P. Domagała, L. Richter

HIERARCHICZNA KLASYFIKACJA GŁOSEK JĘZYKA POLSKIEGO Z WYKORZYSTANIEM TECHNIK OPTYMALIZACJI PRZESTRZENI PARAMETRÓW

18/1995

WARSZAWA 1995

ISSN 0208-5658

Praca wpłynęła do Redakcji dnia 6 lutego 1995 r.

Na prawach rękopisu

Instytut Podstawowych Problemów Techniki PAN
Nakład 100 egz. Ark. wyd. 1,5 Ark. druk. 2,0
Oddano do drukarni w maju 1995 r.

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

Piotr Domagała
Lutosława Richter
Zakład Fonetyki Akustycznej
IPPT PAN

HIERARCHICZNA KLASYFIKACJA GŁOSEK JĘZYKA POLSKIEGO Z WYKORZYSTANIEM TECHNIK OPTYMALIZACJI PRZESTRZENI PARAMETRÓW

Streszczenie.

W oparciu o pewną kombinację liniową wartości funkcji autokorelacji sygnału przeprowadzono klasyfikację wszystkich głosek języka polskiego z wyjątkiem spółgłosek zwarto-trących. Materiał badawczy stanowiły samogłoski izolowane oraz logatomy, wymówione przez dwie osoby (głos męski i kobiecy). Klasyfikacja przebiegała wielostopniowo i polegała na przyporządkowaniu obszarów przestrzeni parametrów określonym populacjom poprzez wyznaczenie ciągów wektorów oraz wartości dyskryminacyjnych uzyskanych w wyniku rzutowania obiektów na kierunki zgodne z kierunkami wektorów własnych dla macierzy WITHIN i BETWEEN.

1. Wstęp.

Przedmiot pracy stanowi przeprowadzenie automatycznej klasyfikacji głosek języka polskiego dla celów automatycznego rozpoznawania mowy. Zastosowano hierarchiczną metodę klasyfikacji obiektów z p-wymiarowej przestrzeni parametrów w oparciu o analizę dyskryminacyjną. Praca stanowi ostatni etap szerzej zakrojonych badań, z których część pierwsza objęła klasyfikację głosek zwartych ([3],[7]), część druga klasyfikację głosek trących ([2],[6]). Opracowanie poświęcone spólgłoskom zwartym, niezależnie od uzyskanych wyników, spełniało rolę testu dla zastosowanej metody parametryzacji oraz klasyfikacji. Ponieważ grupa głosek zwartych napotyka na szczególne trudności w procesie automatycznego rozpoznawania mowy, wynikające ze specyfiki charakteryzujących je przebiegów akustycznych, uzyskanie

pozytywnych wyników rokowało pomyślnie o możliwości klasyfikacji pozostałych głosek.

Po zamknięciu etapu pierwszego (głoski zwarte) i drugiego (głoski trące) pozostały do zbadania w etapie trzecim samogłoski, spółgłoski nosowe oraz głoski /l/, /r/, /j/, /w/. W niniejszej pracy pominięto spółgłoski zwarto-trące, których segmenty składowe przy zastosowanej metodzie parametryzacji będą traktowane tak samo, jak głoski trące lub segmenty zwarcia w głoskach zwartych. Właściwa identyfikacja głosek zwarto-trących wymagać będzie wprowadzenia na wyższym poziomie dodatkowych reguł uwzględniajacych między innymi parametr czasu.

Uznano, iż nie ma potrzeby przeprowadzać klasyfikacji wyłącznie dla grupy głosek pozostałych do zbadania w trzecim etapie, gdyż po doświadczeniach przeprowadzonych z głoskami zwartymi i trącymi, również w tym przypadku można spodziewać się pomyślnych wyników dyskryminacji. Postanowiono więc przeprowadzić klasyfikację całościową, to znaczy wszystkich głosek języka polskiego (z pominięciem jedynie klasy zwarto-trących), co było ostatecznym celem autorów.

2.Wyniki klasyfikacji dla głosek zwartych i trących.

Ponieważ klasyfikacja głosek ma prowadzić w założeniu autorów do automatycznego rozpoznawania mowy niezależnie od kontekstu fonetycznego oraz od mówcy, w materiale doświadczalnym, składającym się z logatomów uwzględniono wszystkie alofony kontekstowe głosek zwartych ([3],[7]) oraz trących ([2],[6]). Łączna liczba logatomów wyniosła 155 dla głosek zwartych, 166 dla głosek trących. Dla celów analizy nagrano 20 głosów (10 męskich i 10 kobiecych) w przypadku zwartych oraz 10 głosów (5 męskich i 5 kobiecych) w przypadku trących. Dyskryminację zwartych ([3], [7]) przeprowadzono w oparciu o segment plozji. Dla głosów kobiecych uzyskano rozpoznawalność w granicach od 70% dla /p/ do 92% dla /ɟ/, średnio 80%, w głosach męskich od 68% dla /b/ do 89% dla /ɟ/ - średnio 79%, natomiast w głosach połączonych od 62% dla /p/ do 84% dla /ɟ/, co daje średnią całkowitą 76%. Zdecydowanie lepsze wyniki uzyskiwane dla głosek /ɟ/ oraz /c/ związane są

niewątpliwie z ich mniejszym zróżnicowaniem kontekstowym, wynikającym z dotyczących ich ograniczeń fonotaktycznych.

Nieco gorsze wyniki klasyfikacji uzyskano dla głosek trących ([2],[6]) - w głosach kobiecych od 46% dla /z/ do 81% dla /s/, średnio 60%, w głosach męskich od 57% dla /ʒ/ do 86% dla /v/, średnio 69%, w głosach połączonych od 36% dla /ʒ/ do 81% dla /v/, średnio 61%.

Uzyskane wyniki klasyfikacji można uznać za reprezentatywne ze względu na wysoką liczebność badanych populacji oraz uwzględnienie zróżnicowania kontekstowego i osobniczego. Należy podkreślić, iż zróżnicowania te, zwiększając rozproszenie wewnątrzpopulacyjne, w znacznym stopniu utrudniały proces klasyfikacji.

