2.23 — akustyka mowy

H. Kubzdela, L. Richter

BADANIE EFEKTYWNOŚCI UKŁADU DO AUTOMATYCZNEGO ROZPOZNAWANIA MOWY

16/1990

WARSZAWA 1990

ISSN 0208-5658

Praca wpłynęła do Redakcji dnia 20 listopada 1989 r.

N a p r a w a c h r ę k o p i s u

Instytut Podstawowych Problemów Techniki PAN

Nakład 120 egz. Ark.wyd.1,6 Ark.druk. 2
Oddano do drukarni w kwietniu 1990 r.
Nr zamówienia 160/90

Warszawska Drukarnia Naukowa, Warszawa,
ul. Śniadeckich 8

Henryk Kubzdela
Lutosława Richter
Zakład Fonetyki Akustycznej
IPPT PAN

BADANIE EFEKTYWNOŚCI UKŁADU DO AUTOMATYCZNEGO ROZPOZNAWANIA MOWY OPARTEGO NA ANALIZATORZE WIDMA BINARNEGO W ZAKRESIE DYSTYNKTYWNOŚCI AKUSTYCZNO-FONETYCZNEJ DLA CELÓW STEROWANIA ZA POMOCĄ GŁOSU.[1)]

Streszczenie.

Wyniki uzyskiwane w zakresie globalnego rozpoznawania wyrazów wykazują związek z materiałem językowym tworzącym testowane wypowiedzi. Odgrywa tu rolę bliskość głosek wchodzących w skład zmagazynowanych wzorców. Przeprowadzono doświadczenie mające na celu stwierdzenie, w jakim stopniu podlegają błędnym identyfikacjom głoski, pomiędzy którymi zachodzi podobieństwo fonetyczno-akustyczne, wizualne lub percepcyjne. Materiał doświadczalny obejmował osiem zestawów logatomów, w obrębie których przeprowadzano porównania pomiędzy określonymi głoskami. Testowano trzy repetycje każdego logatomu wymówione przez jedną osobę. Rozpoznawanie przebiegało w oparciu o obrazy binarne wypowiedzi, wyznaczone przy zastosowaniu tzw. funkcji wagi widma, z wykorzystaniem czterech wariantów uzyskiwania obrazów. Każda z wypowiedzi traktowana była jeden raz jako obiekt rozpoznawany, a wielokrotnie jako wzorzec, do którego przyrównywano inne obiekty. Wyniki przybrały postać macierzy odległości pomiędzy obrazem binarnym rozpoznawanego obiektu a przyrównywanymi do niego obrazami wzorców. W oparciu o przyjęte wartości progowe dokonano oceny błędnych identyfikacji.Uzyskane dane na temat bliskości głosek mogą mieć zastosowanie przy sporządzaniu zbiorów haseł wykorzystywanych w układach rozpoznających.

Wstęp.

Prace z zakresu automatycznego rozpoznawania izolowanych wyrazów prowadzone w Zakładzie Fonetyki Akustycznej IPPT PAN koncentrowały się do tej pory głównie wokół zagadnień związanych z efektywnością stosowanej metody. Dotyczyły założeń ideowych modelu, rozwiązań technicznych oraz programowych, które podlegały kolejnym udoskonaleniom ze strony autora [1], [2], [3], [4]. Reprezentacja wypowiedzi w formie obrazów binarnych, na podstawie których odbywa się identyfikacja

---

1) Praca wykonana w ramach CPBP 02.13

wyrazów, przybierała różne warianty, zgodnie z dążeniem do optymalizacji metody.Testowaniu podlegały różne typy spektrogramów binarnych, to jest obrazów odzwierciedlających istotne cechy widmowe sygnału mowy. Pierwotnie wykorzystywane spektrogramy składające się z widm binarnych 63-parametrycznych zostały zastąpione przez spektrogramy złożone z widm 16-parametrycznych. Te ostatnie okazały się dostatecznie reprezentatywne dla izolowanych wypowiedzi wyrazów wymawianych przez jeden głos. Obecnie stosowana wersja reprezentacji sygnału mowy posługuje się obrazem binarnym utworzonym przy zastosowaniu tzw. funkcji wagi widma. Taka forma parametryzacji ignoruje w znacznym stopniu widmowe cechy osobnicze głosu, przez co w pewnym stopniu umożliwia posługiwanie się w rozpoznawaniu wspólnym zbiorem wzorców pochodzących od różnych osób.

Dotychczasowe dążenia, mające na celu zwiększenie poprawności rozpoznawania, skupiały się na opracowywaniu różnych wersji modelu. Należy jednak uwzględnić fakt, że przedmiotem rozpoznawania jest mowa, w związku z czym istotną rolę musi odgrywać materiał językowy, który ma również swój udział w ostatecznym kształcie uzyskanych wyników. Analiza błędów powstałych w trakcie identyfikacji umożliwi zapoznanie się z uwarunkowaniami lingwistycznymi oddziaływującymi na wyniki rozpoznawania. Okazjonalnie czynione obserwacje pozwalają zakładać, iż możliwość wystąpienia błędów dotyczy określonych grup głosek. Decyduje o tym podobieństwo fonetyczno-akustyczne segmentów, z których pochodzą porównywane fragmenty obrazów. Systematyczne zbadanie zależności pomiędzy strukturą fonetyczną zbioru wypowiedzi poddanych rozpoznawaniu a efektywnością układu, co jest przedmiotem niniejszej pracy, pozwoli określić stopień bliskości głosek w aspekcie automatycznego rozpoznawania wyrazów. Uzyskane wyniki będą miały praktyczne zastosowanie przy tworzeniu zbiorów haseł dla użytkowych modeli rozpoznajników wyrazów. Pozwolą na kontrolowanie struktury głoskowej wypowiedzi w takim kierunku, by w miarę możności nie dochodziło do błędnych identyfikacji, spowodowanych bliskością głosek wchodzących w skład wyrazów słownika.

Rezultaty doświadczenia, dostarczając informacji o rodzajach najczęstszych błędów, pozwolą również ukierunkować dalsze prace nad udoskonaleniem układu.

Dobór materiału

Przystępując do opracowania materiału doświadczalnego przyjęto zasdę, iż porównywane będą ze sobą głoski potencjalnie narażone na mieszanie ze sobą w procesie identyfikacji. Z góry należy przypuszczać, że nie dojdzie do mieszania pomiędzy np. /p/ oraz /ʃ/ ze względu na ich całkowicie odmienne charakterystyki widmowe, co pozwala nie uwzględniać tego rodzaju przypadków przy tworzeniu zestawu doświadczalnego. Aby zapewnić czytelną interpretację uzyskanych wyników nie sporządzono jednej listy logatomów, wspólnej dla wszystkich uwzględnionych głosek, lecz kilka osobnych zestawów, w obrębie których koncentrowano się na wybranych cechach dystynktywnych. Rozpoznawanie dotyczyło więc każdorazowo określonej grupy wypowiedzi.

Grupy od 1 do 7 obejmowały po szesnaście logatomów, grupa 8 - osiem logatomów. Wszystkie logatomy posiadały budowę CVC. W pierwszych siedmiu grupach elementem stałym była samogłoska, natomiast zmiennym spółgłoski: nagłosowa i wygłosowa. Zarówno w nagłosie, jak wygłosie występowały po cztery kombinacje spółglosek, wchodzące z sobą w opozycje. Dawało to w efekcie czterokrotne występowanie w obrębie grupy każdej spółgłoski nagłosowej Cn oraz czterokrotnie każdej spółgłoski wygłosowej Cw. Ilustruje to poniższy schemat:

$Cn_1V\ Cw_1$	$Cn_1V\ Cw_2$	$Cn_1V\ Cw_3$	$Cn_1V\ Cw_4$
$Cn_2V\ Cw_1$	$Cn_2V\ Cw_2$	$Cn_2V\ Cw_3$	$Cn_2V\ Cw_4$
$Cn_3V\ Cw_1$	$Cn_3V\ Cw_2$	$Cn_3V\ Cw_3$	$Cn_3V\ Cw_4$
$Cn_4V\ Cw_1$	$Cn_4V\ Cw_2$	$Cn_4V\ Cw_3$	$Cn_4V\ Cw_4$

Przy takiej strukturze grupy każdy testowany logatom może być teoretycznie zidentyfikowany jako którykolwiek z szesnastu znajdujących się w grupie. To oznacza, że błąd może wystąpić na pozycji Cn, Cw lub równocześnie Cn i Cw.

