poziomu formantów na percepcję syntetycznych dźwięków samogłoskowych*, Prace IPPT 41/1991, Warszawa 1991.

[6]. WINOGRAD, S., *On computing the discrete Fourier transform*, Math. Comp., vol. 32, Jan. 1978, s. 175-199.