3. Materiał językowy.

W pracy niniejszej przyjęto nieco inne kryteria doboru materiału, niż na etapach poprzednich z tego względu, iż klasyfikacja miała objąć wszystkie głoski. Zachowanie dotychczasowych proporcji w zakresie liczebności materiału (np. w przypadku trących było to 18 183 obiektów dla głosów kobiecych i 16 172 obiektów dla głosów męskich) wymagałoby rozbudowania bazy danych do ogromnych rozmiarów, zwiększając w sposób nieefektywny nakład naniesionej pracy. Ponieważ próba klasyfikacji wszystkich głosek ma charakter pilotażowy, postanowiono ją przeprowadzić na dwóch głosach (jeden kobiecy i jeden męski) oraz materiale fonetycznym nie zróżnicowanym kontekstowo. Jedyny wyjątek uczyniono dla głoski /x/, która wyróżnia się szczególną podatnością na alofonizację kontekstową. Materiał językowy stanowiły samogłoski izolowane oraz logatomy o budowie VCV, w których element C reprezentowany był przez poszczególne spółgłoski dźwięczne i bezdźwięczne zwarte i trące, spółgłoski nosowe oraz /l/, /r/, /j/, /w/. W każdym przypadku obustronny kontekst samogłoskowy stanowiła głoska /e/. Szerszy kontekst, poprzez dodanie samogłosek skrajnych /i/, /u/ zapewniono jedynie głosce /x/. Materiał językowy był wprowadzany bezpośrednio do pamięci komputera IBM PC AT za pośrednictwem

karty wejścia głosowego, wchodzącej w skład systemu do analizy sygnału mowy VOLYZER. Całość materiału została powtórzona pięciokrotnie przez każdą z osób.

4. Parametryzacja sygnału mowy.

Dla sygnału mowy wprowadzanego do pamięci komputera zastosowano 13-bitowy przetwornik A/C i częstotliwość próbkowania wynoszącą 10 kHz. Sygnał, ograniczony do 5 kHz, został poddany preemfazie.

Sposób parametryzacji sygnału, opisany szczegółowo w pracach [1], [3], wykorzystuje pewną kombinację liniową wartości funkcji autokorelacji dla ograniczonych czasowo, dyskretnych fragmentów sygnału (fram). Przyjęta długość framy wynosiła 128 próbek (12,8 ms), przy czym krok, z jakim przesuwana była frama wzdłuż osi czasu, wynosił 64 próbki. Sygnał mowy po parametryzacji był zapamiętywany w postaci ciągu 12-elementowych wektorów, reprezentujących 12 wartości funkcji autokorelacji dla każdej kolejnej framy.

5. Baza danych.

Utworzono bazę danych obejmującą pliki tekstowe, po jednym dla każdej klasy głosek. W przypadku zwartych odrębne klasy głosek stanowiły segmenty plozji dla każdej z głosek oraz segment zwarcia dźwięcznego. Kolejne wiersze w pliku odnoszą się do kolejnych fram (mikrosegmentów), reprezentujących określoną klasę głosek. Dane do pliku wprowadzał specjalnie przygotowany "program narzędziowy", który umożliwiał wycinanie wybranych fragmentów sygnału - głosek lub ich segmentów składowych (zwarcie, plozja) w przypadku głosek zwartych, z rozbiciem na poszczególne, następujące po sobie mikrosegmenty. Segmentację przeprowadzano w oparciu o zwizualizowane wartości funkcji autokorelacji dla kolejnych fram (tzw. korelogramy), zsynchronizowane na ekranie z postacią czasową sygnału. Krzywe obrazujące rozkład wartości autokorelacji dla każdej framy (tzw. "sekcje") posiadają na ogół odmienną postać w obrębie kilku początkowych oraz końcowych

mikrosegmentów głoski w porównaniu z jej pozostałym fragmentem. W przypadku, gdy wartości funkcji autokorelacji dla mikrosegmentów reprezentujących stadia przejściowe odbiegały w widoczny sposób od rozkładów charakterystycznych dla danej głoski, nie uwzględniano ich w materiale uczącym. Każdy mikrosegment, wchodzący w skład wyciętego fragmentu, po wprowadzeniu do właściwego pliku reprezentowany był przez wektor, obejmujący 12 wartości funkcji autokorelacji oraz numery identyfikacyjne mikrosegmentu.

6. Założenia teoretyczne klasyfikacji głosek.

Teoretyczne podstawy klasyfikacji z wykorzystaniem techniki optymalizacji parametrów opracowane w oparciu o literaturę przedmiotu [4],[5] zostały szczegółowo przedstawione przez autorów w pracy [2]. W najogólniejszych zarysach przedstawiają się one następująco: Wektor zbudowany z p elementów wygodnie jest przedstawić jako punkt umieszczony w p-wymiarowej przestrzeni. Zbiór wielu takich punktów tworzy w przestrzeni skupienie, dla którego obszar ufności o zadanym poziomie procentowym stanowi p-wymiarową elipsoidę. Wyznaczenie kierunków jej osi pozwala na znalezienie nowych, wzajemnie nieskorelowanych cech, będących liniowymi kombinacjami cech pierwotnych. Oś związana z największą wariancją wyznacza kierunek najdłuższej osi elipsoidy, na którym rozrzut elementów jest największy. Jest on zgodny z kierunkiem wektora własnego macierzy kowariancji odpowiadającego maksymalnej wartości własnej. Wektory i wartości własne są wynikiem rozwiązania zagadnienia własnego

$$\Sigma\psi - \lambda\psi = 0$$

gdzie ψ jest wektorem własnym, a λ wartością własną. Jeżeli k p-wymiarowych populacji posiada równe macierze kowariancji i różne wektory średnie, to w wyniku transformacji do przestrzeni zmiennych dyskryminacyjnych elipsoidy ufności rozrzucone w przestrzeni parametrów pierwotnych zostaną przekształcone w kule rozmieszczone w przestrzeni zmiennych dyskryminacyjnych.

Rozwiązując uogólnione zagadnienie własne dla macierzy rozproszeń międzypopulacyjnych (BETWEEN) i macierzy rozproszeń wewnątrzpopulacyjnych (WITHIN) otrzymujemy wektor własny, wyznaczający optymalną oś dyskryminacyjną, na którą rzutowane obiekty należące do jednej populacji są możliwie najbardziej skupione, natomiast same populacje są możliwie najbardziej od siebie oddalone. Jest to konfiguracja najkorzystniejsza dla klasyfikacji obiektów, ponieważ spoglądamy na przestrzeń pod takim kątem, który umożliwia najlepszą dyskryminację populacji.