W grupie 8, dotyczącej błędów samogłoskowych, elementem stałym były spółgłoski nagłosowa i wygłosowa, zaś zmiennym - samogłoska. Liczebność tej grupy była zdeterminowana przyjętą liczbą samogłosek objętych badaniem: sześć fonemów ustnych i

dwa nosowe: /ẽ/, /õ/.

Kryteria doboru spółgłosek w grupach 1-7 przyjęto kierując się podobieństwem ich cech widmowych, jak również w oparciu o wyniki prac z zakresu wizualnego rozpoznawania mowy oraz psychofonetycznej klasyfikacji spółgłosek [9], [10]. Przyjęto założenie, że błędy stwierdzone w wyniku doświadczeń nad wzrokowym rozpoznawaniem spektrogramów mogą również wystąpić w trakcie automatycznego rozpoznawania z uwagi na to, iż w obu tych przypadkach o wyniku identyfikacji decyduje podobieństwo wizualne głosek. Wykorzystano matryce błędów uzyskane w wyniku przeprowadzonych doświadczeń z odczytywaniem spektrogramów [10]

Kolejnym czynnikiem, który wzięto pod uwagę przy tworzeniu grup logatomowych, było podobieństwo percepcyjne wyznaczone na podstawie testów odsłuchowych z zastosowaniem psychometrycznych metod klasyfikacji spółgłosek [9]. Stopień bliskości spółgłosek wyznaczony w oparciu o podobieństwo percepcyjne pokrywa się w znacznej mierze ze stopniem bliskości wynikającym z podobieństwa wizualnego.

Poniższa tabela podaje skład fonemowy wszystkich grup z uwzględnieniem pozycji w logatomie.

Tab. 1

Głoski wchodzące w skład logatomów w poszczególnych zestawach doświadczalnych.

Grupa	Spółgłoska nagłosowa				Samogłoska	Spółgłoska wygłosowa			
	Cn_1	Cn_2	Cn_3	Cn_4	V	Cw_1	Cw_2	Cw_3	Cw_4
1	b	d	g	v	o	p	t	k	x
2	z	ʒ	ʒ	ʃ	a	t͡s	t͡ɕ	t͡ʃ	t
3	t͡ɕ	d͡ʒ	t	d	a	m	n	w	l
4	t͡s	s	t͡ɕ	ɕ	o	r	l	m	ɲ
5	ʒ	d͡ʒ	ɲ	j	e	k	t͡s	f	x
6	v	z	n	w	a	j	m	ɲ	n
7	f	x	t͡ʃ	ʃ	u	s	ɕ	ʃ	t͡ʃ
8	Cn				V_1 V_2 V_3 V_4	Cw			
	m				i ɨ e ẽ	s			
					V_5 V_6 V_7 V_8				
					a o õ u				

Prawie wszystkie spółgłoski występują w materiale doświadczalnym więcej niż jeden raz, co jest spowodowane koniecznością przeprowadzania porównań między nimi w różnych kombinacjach.

O umieszczeniu danej spółgłoski w określonych zestawach decydowały różne przesłanki. Na przykład:

/d/ umieszczono w grupie 1 wraz z innymi zwartymi dźwięcznymi, z którymi było wzajemnie mieszane w trakcie rozpoznawania wizualnego, o czym zadecydowało niewątpliwie ich silne podobieństwo w zakresie cech widmowych. Równocześnie spółgłoski te wykazują bliskie podobieństwo percepcyjne.

/d/ umieszczono również w grupie 3, gdzie wchodzi w opozycję z /t/ - ten fakt umożliwia zbadanie rozróżnialności spółgłosek dźwięcznych i bezdźwięcznych - oraz w opozycję z /dʒ/ - zbliżone cechy fonetyczno-akustyczne tych głosek mogą doprowadzić do powstania błędów identyfikacji, analogicznie do stwierdzonych w trakcie odczytywania spektrogramów.

/x/ umieszczono w grupie 1 wraz ze zwartymi bezdźwięcznymi /p/, /t/, /k/ ze względu na ich bliskie podobieństwo percepcyjne.

/x/ wykazuje największe podobieństwo fonetyczno-akustyczne i percepcyjne z /f/, stąd wykorzystano kontrast tych dwóch spółgłosek w grupie 5 (wygłos) i 7 (nagłos). Powtórne umieszczenie ich w grupie 7 miało na celu stwierdzenie, czy poprawność rozpoznawania danej głoski wiąże się również z jej pozycją w wypowiedzi.

Zastosowane wobec materiału doświadczalnego kryteria językowe: wysoki stopień podobieństwa fonetyczno-akustycznego, wizualnego lub percepcyjnego, często występujących równocześnie, jak również bardzo mała rozciągłość wypowiedzi obejmujących zaledwie trzy głoski, stawiały wysokie wymagania wobec układu rozpoznającego. Jednakże oparcie się na ostrych kryteriach pozwoli uzyskać istotną informację na temat efektywności układu w zakresie dystynktywności fonetyczno-akustycznej.

Opis doświadczenia.

Celem eksperymentu było uzyskanie odpowiedzi na pytanie, ile razy została rozpoznana błędnie każda z głosek ujętych w materiale doświadczalnym oraz w przypadku zaistniałego błędu, która z głosek z testowanej grupy została odebrana na miejsce nadanej. W rozpatrywanej metodzie rozpoznawania identyfikacja rozpoznawanego obrazu zwanego obiektem następuje poprzez ocenę, do którego ze zgromadzonych wcześniej obrazów traktowanych jako wzorce obiekt wykazuje największe podobieństwo. Zamiast pojęcia podobieństwo zastosowano pojęcie odległości pomiędzy obrazami. Wartość tej odległości wyraża się liczbą dwuelementowych segmentów testowanego obrazu (obiektu), dla których nie znaleziono podobnych segmentów w obrazie przyrównywanym (wzorcu). Warunkiem istnienia wzajemnego podobieństwa dwóch segmentów jest spełnienie następującego kryterium:

$$\frac{nz(1)}{\Sigma(1)_{st}} < k$$

gdzie nz(1) oznacza liczbę niezgodnie występujących jedynek w parze porównywanych segmentów,

$\Sigma(1)$ oznacza sumę wszystkich jedynek w rozpatrywanym segmencie testowanego obrazu,

k jest progiem określającym granicę podobieństwa segmentów.

Przyjęto $k=\frac{1}{2}$. Oznacza to, że dwa segmenty różnych obrazów uważa się za podobne, jeśli liczba niezgodnie występujących w nich jedynek stanowi mniej niż połowę liczby jedynek w segmencie obrazu testowanego (obiektu). O wyniku identyfikacji decyduje wartość odległości pomiędzy obrazami binarnymi.

Rolę wzorców i obiektów spełniały w doświadczeniu obrazy binarne trzech wypowiedzi każdego logatomu. Testowano obraz każdej z trzech wypowiedzi każdego logatomu osobno dla każdej grupy logatomów. Oznacza to, że testowanie obrazu wypowiedzi danego logatomu następowało przy użyciu zbioru obrazów wypowiedzi wyłącznie logatomów z tej grupy. Obraz każdego logatomu danej grupy występował jeden raz jako testowany obiekt, zaś wielokrotnie jako wzorzec, do którego przyrównywano jako obiekty wszystkie testowane wypowiedzi tej grupy.

Wypowiedzi dostarczył jeden głos (**kobiecy**). Ponieważ każdy

logatom został wymówiony trzykrotnie, rozpoznanie w obrębie grupy dotyczyło 48 wypowiedzi (16 logatomów x 3 powtórzenia). Dla każdego logatomu należało więc wyznaczyć 48 wskaźników odległości (48 wzorców w grupie), z czego jedna wartość odnosiła się do porównania obiektu samego z sobą, dwie wartości do porównania obiektu z dwoma innymi **replikacjami** testowanego logatomu, zaś pozostałe (w liczbie 45) do porównania obiektu z wzorcami wypowiedzi wszystkich innych logatomów w grupie.

Przed przystąpieniem do zasadniczej części pracy należało materiał doświadczalny utrwalić na taśmie magnetofonowej. W kabinie bezechowej odczytano trzykrotnie wszystkie logatomy z każdej grupy, przy czym kolejne powtórzenie następowało po odczytaniu wszyskich grup 1-8. Utrzymywano równe tempo i głośność wypowiedzi oraz równą intonację, zachowując pomiędzy sąsiadującymi logatomami 2-sekundowe przerwy.

Dla każdej wypowiedzi sporządzono po cztery obrazy binarne. Każdy obraz wyznaczony był według jednej z dwóch metod przekształcenia widma amplitudowego w wektor binarny, zwany też umownie widmem binarnym,

- pierwszej, będącej połączeniem metody wypukłości obwiedni widma i metody maskowania oraz
- drugiej, będącej metodą maskowania.

Metody te przedstawiono szerzej w pracach [4], [7], [8]. Według metody wypukłości obwiedni widma poszczególne parametry wektora binarnego otrzymują wartość 1 lub 0 zależnie od istnienia lub braku odpowiedniej wypukłości obwiedni widma w przyporządkowanych tym parametrom punktach widma. Według metody maskowania o wartości parametru wektora binarnego decyduje położenie obwiedni widma względem tak zwanego poziomu maskowania, zależnego w danym punkcie widma od wartości współczynnika maskowania i od pola powierzchni pod obwiednią, w przedziale rozciągającym się z lewej strony tego punktu. Każdy obraz binarny wyznaczony był przy użyciu jednej z dwóch wartości współczynnika maskowania równych 1/16 i 1/32.

Z nagranych wypowiedzi sporządzono 16-elementowe zbiory obrazów binarnych. Na wstępie dla każdej analizowanej wypowiedzi wyświetlał się na ekranie monitora spektrogram z

4-stopniową skalą poziomów, co pozwalało operatorowi ocenić jakość wypowiedzi. W razie stwierdzenia usterek analizę przeprowadzano ponownie, po czym następowało wyznaczenie obrazów binarnych wypowiedzi. Po tej operacji na ekranie ukazywały się równocześnie cztery obrazy binarne uzyskane według wyżej podanych czterech wariantów ich wyznaczania. Cztery wersje obrazów wypowiedzi dla każdego z czterech kolejnych logatomów gromadzono w jeden zbiór i przesyłano na dyskietkę. Utworzono łącznie 34 zbiory obrazów wypowiedzi. Na etapie identyfikacji pobierano z dyskietki żądane obrazy, wyznaczając odległości pomiędzy nimi. Ostatni etap stanowił wydruk macierzy odległości. W wyniku przeprowadzonych we wszystkich grupach porównań uzyskano dla każdego z czterech wariantów tworzenia obrazów binarnych po 7 macierzy odległości o wymiarach 48 x 48 oraz po jednej macierzy o wymiarach 24 x 24. Większe macierze dotyczyły grup 1-7 (16-logatomowych), a mniejsza macierz grupy 8 (8-logatomowej). Łącznie uzyskano 32 macierze.

Omówienie wyników.

W każdym wierszu każdej z uzyskanych macierzy występuje wartość odległości równa 0. Odnosi się ona do porównania obiektu z samym sobą. Zwiększanie się wartości odległości oznacza zmniejszanie podobieństwa pomiędzy porównywanymi obrazami. W idealnym przypadku najmniejszą wartość w całym wierszu powinna osiągnąć którakolwiek z dwóch repetycji logatomu reprezentowanego przez rozpoznawany obiekt. Taka sytuacja oznaczałaby, iż spośród 48 wzorców, z którymi porównywano testowany logatom, najbardziej podobny (poza obrazem samego obiektu) okazał się obraz tejże repetycji. Wskaźnik odległości dla tej wypowiedzi może przybrać również wartość 0, co oznacza, że układ odebrał jej obraz jako w praktyce identyczny z obrazem obiektu. W przypadku, gdy układ rozpoznający wybierze spośród 47 wzorców repetycję testowanego logatomu, jako najbardziej zbliżoną do obiektu, wynik rozpoznawania jest poprawny. Wartość odległości pomiędzy obrazem binarnym obiektu a obrazem wzorca reprezentującego jedną z dwóch repetycji testowanego logatomu (przy czym chodzi

każdorazowo o wartość niższą spośród obu rozpatrywanych) należy więc traktować jako wartość progową, w stosunku do której wszystkie pozostałe w wierszu 46 wskaźników odległości powinno przyjąć wartości wyższe.

Sytuacja, gdy w wierszu pojawiają się wartości niższe lub równe progowej, oznacza rozpoznanie błędne. W tym przypadku obrazy innych logatomów zostały zidentyfikowane jako bardziej podobne do obiektu, aniżeli obraz którejkolwiek z dwóch repetycji testowanego logatomu, względnie podobne w takim samym stopniu, co prowadzi do niejednoznacznej interpretacji wyników. Przykłady identyfikacji poprawnych oraz błędnych zamieszczono w tab.2 dla dowolnie wybranych dwunastu wierszy macierzy wyznaczonej dla grupy 6, dla obrazów według wariantu WOW3. W rubryce odnoszącej się do rozpoznań błędnych podano wraz z wartością odległości logatomy, które jej dotyczą.

Tab.2.

Interpretacja wyników rozpoznawania w oparciu o kryterium progowe dla wierszy 25-36 macierzy MTG6WOW3.

Wypowiedź testowana	Wartość progowa odległości	Identyfikacja błędna	
		Odległość wzorca od obiektu mniejsza od wart.prog.	Odległość wzorca od obiektu równa wart. progowej
naj 1	5	-	5 vaj
naj 2	12	6 waj	12 naj, waj, waj
naj 3	22	6 waj 17 vaj, vaj, vaj, naj, waj	22 van, zan, nam naj, nan, waj wan
nam 1	0	-	-
nam 2	0	-	0 vam, wam
nam 3	0	-	0 vam, wam
naj 1	10	5 nan, nan	10 vaj, waj
naj 2	6	-	-
naj 3	5	-	5 waj
nan 1	5	0 naj	5 naj
nan 2	17	6 naj 11 vaj, van, naj, wan	17 vaj, zan, nam nam
nan 3	0	-	-

W oparciu o kryterium progowe przeprowadzono ocenę wyników rozpoznawania zamieszczonych w macierzach odległości. Dla każdej testowanej wypowiedzi sporządzono zestawienie logatomów, których obrazy binarne wykazały odległości mniejsze lub równe progowej, co zgodnie z przyjętymi kryteriami uznane jest za identyfikację błędną. Zebrany materiał stanowił podstawę do ustalenia, które spółgłoski mieszały się z sobą w procesie rozpoznawania.

Należy zdać sobie sprawę, iż pomieszanie dwóch spółgłosek, np. /n/ oraz /ɲ/ w przypadku odebrania logatomu /nan/ jako /naɲ/, nie musi oznaczać, że błąd wyniknął na skutek szczególnie silnego podobieństwa tych dwóch segmentów. Równie możliwa jest sytuacja, iż rozpatrywane spółgłoski różniły się wyraźnie między sobą, jednakże pozostałe odpowiadające sobie fragmenty porównywanych obrazów binarnych okazały się daleko bardziej do siebie podobne, niż w każdym z dwóch obrazów repetycji. Odnosząc to do powyższego przykładu mogło zdarzyć się tak, iż różnica pomiędzy /n/ oraz /ɲ/ wywarła mniejszy wpływ na wartość średniej odległości obrazów, niż podobieństwo pomiędzy fragmentami /na/ w obu obrazach. Uzyskanie szczegółowych odpowiedzi na pytanie, w którym fragmencie obrazów stwierdzono ekstremalne wartości podobieństw lokalnych nie było przedmiotem zainteresowań na obecnym etapie pracy. Badanie dotyczyło podatności głosek na błędną identyfikację w warunkach globalnego rozpoznawania wypowiedzi.