7. Metoda klasyfikacji.

W oparciu o powyższe podstawy teoretyczne przyjęto algorytm klasyfikacji oracowany przez Domagałę [1]. Musi on być poprzedzony etapem uczenia, który realizowany jest w następujących krokach:

1. Obliczenie macierzy BETWEEN i WITHIN.
2. Rozwiązanie uogólnionego zagadnienia własnego dla macierzy.
3. Zapamiętanie w tablicy wektora własnego odpowiadającego maksymalnej wartości własnej.
4. Dokonanie rzutowania wszystkich obiektów na kierunek zgodny z kierunkiem wektora własnego.
5. Wyznaczenie progowej wartości dyskryminacyjnej rozdzielającej analizowane populacje na dwa podzbiory populacji składowych (niekoniecznie rozłączne).
6. Powrót do punktu 1 i powtórzenie całej operacji dla każdego z uzyskanych podzbiorów.

Procedura kończy się wtedy, gdy wszystkie populacje zostaną rozgraniczone, lub gdy rzuty populacji wzajemnie się pokrywają. W przypadku częściowego pokrywania się populacji można przeprowadzić ich redukcję odrzucając w dalszej analizie te obiekty, które dają się sklasyfikować. Końcowym efektem przedstawionej procedury, nie zawsze możliwym do osiągnięcia, jest więc przydzielenie określonego obszaru w przestrzeni parametrów tylko jednej populacji.

Proces uczenia jest pracochłonny, ale jednorazowy. W jego wyniku otrzymujemy zbiór wektorów własnych, przyporządkowanych

kolejno otrzymywanym grupom populacji i odpowiadający im zbiór wartości dyskryminujących.

Proces klasyfikacji pojedynczego obiektu reprezentującego dowolny mikrosegment przebiegać będzie następująco:

1. Skalarne mnożenie wektora parametrów pierwotnych przez odpowiedni wektor własny uzyskany na etapie uczenia.
2. Porównanie wyniku mnożenia z odpowiadającą wektorowi własnemu wartością dyskryminacyjną i na tej podstawie przyporządkowanie obiektu do właściwej grupy populacji.
3. Dobranie kolejnego wektora własnego i powrót do punktu 1.

Procedura klasyfikacyjna ulega zakończeniu, gdy obiekt zostanie zakwalifikowany do nierozdzielnej populacji.

8. Realizacja procedury klasyfikacyjnej.

Na rys. 1 przedstawiono rzut wszystkich obiektów (mikrosegmentów) reprezentujących podzbiór / = l r w i u / (klasa nr 25) na kierunek optymalny (oś pionowa) dla równoczesnej dyskryminacji wszystkich sześciu populacji wchodzących w jej skład. Każdy obiekt reprezentowany jest przez punkt, którego położenie na osi poziomej, w ramach obszaru przynależnego dla danej populacji jest losowe. Pionowe odcinki, nałożone na rozproszone obiekty, odpowiadają odchyleniu standardowemu dla populacji. Dla prezentowanej klasy przyjęto wartość dyskryminacyjną równą 0,03, która pozwala przydzielić populację /ɨ/ (klasa 47) do obszaru przestrzeni leżącego poniżej, a populację / = ɨ u / (klasa 48) do obszaru leżącego powyżej tej wartości. Populacje / l r /, których obiekty znalazły się zarówno poniżej, jak i powyżej wartości granicznej, zostały zaliczone do obu nowopowstałych klas głosek, to jest do klasy 47 oraz 48. W niektórych przypadkach korzystne okazało się rzutowanie na płaszczyznę wyznaczoną przez dwa wektory własne, odpowiadające dwóm najwiekszym wartościom własnym. Wówczas podział przestrzeni nastąpił w oparciu o prostą.

W wyniku realizowania kolejnych kroków procedury następowała stopniowa redukcja liczby głosek wchodzących w skład podzbiorów. Rys. 2 przedstawia rzut obiektów należących do dwuelementowej

klasy 38, zawierającej populacje / a o /. Przyjęcie wartości progowej równej -0,01 tworzy dwie klasy rozłączne - / o / (klasa 67) oraz / a / (klasa 68), wyznaczając każdej z populacji określony, przynależny tylko jej obszar w przestrzeni parametrów. Nie wszystkie podzbiory klas, kończące procedurę, okazały się być jednoelementowe. Miało to miejsce wówczas, gdy rzuty obiektów należących do różnych populacji wzajemnie się pokrywały. Taki zbiór nierozdzielny może obejmować od dwóch populacji, np. / l r / (klasa 11), / s x / (klasa 200) do pięciu populacji , np. / s z ʃ ʒ x / (klasa 135), a w skrajnym przypadku siedem populacji / c ɟ s z ʃ ʒ x / (klasa 51). W skład takich nierozdzielnych klas wchodzą głoski o zbliżonych cechach akustycznych.

Fakt przynależności głoski do klasy nierozdzielnej nie wyklucza możliwości jej ostatecznego wyodrębnienia, o ile głoska znalazła się równocześnie w zbiorze o innej zawartości. Np. głoska / ʃ /, niezależnie od tego, że wchodzi w skład klas nierozdzielnych 135 i 51, stworzyła odrębną, samodzielną klasę na najniższym poziomie dzięki temu, że w wyniku podziału klasy 16 / t d c ɟ f s z ç ʑ ʃ ʒ x i / znalazła się nie tylko w klasie 29 / t d c ɟ f s z ʃ ʒ x /, której podział uległ częściowemu zablokowaniu, lecz również w klasie 30 / t d c ɟ f ç ʑ ʃ ʒ i /, z której udało się wydzielić każdą głokę składową.