W myśl przyjętych założeń stosowane w dalszym opisie sformułowania typu: głoska /n/ została błędnie zidentyfikowana jako /ɲ/ nie mogą być traktowane jako równoznaczne ze stwierdzeniem, iż błędne rozpoznanie logatomu jest następstwem silnego podobieństwa lokalnego pomiędzy fragmentami obrazów odpowiadającymi tym spółgłoskom, lecz oznaczają, że ewentualne różnice między tymi spółgłoskami byłyby na tyle nieistotne, że nie odegrały roli przy ich porównywaniu w szerszym kontekście fonetycznym.

W oparciu o sporządzone zestawienia błędów popełnionych przy rozpoznawaniu logatomów uzyskano dane dotyczące poszczególnych głosek z uwzględnieniem osobno pozycji nagłosowej i wygłosowej. Dane te odnoszą się do określonych

zestawów głoskowych (grupy 1-8) oraz określonych wariantów uzyskiwania obrazów binarnych (WOW1, WOW3, PMW1, PMW3). Wartości liczbowe podają oddzielnie dla każdej pozycji, ile razy dana głoska została odebrana jako dowolna inna głoska z grupy. Ostateczne wyniki przybrały postać wartości procentowych, określających udział błędnych identyfikacji poszczególnych głosek w określonej pozycji w stosunku do liczby głosek w macierzy, z którymi przeprowadzono porównanie. Wyniki zaprezentowano w postaci graficznej na rycinach. Każdą z testowanych głosek reprezentują cztery diagramy odnoszące się do czterech zastosowanych wariantów obrazów binarnych. Wysokość każdego z nich oznacza 100% wszystkich wyników rozpatrywanych w odniesieniu do testowanej głoski.

Zakres występowania błędów jest różny dla poszczególnych zestawów. Najliczniej występowały błędy w nagłosie grupy 1, 3 i 5. Grupa 1 obejmuje głoski zwarte dźwięczne, których charakterystyki widmowe są bardzo do siebie zbliżone, a przyjęta częstość próbkowania sygnału bywa niewystarczająca ze względu na jego impulsowy przebieg oraz spółgłoskę /v/, która wykazuje znaczne podobieństwo wizualne ze zwartymi dźwięcznymi. Spółgłoski z grupy 3 - zwarte /t/, /d/ i zwarto-trące /t͡ɕ/, /d͡ʑ/ posiadają bardzo zbliżone cechy fonetyczno-akustyczne. W nagłosie grupy 5 znalazły się głoski /ʑ/, /d͡ʑ/, /ɲ/, /j/ wykazujące znaczne podobieństwo percepcyjne.

Zróżnicowaniu w zakresie ilości stwierdzonych błędów podlegają również głoski w obrębie grup. Np. w wygłosie grupy 2 /t͡s/ było rozpoznawane z minimalną ilością błędów, natomiast /t/ uzyskało w metodzie PMW1 łącznie 23% błędów, przy czym błędy rozłożyły się równomiernie pomiędzy /t͡s/, /t͡ɕ/ i /t͡ʃ/. Najczęściej głoski były mieszane z pozostałymi w grupie w równym mniej więcej stopniu, chociaż zdarzało się, że mieszanie następowało częściej z określonymi głoskami, np. /t͡ɕ/ (grupa 2) sporadycznie odbierano jako /t͡s/, zdecydowanie częściej jako /t͡ʃ/ i /t/. Preferowanie określonych głosek w przypadku wystąpienia błędnej identyfikacji szczególnie wyraźnie wystąpiło w grupie 8, obejmującej samogłoski. Jedynie /õ/ bywało odbierane jako każda z pozostałych samogłosek.

Dane z wykresów pozwalają stwierdzić, iż poprawność rozpoznawania w obrębie grupy, pomimo zastosowania ostrych kryteriów językowych, była dość wysoka. Liczba błędnie odebranych określonych głosek na miejsce nadanych mieści się w granicach kilku procent. Przykładowo w wygłosie grupy 1 /k/ odebrano jako /p/ w następującej liczbie przypadków w zależności od zastosowanego wariantu: 9%, 3%, 7%, 7%, /k/ jako /t/: 4%, 2%, 4%, 3%, /k/ jako /x/: 1%, 0%, 1%, 0%. Przekroczenie 10% udziału błędów stwierdzono w odniesieniu do nielicznych przypadków. I tak liczba błędów polegających na odebraniu /ʒ/ jako /ʑ/ wyniosła 14% dla jednego z wariantów, /d͡ʒ/ jako /t͡ɕ/ również 14%, /f/ jako /t͡s/ 12%, /f/ jako /x/ 11%. W odniesieniu do samogłosek najwyższa wartość błędnych identyfikacji wyniosła 10% i dotyczy rozpoznania /ɨ/ jako /e/.

Oceniając poprawność rozpoznawania w niniejszym doświadczeniu należy zawsze brać pod uwagę pary głosek: nadaną i odebraną. Łączna liczba błędnych identyfikacji w odniesieniu do określonej głoski nie stanowi w tym przypadku miarodajnej informacji, gdyż jest ściśle uzależniona od tego, które głoski znalazły się w rozpatrywanej grupie. Np. głoska /w/ w grupie 6 uzyskała łącznie 6%, 1%, 10%, 7% błędnych rozpoznań, zaś w grupie 3: 10%, 11%, 14%, 19%. Na tej podstawie nie sposób wnioskować, czy /w/ jest łatwo rozpoznawalną głoską, natomiast można mówić o większej podatności /w/ na błędną identyfikację, w przypadku gdy wzorce zawierają głoski z grupy 3, aniżeli wówczas, gdy są to głoski z grupy 6. Nieoczekiwany wynik uzyskano dla spółgłoski /d/, która wykazała większą liczbę błędnych identyfikacji w zestawieniu z głoskami /t͡ɕ/, /d͡ʑ/, /t/ (grupa 3), niż z głoskami /b/, /g/, /v/ (grupa 1). Należy spodziewać się, że rozszerzenie materiału doświadczalnego poprzez porównywanie z sobą innych kombinacji spółgłosek może przynieść nowe informacje na temat stopnia ich bliskości.

Zakres popełnianych błędów nie zawsze wykazuje symetryczny rozkład dla danej pary głosek. Np. liczba błędów polegających na odebraniu /t/ na miejsce /t͡s/ wyniosła dla czterech wariantów: 1%, 1%, 2%, 1%, natomiast w sytuacji odwrotnej /t͡s/ na miejsce /t/: 7%, 5%, 7%, 8%.. Dla określenia bliskości głosek ma więc znaczenie, która z nich wchodzi w skład obiektu,

a która w skład wzorca.

Wyniki uzyskane przy zastosowaniu poszczególnych wariantów tworzenia obrazów binarnych nie wykazują wyraźnego zróżnicowania. Żaden z wariantów nie okazał się w widoczny sposób korzystniejszy dla uzyskania lepszej poprawności rozpoznawania w zakresie użytego materiału doświadczalnego. Można jedynie stwierdzić nieznaczną przewagę wariantu WOW3 nad pozostałymi oraz łącznie potraktowanych metod WO nad metodami PM.

Uwagi końcowe.

Przeprowadzone doświadczenie pozwoliło ocenić wpływ czynnika lingwistycznego na poprawność globalnego rozpoznawania wyrazów według zastosowanej metody. Czynnik ten odgrywa ważną rolę ze względu na podobieństwa fonetyczno-akustyczne, wizualne lub percepcyjne zachodzące pomiędzy głoskami, a znajdujące swe odbicie w porównywanych obrazach binarnych. Podatność na błędną identyfikację jest w dużym stopniu uwarunkowana bliskością głosek wchodzących w skład wypowiedzi wzorcowych. Stąd istotne znaczenie posiada struktura fonetyczna wzorców wykorzystywanych w procesie identyfikacji, z którą związana jest możliwość mieszania rozpoznawanych haseł w przypadku nieuwzględnienia bliskości poszczególnych głosek.

Dane z rycin określające bliskość badanych głosek w aspekcie globalnego rozpoznawania wypowiedzi mogą stanowić ważną pomoc przy sporządzaniu słowników wykorzystywanych w układach rozpoznających. Odpowiedni dobór głosek tworzących hasła pozwoli zmniejszyć zakres błędnych identyfikacji.

Przeprowadzona analiza błędów może być również wykorzystana przy dalszych pracach nad udoskonalaniem systemu do rozpoznawania wyrazów.

Wszystkie poczynione tu obserwacje odnoszą się do jednego głosu. Przewiduje się przeprowadzenie dalszych tego rodzaju badań na większej liczbie głosów.

Bibliografia.

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

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

[3] KUBZDELA, H., Badania nad udoskonaleniem spektrogramów binarnych, Prace IPPT 24/1983, Warszawa.

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

[5] KUBZDELA, H., Model automatycznego rozpoznawania wyrazów na podstawie uproszczonych spektrogramów binarnych, w: Wizualizacja mowy i jej zastosowania (red. W. Jassem), Prace IPPT 1987, Warszawa, ss. 21-42.

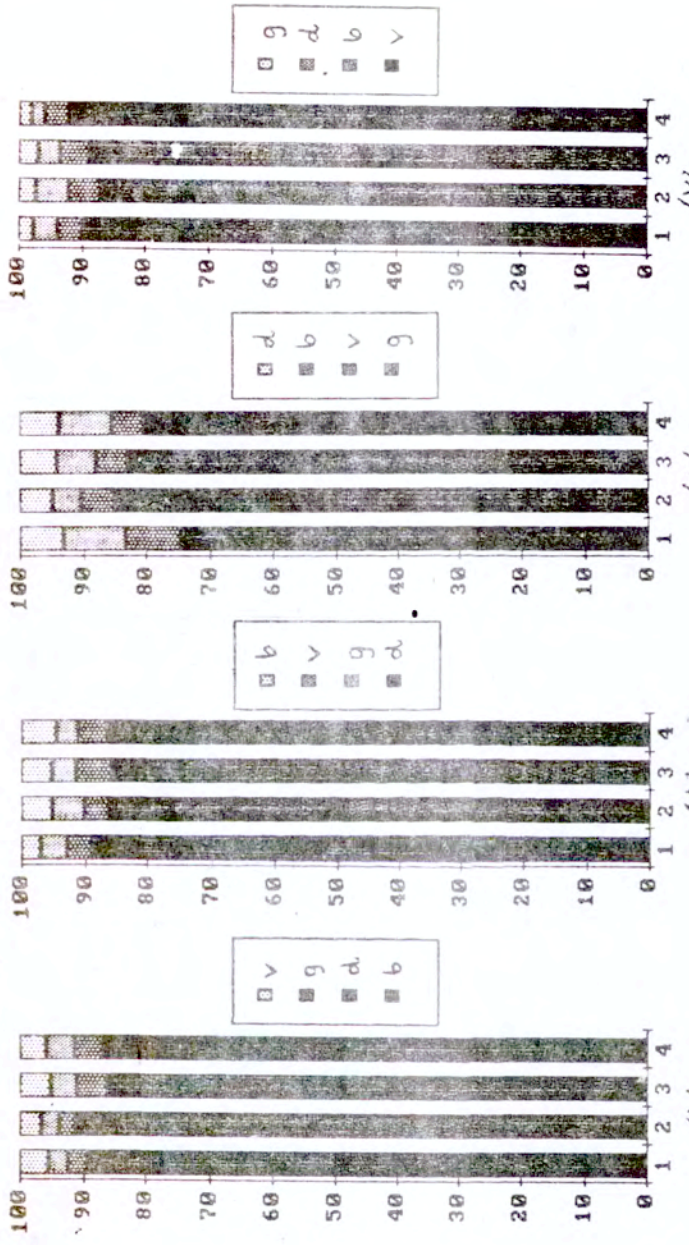
[6] KUBZDELA, H., Udoskonalenie reprezentacji sygnału mowy w formie obrazów binarnych, Prace IPPT 24/1987, Warszawa.

[7] KUBZDELA, H., Próby klasyfikacji dwuelementowych fragmentów w obrazie binarnym mowy opartym na pojęciu funkcji wagi widma, Prace IPPT 13/1988, Warszawa.

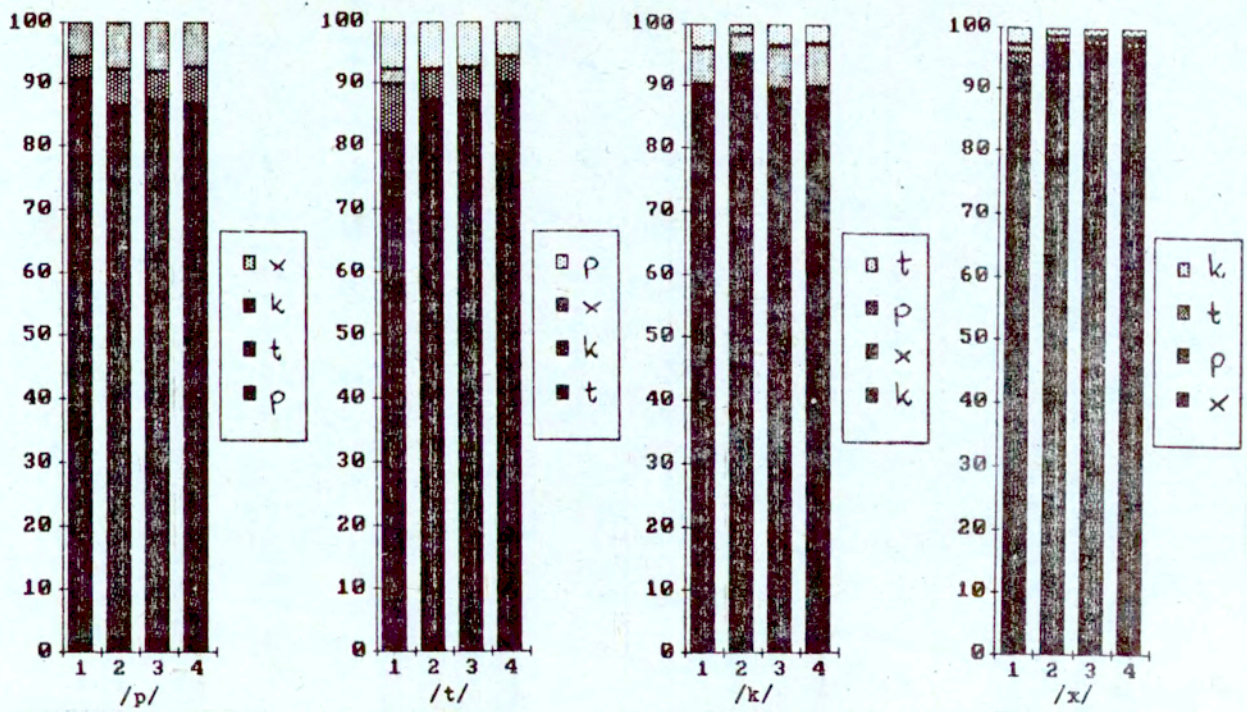
[8] KUBZDELA, H., Verwendungsmoglichkeit der binaren Darstellung von Spektralmerkmalen in Worterkennung, Proceedings of the 13th International Congress of Acoustics in Belgrade, vol.2, pp. 423-426.

[9] ŁOBACZ, P., Psychofonetyczne metody klasyfikacji elementów segmentalnych, (w druku).