Powstałe w procesie realizacji procedury klasyfikacyjnej drzewo, obejmujace 12 poziomów, na których dokonywano kolejnych podziałów, zostało przedstawione w tab. 1. Dla każdej z klas podano przyjętą dla niej wartość dyskryminującą populacje składowe, względnie dwie wartości, jeśli podziału przestrzeni dokonano za pomocą prostej. W kolumnie I znalazły się nowopowstałe klasy, utworzone z obiektów leżących poniżej wartości progowej, w kolumnie II klasy, utworzone z obiektów leżących powyżej wartości progowej. Niekiedy dochodzi do powtarzania się klas, to znaczy wyodrębnienia takich samych podzbiorów na różnych ścieżkach podziału. W takich przypadkach podano w tabeli, której klasie, wcześniej wyłonionej, odpowiada klasa powtarzająca się. Dalszy podział klasy kontynuowano w miejscu drzewa, gdzie pojawiła się ona po raz pierwszy.

Klasy uznane za nierozdzielne zostały wyróżnione w tabeli przez umieszczenie ich w pogrubionych ramkach, zaś klasy jednoelementowe zostały wyróżnione powiększoną czcionką.

Każda z głosek, zajmujac okreslone pozycje na poszczególnych poziomach drzewa klasyfikacyjnego, może być przedstawiona za pomocą właściwego tylko dla niej kodu, pozwalającego odróżnić ją od pozostałych głosek. Kod klasyfikacyjny może stanowić podstawę do identyfikacji głosek w procesie automatycznego rozpoznawania mowy. W tab. 2 zamieszczono kody uzyskane dla samogłosek. Znak "+" oznacza, iż w wyniku podziału klasy o podanym numerze, dana populacja znalazła się powyżej wartości progowej, za "-", że znalazła się poniżej tej wartości. W przypadku, gdy obiekty rozproszone są po obu stronach wartości progowej, głoska jest oznaczana zarówno symbolem "+", jak "-", a równocześnie pojawia się nowa, równoległa ścieżka podziału.

Najszybciej klasyfikacja przebiega dla samogłosek / a /, / o /, / u /. Każda z nich posiada po jednym kodzie, mieszczącym się w obrębie pięciu do siedmiu poziomów. Najbardziej skomplikowana sytuacja występuje w przypadku samogłoski / ɨ /. Charakteryzują ją aż cztery kody, według których może przebiegać klasyfikacja tej głoski.

9. Wyniki klasyfikacji mikrosegmentów.

Dla każdego obiektu, reprezentującego dowolny mikrosegment w materiale doświadczalnym, przeprowadzono klasyfikację do odpowiedniej populacji, w oparciu o wartości wektorów własnych i wartości dyskryminujące, uzyskane na etapie uczenia. Tab. 3 podaje procentowy udział rozpoznań poprawnych oraz błędnych dla wszystkich obiektów, reprezentujących wszystkie populacje. Suma wartości, odnoszących się do określonej głoski, przekracza często 100% (dotyczy to zwłaszcza zwartych i trących), ponieważ błędnie zaklasyfikowane obiekty lokowały się niekiedy w przestrzeni należącej do klasy nierozdzielnej, a wówczas program zaliczał je po kolei do każdej populacji składowej. Liczby umieszczone w pogrubionych ramkach oznaczają procentowo wyrażoną liczebność obiektów zaklasyfikowanych poprawnie.

Bardzo dobre wyniki uzyskano dla samogłosek oraz zwarcia dźwięcznego. Z wyjątkiem / ɨ / (86%), wykazują one blisko 100%-wą rozpoznawalność. Dla społgłosek zwartych można przyjąć rozpoznawalność w granicach 60 - 70%, z wyjątkiem /ɟ/ (47%), dla spółgłosek nosowych - 70%, dla trących w granicach od 70 do 80%, z wyjątkiem /v/, /ʒ/, które nie osiągnęły 50% rozpoznawalnoci.

Wyniki uzyskane w niniejszej pracy dla głosek zwartych są nieco gorsze od tych, które uzyskano w pierwszym etapie badań, gdzie materiał doświadczalny stanowiły wyłącznie zwarte. Z kolei dla trących wyniki zaprezentowane w tab. 3 są nieco lepsze od tych, które uzyskano dla materiału zawierajacego wyłącznie głoski trące. Rezultaty klasyfikacji dla poszczególnych grup głosek są jednak porównywalne, pomimo że na ostatnim etapie badań usunięto źródło zmienności w postaci zróżnicowanego kontekstu fonetycznego, a równocześnie zwiększono liczbę populacji głosek podlegających klasyfikacji.

Błędne klasyfikacje dotyczące głosek zwartych polegały najczęściej na zaliczeniu ich do populacji głosek trących i na odwrót - głosek trących do głosek zwartych. W dużym stopniu związane jest to z wyłonieniem kilku klas nierozdzielnych, których elementy składowe stanowiły równocześnie głoski zwarte i trące. Np. /c/ oraz /ɟ/ zostały między innymi zaklasyfikowane jako /s/, /z/, /ʃ/, /ʒ/, /x/, ponieważ należą do tej samej klasy nierozdzielnej /c ɟ s z ʃ ʒ x/. Niezależnie od przypadków tego rodzaju, błędne rozpoznania mikrosegmentów głosek zwartych związane były często z zaklasyfikowaniem ich jako określonej głoski trącej. Bardzo niski procent poprawnych rozpoznań dla /ɟ/ wiąże się z wysoką liczbą błędnych rozpoznań tej głoski jako /z/.

Spośród głosek trących najgorsze wyniki klasyfikacji stwierdzono dla /v/ oraz /ʒ/. Błędne rozpoznania /v/ polegały najczęściej na zaliczeniu reprezentujących tę głoskę obiektów do populacji /=/ (zwarcie dźwięczne), zaś błędne rozpoznania /ʒ/ na zaliczeniu obiektów do populacji /ç/ i /ʒ/ (akustycznie głoski najbardziej zbliżone do /ʒ/).

Duża liczba błędnych rozpoznań głosek /r/ i /l/ związana jest z jednej strony z obecnością klasy nierozdzielnej /r l/, z drugiej z częstym zaklasyfikowaniem obiektów /r/ do populacji

/v/, a obiektów /l/ do populacji /w/.

Przedstawiona metoda klasyfikacji głosek z wykorzystaniem technik optymalizacji przestrzeni parametrów okazuje się być w różnym stopniu skuteczna dla różnych typów głosek. W pełni zadowalające rezultaty uzyskano dla samogłosek. Otrzymanie porównywalnych wyników dla spółgłosek, przy optymalnym nakładzie kosztów (w sensie metodologicznym, mocy obliczeniowych i koniecznego sprzętu), jest prawdopodobnie niemożliwe. Konieczne byłoby filtrowanie uzyskanych wyników na wyższych poziomach - fonotaktycznym, syntaktycznym, semantycznym itd., poprzez konstruowanie funkcjonalnych modeli mechanizmów percepcji mowy, jakie stosowane są w psycholingwistyce.