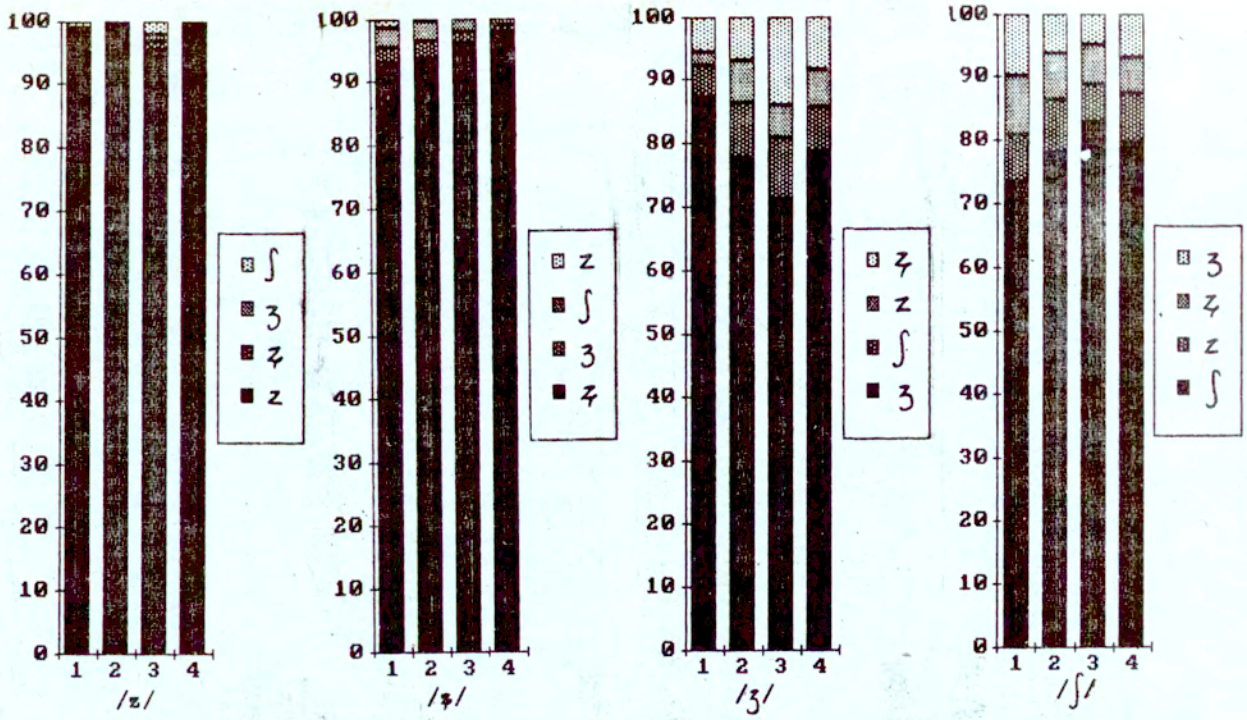
[10] RICHTER, L., Wizualne rozpoznawanie wybranych wyrazów w oparciu o informacje segmentalne zawarte w spektrogramach komputerowych, w: Wizualizacja mowy i jej zastosowania (red. W. Jassem), Prace IPPT 1987, Warszawa, ss. 135-180.



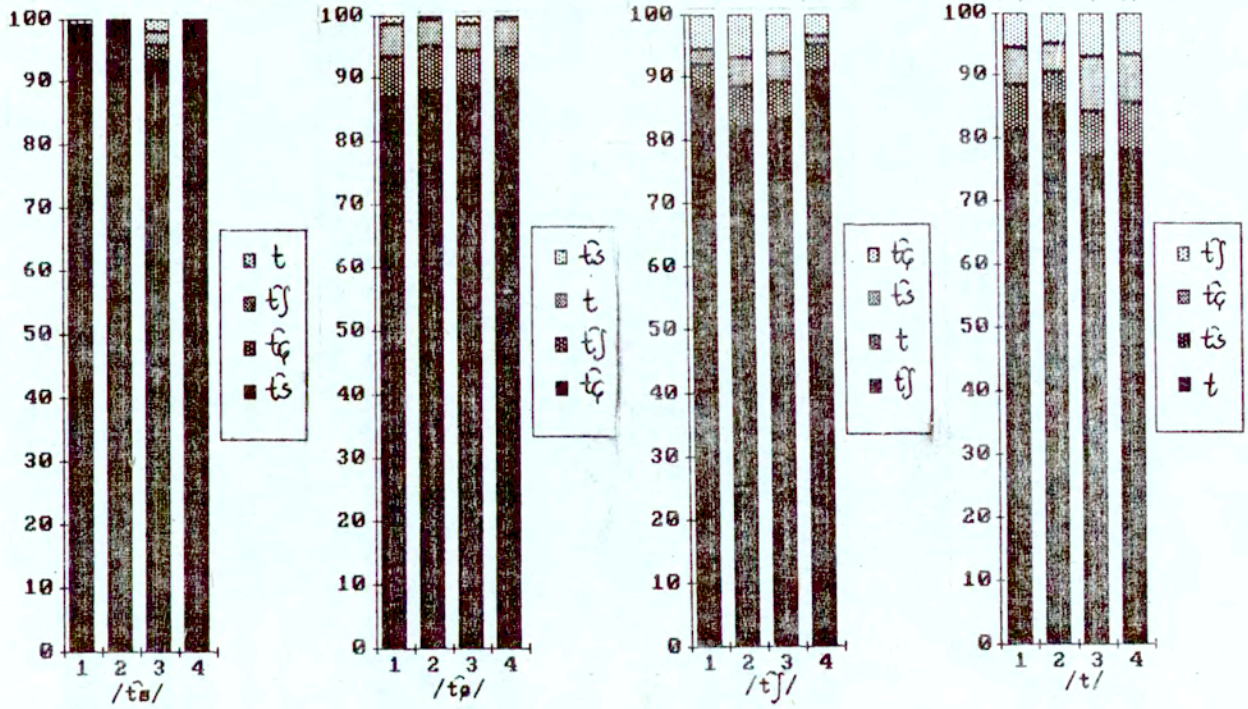
Ryc.1. Procentowy rozkład rozpoznań głosek: /b/, /d/, /g/, /v/ (grupa 1, nagłos) dla czterech wariantów obrazów: 1 (WOW1), 2 (WOW3), 3 (PMW1), 4 (PMW3).



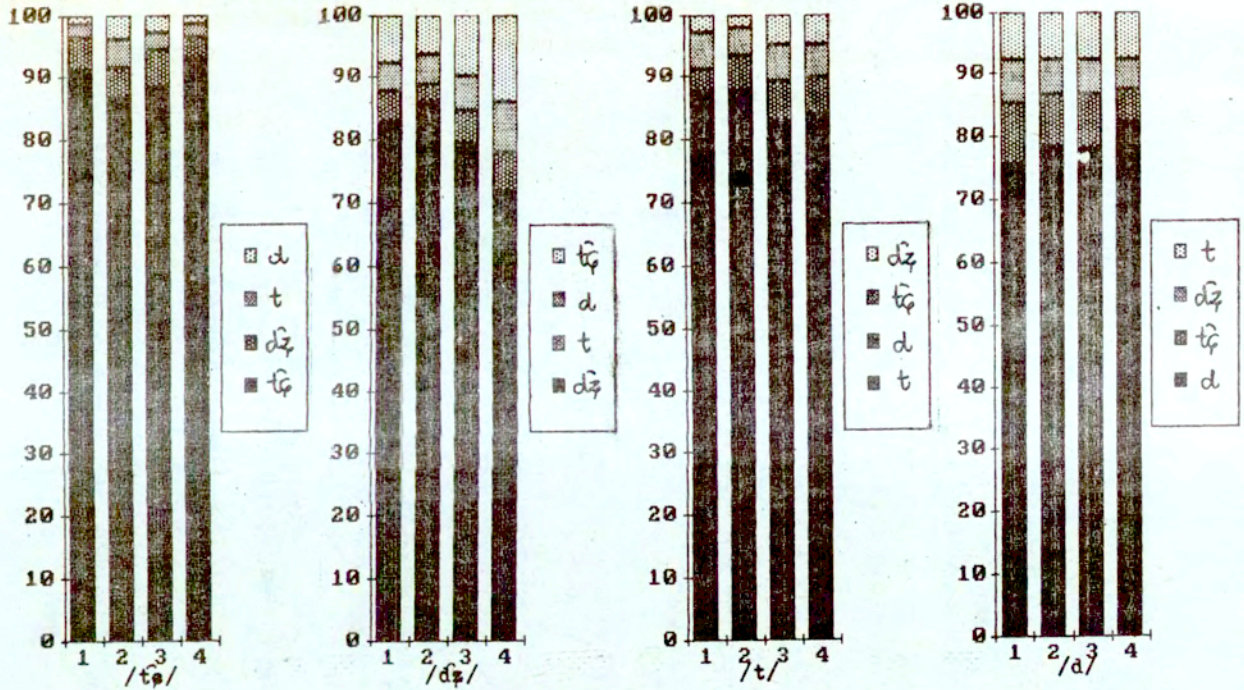
Ryc.2. Procentowy rozkład rozpoznań głosek: /p/, /t/, /k/, /x/ (grupa 1, wygłos) dla czterech wariantów obrazów: 1 (WOW1), 2 (WOW3), 3 (PMW1), 4 (PMW3).



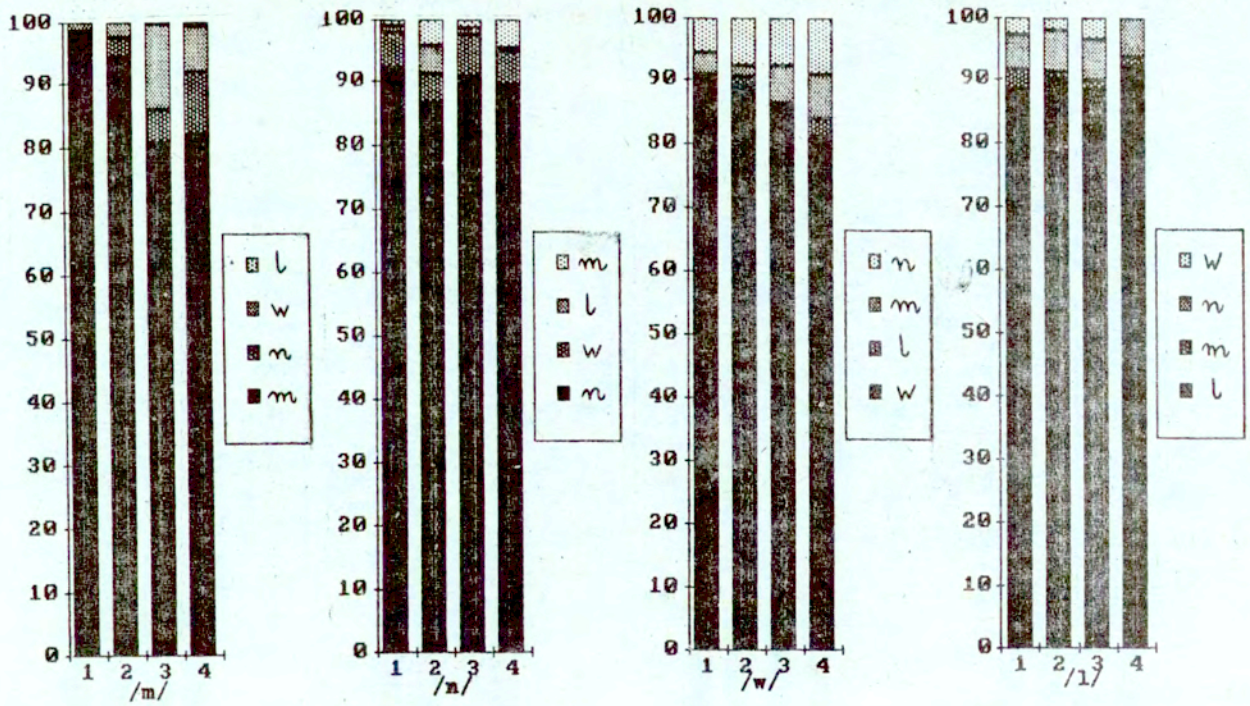
Ryc.3. Procentowy rozkład rozpoznań głosek: /z/, /ʑ/, /ʒ/, /ʃ/ (grupa 2, nagłos) dla czterech wariantów obrazów: 1 (WOW1), 2 (WOW3), 3 (PMW1), 4 (PMW3).



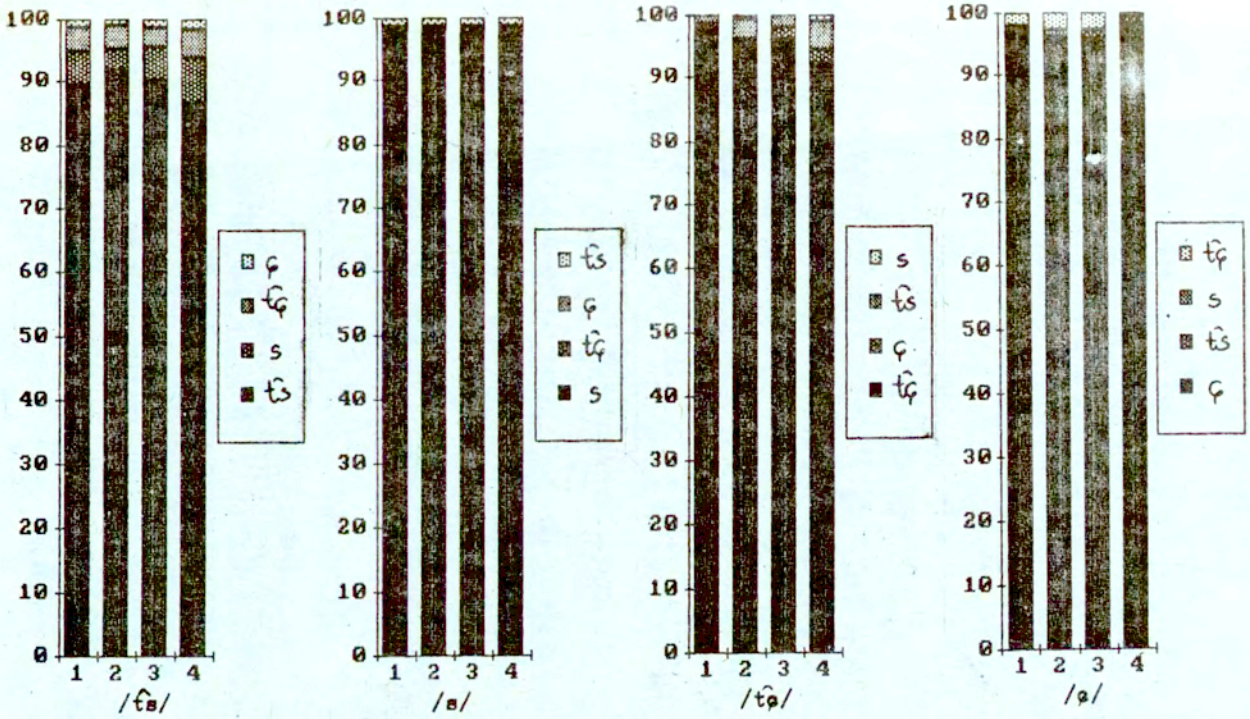
Ryc.4. Procentowy rozkład rozpoznań głosek: /t͡s/, /t͡ɕ/, /tʃ/, /t/ (grupa 2, wygłos) dla czterech wariantów obrazów: 1 (WOW1), 2(WOW3), 3 (PMW1), 4 (PMW3).