10. Obserwacja przebiegu klasyfikacji głosek w obrębie dwóch poziomów klasyfikacyjnych.

Opracowano program, pozwalający śledzić przebieg klasyfikacji mikrosegmentów w obrębie dwóch poziomów klasyfikacyjnych. Stanowi on pierwszy krok w kierunku automatycznego rozpoznawania wyrazów. Działńie programu oparte jest na zasadzie kodowania mikrosegmantów w oparciu o podział klas na podzbiory znajdujące się poniżej lub powyżej wartości progowej (patrz tab.2). Program pozwala startować z dowolnej klasy i schodzi zawsze o dwa poziomy w dół. Wynik jego działania zilustrowano na rys.3 - wymówiony wyraz /ʃosa/ *szosa* i rys.4 - wymówiony wyraz /mul/ *mól*. W dolnej części rysunku znajduje się korelogram, będący zapisem 12 wartości funkcji autokorelacji sygnału dla kolejnych mikrosegmentów, w których określone zakresy wartości zostały zastąpione kolorami. Jak wspomniano w rozdziale 5, taka forma zapisu służyła do przeprowadzenia segmentacji w materiale doświadczalnym. Słupki umieszczone powyżej korelogramu odpowiadają średniej energii akustycznej sygnału dla każdego mikrosegmentu. W środkowej części rysunku znajdują się dwa paski, których odpowiednie fragmenty mają postać przerywaną lub ciągłą. Przedstawiają one wyniki klasyfikacji analizowanych obiektów do podzbiorów leżących poniżej (pasek przerywany) lub powyżej (pasek ciągły) wartości dyskryminującej, wyznaczonej dla danej klasy.

Te same paski, tyle że złączone, widoczne są w górnej części rysunku. W przypadku wyrazu /ʃosa/ *szosa* (rys.3), gdzie startowano od klasy nr 1, po pierwszym podziale wszystkie mikrosegmenty, reprezentujące głoski składowe, znalazły się w tym samym podzbiorze, leżącym powyżej wartości progowej. W wyniku drugiego podziału, mikrosegmenty reprezentujące głoskę /ʃ/ oraz /s/ znalazły się powyżej wartości dyskryminacyjnej, zaś mikrosegmenty reprezentujące głoski /o/ oraz /a/ - poniżej tej wartości. W przypadku wyrazu /mul/ *mól* (rys.4), gdzie startowano od klasy nr3, w wyniku pierwszego podziału mikrosegmenty składowe znalazły się w dwóch nowopowstałych klasach - te należące do /m/ w klasie zawierającej obiekty leżące poniżej wartości progowej, należące do /u/ oraz /l/ w klasie zawierającej obiekty leżące powyżej tej wartości. Po kolejnym podziale, mikrosegmenty reprezentujące wszystkie trzy głoski, znalazły się poniżej wartości dyskryminacyjnych wyznaczonych dla obu klas.

Wyniki klasyfikacji przedstawione w sposób zaprezentowany na rysunkach 3 i 4 pozwalają na bieżąco ocenić poprawność klasyfikacji w oparciu o dane ujęte w postaci drzewa klasyfikacyjnego. W obu zamieszczonych przykładach wszystkie głoski na każdym z poziomów zostały zaklasyfikowane poprawnie.

11. Uwagi końcowe.

Błędna klasyfikacja głosek może być w dużej mierze spowodowana wpływem kontekstu fonetycznego. Prześledzenie oddziaływania kontekstu umożliwia program, który wyniki klasyfikacji przedstawia w sposób zaprezentowany na rys. 3 i 4. Daje on możliwość oceny poprawności klasyfikacji przeprowadzanej dla dowolnej głoski w dowolnej klasie. Badania przeprowadzone w tym zakresie mogą wykazać konieczność uwzględnienia na etapie uczenia alofonów kontekstowych dla niektórych głosek.

Rozbudowanie programu, uwzględniające wszystkie poziomy drzewa klasyfikacyjnego, doprowadzi w ostatnim kroku do wyłonienia klas jednoelementowych, reprezentujących określoną głoskę lub jej alofon, a w końcowym efekcie do rozpoznania wyrazu.

Bibliografia.

[1] Domagała,P., 1991, Automatyczne segmentalne rozpoznawanie wyrazów polskich, rozprawa doktorska, IPPT PAN, Warszawa,

[2] Domagała,P., Richter,L., 1994, Automatyczna klasyfikacja spółgłosek trących języka polskiego na bazie optymalizacji przestrzeni parametrów, Prace IPPT 9/1994, Warszawa.

[3] Domagała,P. Richter,L., 1994, Discrimination of Polish Stop Consonants Based on Bursts Analysis, Archives of Acoustics, 19,2,147-159.

[4] Fukunaga,K., 1972, Introduction to Statistical Pattern Recognition, Academic Press, New York.

[5] Krzyśko,M., 1990, Analiza dyskryminacyjna, WNT, Warszawa.

[6] Richter,L., Domagała., 1993, Automatyczna klasyfikacja spółgłosek trących, Materiały XL Otwartego Seminarium z Akustyki, Rzeszów-Polańczyk, 325-328.

[7] Richter,L., Domagała,P., 1993, Discrimination of Polish Stop Consonants Based on Mapped Techniques, EUROSPEECH'93, Proceedings of 3rd European Conference on Speech Communication and Technology, Berlin, 1647-1650.

Tabela 1. Klasy populacji uzyskane w wyniku rzutowania obiektów. Kolumna I - klasy powstałe z obiektów leżących poniżej wartości progowej, II - klasy powstałe z obiektów leżących powyżej wartości progowej. Symbol "=" oznacza zwarcie dźwięczne.

Numer klasy	Populacje składowe	Wartość dyskryminująca	I		II	
			Nr. klasy	Populacje składowe	Nr. klasy	Populacje składowe
	Poziom 1					
1	pbtdcʝkgfv szçʒʃʒxmnʃ =lrwjiɨeao u	1	2	pbtdcʝkgfv szçʒʃʒxlrj iɨeao	3	bvzmnʃ=lrw jɨu
	Poziom 2					
2	pbtdcʝkgfv szçʒʃʒxlrj iɨeao	0,8	4	pbtdcʝkgfv szçʒʃʒxlrj iɨe	5	xlreao
3	bvzmnʃ=lrw jɨu	a=-1 b=2	6	bvz=lrwjɨu	7	vmnʃ=w
	Poziom 3					
4	pbtdcʝkgfv szçʒʃʒxlrj iɨe	0,2	8	ptdcʝkgfvs zçʒʃʒxji	9	pbkgvzxlrj ɨe
5	xlreao	0,2	10	xeao	11	lr
6	bvz=lrwjɨu	0,5	12	bvzrjɨ	13	bv=lrwjɨu
7	vmnʃ=w	a=-3,2 b=2	14	v	15	mnʃ=w

	Poziom 4					
8	ptdc↓kgfvs z¢z∫3xji	-0,3	16	tdc↓fsz¢z∫ 3xi	17	ptdkgvzxji
9	pbkgvzxlrj ↓e	0,3	18	pkgvzxj	19	pbkgvlrj↓e
10	xeao	-0,3	20	xao	21	xe
12	bvzrj↓	0,3	22	bvzrj	23	bvrj↓
13	bv=lrwj↓u	0,8	24	bv=lrwj↓	25	łrw↓u
15	mnſ=w	0 a=1,43 b=0,11	26 27	w =	28	mnſ
	Poziom 5					
16	tdc↓fsz¢z∫ 3xi	0,5	29	tdc↓fsz∫3x	30	tdc↓f¢z∫3i
17	ptdkgfvzxj i	0,65	31	ptdkgfvzxj	32	tji
18	pkgvzxj	0,5	33	pkgvzx	34	vzj
19	pbkgvlrj↓e	0,2	35	pbvlrj	36	pkgvlr↓e
20	xao	0,6	37	x	38	ao
21	xe	0,5	39	x	40	e
22	bvzrj	-0,17	41 =34	vzj	42	bvrj
23	bvrj↓	-0,25	43	vr↓	44 =42	bvrj
24	bv=lrwj↓	0,3	45	bvlrj	46	bv=lrw↓
25	=lrw↓u	0,03	47	lr↓	48	=lrwu
28	mnſ	0,02	49	mſ	50	nſ
	Poziom 6					
29	tdc↓fsz∫3x	a=1,71 b=-1,7	51	c↓sz∫3x	52	tdfsz∫3x
30	tdc↓f¢z∫3i	0,1	53	tdc↓f¢z∫3	54	d3i
31	ptdkgfvzxj	0,5	55	ptdkgfvzx	56	pdvzj

32	tji	0,1	57	ti	58	j
33	pkgvzx	-0,4	59	kgx	60a	pvzx
34	vzj	a=-1,1 b=-1,2	61	j	62	vz
35	pbvlrj	0	63	pvlr	64	pbvrj
36	pkgvlrɨe	-0,5	65	kglre	66	pvlrɨ
38	ao	-0,01	67	o	68	a
42	bvrj	0,2	71	bvr	72	bvj
43	vrɨ	0,4	73	vr	74	rɨ
45	bvlrj	0	77	bvlr	78 =42	bvrj
46	bv=lrwɨ	0,12	79	vlrwɨ	80	bv=w
47	lrɨ	-0,38	81	lɨ	82	lr
48	=lrwu	0,48	83	=lrw	84	wu
49	mɲ	-0,01	85	m	86	ɲ
50	nɲ	0,017	87	ɲ	88	n
	Poziom 7					
52	tdfszʃʒx	0,45	89	tdfsʃʒx	90	tszʃʒx
53	tdcɟfçzʃʒ	-0,05	90a	tcɟçzʃʒ	91	tdfçz
54	dɟi	0,6	92	dɟ	93	ɟi
55	ptdkgfvzx	-0,5	94	tkgx	95	ptdfvzx
56	pdvzj	0,55	96	pdvz	97	dvj
57	tɨ	0,5	98	t	99	i
59	kgx	0,2	100	x	101	kg
60a	pvzx	0,035	60	pvz	215	zx
60	pvz	-0,2	102	pv	103	pz
62	vz	0	104	v	105	z
63	pvlr	0,32	106	pvr	107	plr
64	pbvrj	0,3	108	pbvr	109	bj

65	kglre	0,25	110 =101	kg	111	lre
66	pvlrɨ	0,4	112 =63	pvlr	113	plrɨ
71	bvr	0,25	114 =75	bv	115	r
72	bvj	0,52	75	bv	76	j
73	vr	0,22	116	v	117	r
74	rɨ	0,42	74a	r	74b	ɨ
75	bv	0,05	118	v	119	b
77	bvlr	0,3	120 =71	bvr	121	lr
79	vlrwɨ	0,3	122 =43	vrɨ	123	vlrw
80	bv=w	-0,4	124	=w	125	bv=
81	lɨ	0,068	126	l	127	ɨ
83	=lrw	-0,5	128 =124	=w	129	lrw
84	wu	0,096	130	u	131	w
	Poziom 8					
89	tdfsʃ3x	-0,47	132	tsʃ3x	133	tdfsx
90	tszʃ3x	0,55	134	tszʃx	135	szʃ3x
90a	tcɟɕzʃ3	-0,525	136	cɟɕzʃ3	137	t
91	tdfɕz	0,875	138	tdf	139	tɕz
92	dɟ	-0,45	140	ɟ	141	d
93	ɟi	0,56	142	ɟ	143	i
94	tkgx	0,3	144	tx	145	tkg
95	ptdfvzx	-0,05	146	ptdfzx	147	ptdfv
96	pdvz	0,25	148	pdv	149	z

97	dvj	0,625	150	dv	151	j
101	kg	0,4	152	k	153	g
102	pv	0,17	154	p	155	v
103	pz	-0,5	156	z	157	p
106	pvr	0,22	158 =102	pv	159	pr
107	plr	0,5	160 =159	pr	161	lr
108	pbvr	a=0,33 b=0,17	162 =106	pvr	163 =71	bvr
109	bj	-0,45	164	j	165	b
111	lre	-0,2	166	e	167	lr
113	plrɨ	0,42	168 =159	pr	169 =47	lrɨ
123	vlrw	0,4	170 =73	vr	171 =129	lrw
124	=w	-0,049	172	w	173	=
125	bv=	1,03	174 =75	bv	175	=
129	lrw	-0,24	176	w	177	lr
	Poziom 9					
133	tdfsx	-0,4	178	tfsx	179	tdfx
134	tszʃx	0,3	180	tsʃx	181	szʃ
136	cɬɕzʃʒ	-0,05	182	ɕzʃ	183	cɬʃʒ
139	tɕz	-0,535	184	ɕz	185	t
145	tkg	0,5	188	t	189 =101	kg
146	ptdfzx	0,2	190	ptdfx	191	pzx
147	ptdfv	0,05	192	tdf	193	ptdv

148	pdv	-0,19	194	dv	195 =102	pv
159	pr	0,43	198	p	199	r
	Poziom 10					
178	tfsx	-0,45	200	sx	201	tfx
179	tdfx	-0,25	202	tfx	203	tdf
181	szʃ	-0,52	204	ʃ	205	sz
182	ɕzʃ	-0,33	206	ɕʃ	207 =184	ɕz
183	cɬʃ3	-0,066	208	cʃ3	209	ɬ3
184	ɕz	0,244	210	ɕ	211	ʑ
190	ptdfx	-0,13	212 =179	tdfx	213	ptd
191	pzx	a=-0,4 b=0,4	214	px	215	zx
193	ptdv	0	216	tdv	217	ptv
	Poziom 11					
206	ɕʃ	0,021	222	ɕ	223	ʃ
208	cʃ3	0,165	224	cʃ	225	ʃ3
209	ɬ3	0,65	226	ɬ	227	3
213	ptd	0	228	td	229	p
214	px	-0,25	230	x	231	p
215	zx	0,13	232	x	233	z
216	tdv	0,12	234	td	235	v
217	ptv	-0,1	234a	tv	235a =102	pv
	Poziom 12					
224	cʃ	-0,39	236	ʃ	237	c
225	ʃ3	0,343	238	ʃ	239	3

Tabela 2. Kody klasyfikacyjne dla samogłosek.

i		ɨ	
Numery klas	Kody	Numery klas	Kody
1	-	1	- +
2	-	2, 3	- -
4	-		/ \
	/ \	4, 6	+ - +
8	- +		/ \
16, 17	+ +	9, 12, 13	+ + - +
30, 32	+ -	19, 23, 24, 25	+ - + -
54, 57	+ +	36, 43, 46, 47	+ + - -
93	+	66, 74, 79, 81	+ + - +
		113, 122	+ +
		169 74	- +
		81	+

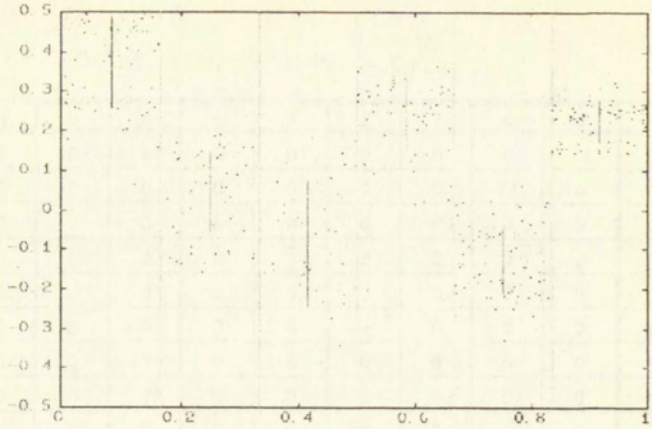
e		a		o		u	
1	-	1	-	1	-	1	+
	/ \	2	+	2	+	3	-
2	- +	5	-	5	-	6	+
4, 5	+ -	10	-	10	-	13	+
9, 10	+ +	20	-	20	+	25	+
19, 21	+ +	38	+	38	-	48	+
36	-					84	-
65	+						
111	-						

Tabela 3. Wyniki klasyfikacji mikrosegmentów w materiale doświadczalnym. Wartości podane w procentach.

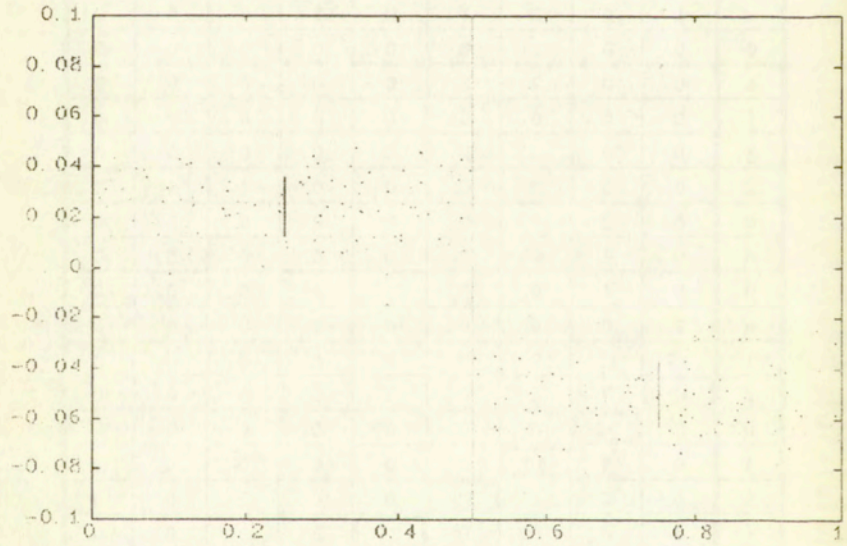
	p	b	t	d	c	ɟ	k	g	f	v	s
p	58	0	16	5	0	0	0	5	11	11	0
b	11	67	0	0	0	0	0	0	0	0	0
t	4	0	72	28	4	4	4	0	36	4	4
d	0	0	77	59	0	0	0	9	64	0	0
c	0	0	0	0	68	68	0	0	0	0	74
ɟ	0	0	0	0	29	47	0	0	0	0	29
k	0	0	8	0	0	0	65	12	0	0	4
g	0	5	5	5	0	0	10	62	0	0	4
f	2	0	83	52	0	0	0	0	67	2	11
v	5	1	21	16	0	0	3	0	5	46	0
s	1	0	25	7	24	24	0	1	10	4	79
z	1	2	6	0	16	16	0	0	1	4	58
ç	0	0	4	0	5	4	2	1	0	0	5
ʑ	0	0	5	1	2	1	0	8	0	0	2
ʃ	0	0	3	0	59	58	0	0	0	0	70
ʒ	0	0	3	3	38	40	0	0	2	0	59
x	0	0	31	3	27	27	3	0	22	1	35
m	0	0	0	0	0	0	0	0	0	6	0
n	0	0	0	0	0	0	0	0	0	0	0
ɲ	0	0	0	0	0	0	0	0	0	0	0
=	0	0	0	0	0	0	0	0	0	0	0
l	0	1	0	0	0	0	0	0	0	0	0
r	2	0	3	3	0	0	2	0	2	13	0
w	0	0	0	0	0	0	0	0	0	0	0
j	0	5	1	1	0	0	0	0	0	1	0
i	0	0	0	0	0	0	0	0	0	0	0
ɨ	1	0	0	0	0	0	0	0	0	3	0
e	0	0	2	0	0	0	0	0	2	0	0
a	0	0	0	0	0	0	0	0	0	0	0
o	0	0	0	0	0	0	0	0	0	0	0
u	0	0	0	0	0	0	0	0	0	0	0

	z	ç	ʑ	ʃ	ʒ	x	m	n	ɲ	=	l
p	0	0	0	0	0	5	0	0	0	0	0
b	0	0	0	0	0	0	0	0	0	11	0
t	8	0	4	4	4	20	0	0	0	0	0
d	0	0	0	0	0	9	0	0	0	0	5
c	74	13	0	74	81	74	0	0	0	0	0
ɟ	59	6	6	29	29	35	0	0	0	0	0
k	0	0	0	4	4	19	0	0	0	0	0
g	0	0	0	0	0	14	0	0	0	0	0
f	3	5	0	8	8	33	0	0	0	0	0
v	1	0	0	0	0	0	0	0	0	17	0
s	61	1	0	63	52	72	0	0	0	0	0
z	83	0	0	47	44	48	0	0	0	0	0
ç	4	70	8	12	6	5	0	0	0	0	0
ʑ	1	20	48	5	13	2	0	0	0	0	0
ʃ	67	7	4	83	64	66	0	0	0	0	0
ʒ	61	0	9	59	77	58	0	0	0	0	0
x	27	0	0	28	27	78	0	0	0	0	0
m	0	0	0	0	0	0	68	10	8	9	0
n	0	0	0	0	0	0	3	71	10	14	0
ɲ	0	0	0	0	0	0	16	12	69	0	0
=	0	0	0	0	0	0	0	1	1	96	2
l	0	0	0	0	0	0	0	0	0	1	48
r	0	0	0	0	0	7	0	0	0	2	32
w	0	0	0	0	0	0	1	0	2	1	6
j	6	0	0	0	0	0	0	0	0	0	0
i	0	0	0	0	0	0	0	0	0	0	0
ɨ	0	0	0	0	0	0	0	0	0	0	4
e	0	0	0	0	0	4	0	0	0	0	0
a	0	0	0	0	0	0	0	0	0	0	0
o	0	0	0	0	0	0	0	0	0	0	0
u	0	0	0	0	0	0	0	0	0	0	0

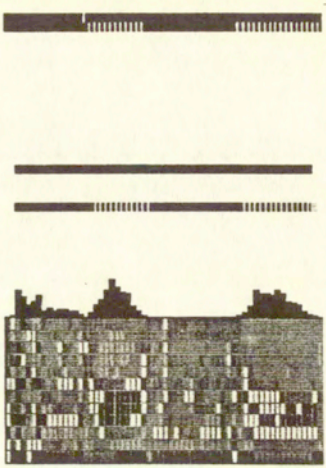
	r	w	j	i	ɨ	e	a	o	u
p	11	0	0	0	0	0	0	0	0
b	0	11	0	0	0	0	0	0	0
t	0	0	0	0	0	4	0	0	0
d	5	0	0	5	0	0	0	0	0
c	0	0	0	0	0	0	0	0	0
ɟ	0	0	0	6	0	0	0	0	0
k	0	0	0	0	0	0	0	0	0
g	0	0	0	0	0	0	5	0	0
f	0	0	0	0	0	0	0	0	0
v	0	5	3	0	0	0	0	0	0
s	0	0	0	0	0	0	0	0	0
z	1	0	2	1	0	0	0	0	0
ɕ	0	0	2	0	0	0	0	0	0
ʑ	0	0	2	1	0	0	0	0	0
ʃ	0	0	0	0	0	0	0	0	0
ʒ	0	0	3	2	0	0	0	0	0
x	0	0	0	0	0	0	4	4	0
m	0	0	0	0	0	0	0	0	0
n	0	2	0	0	0	0	0	0	0
ɲ	0	4	0	0	0	0	0	0	0
=	2	0	0	0	0	0	0	0	0
l	49	13	8	0	0	0	0	0	0
r	42	7	2	0	3	0	0	0	0
w	7	72	0	0	0	0	0	0	11
j	0	8	80	0	0	0	0	0	0
i	0	0	1	99	0	0	0	0	0
ɨ	4	0	0	0	86	3	0	0	0
e	0	0	0	0	0	96	0	0	0
a	0	0	0	0	0	0	100	0	0
o	0	0	0	0	0	1	0	99	0
u	0	5	0	0	0	0	0	0	95



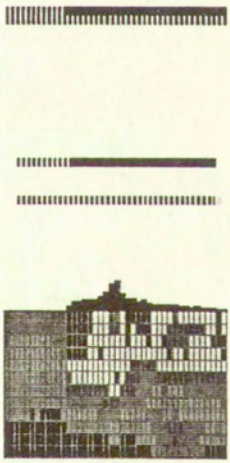
Rys. 1.Rzuty obiektów należących do klasy 25 /=lrwiu/



Rys.2. Rzuty obiektów należących do klasy 38 /ao/.



Rys.3. Wyraz /ʃosa/ - szosa.



Rys.4. Wyraz /mul/ - mól.