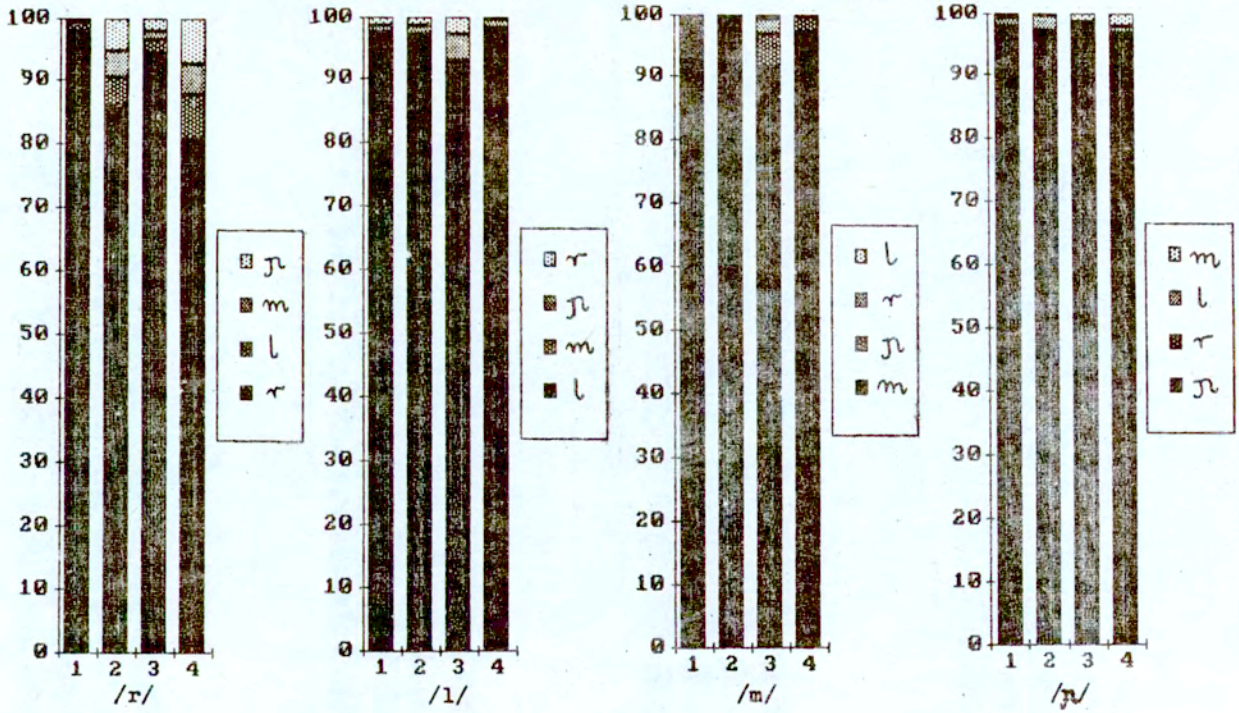
Ryc.5. Procentowy rozkład rozpoznań głosek: /tɕ/, /dʑ/, /t/, /d/ (grupa 3, nagłos) dla czterech wariantów obrazów: 1 (WOW1), 2 (WOW3), 3 (PMW1), 4 (PMW3).



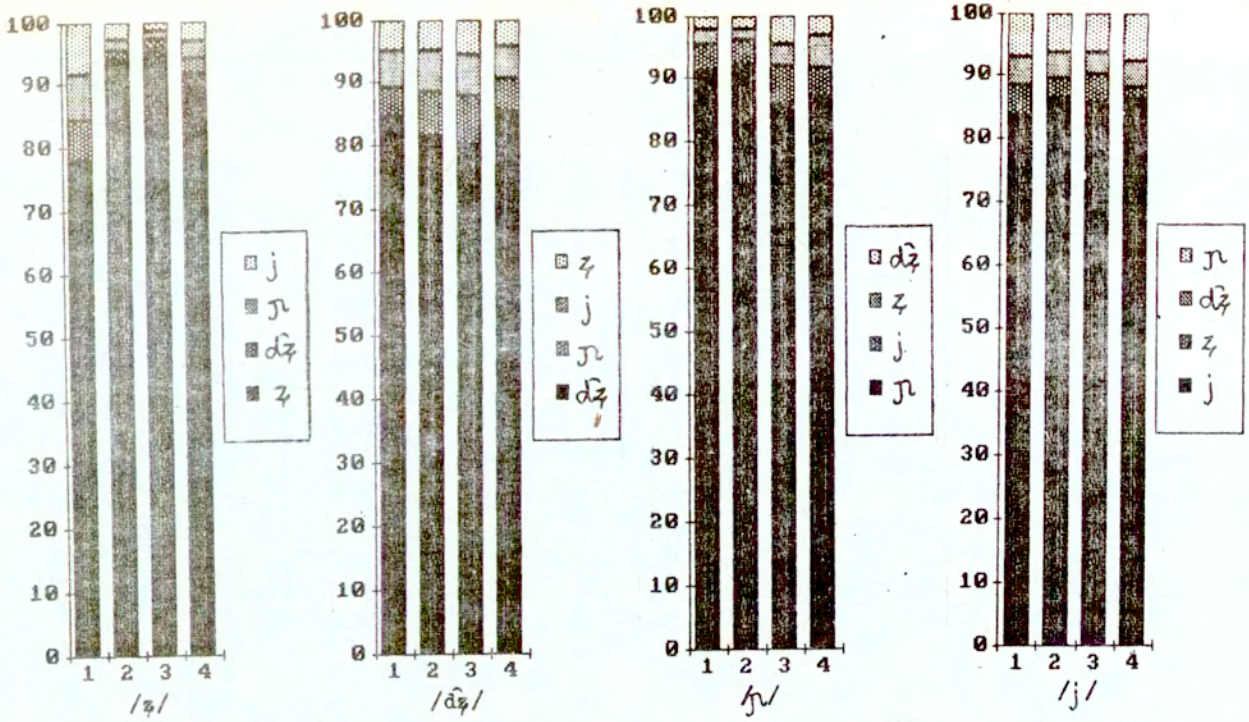
Ryc.6. Procentowy rozkład rozpoznań głosek: /m/, /n/, /w/, /l/ (grupa 3, wygłos) dla czterech wariantów obrazów: 1 (WOW1), 2 (WOW3), 3 (PMW1), 4 (PMW3).



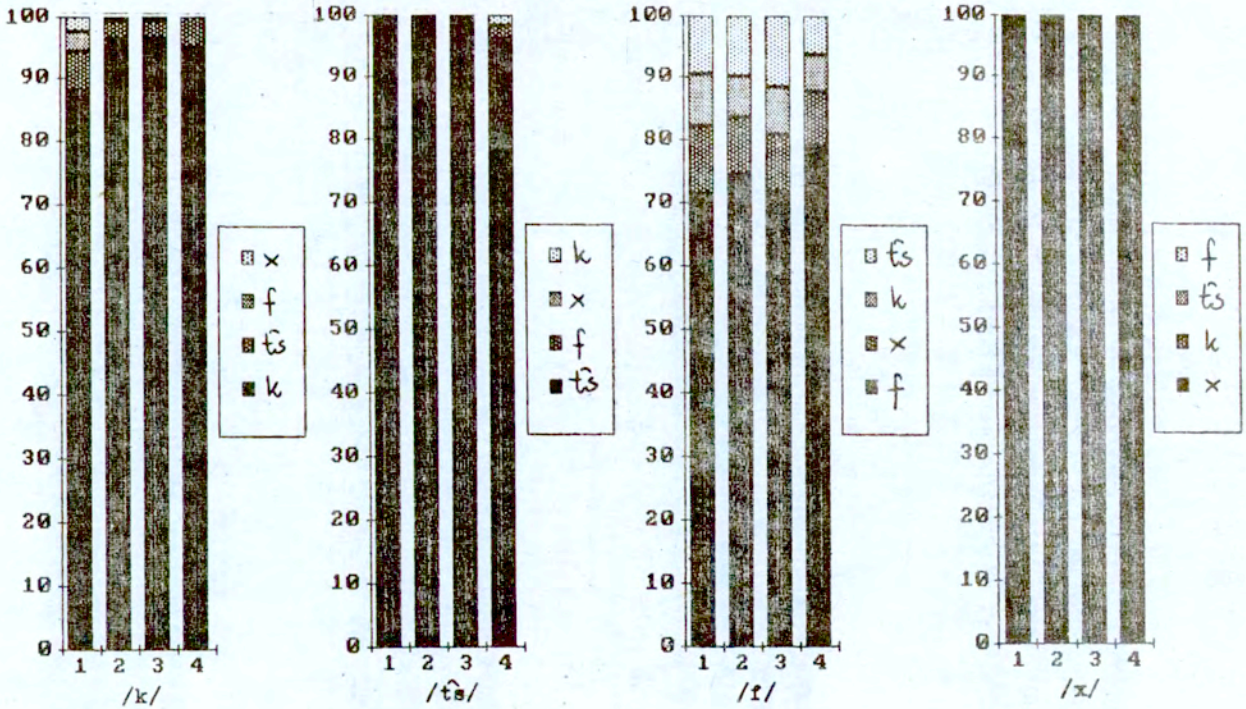
Ryc.7. Procentowy rozkład rozpoznań głosek: /ʧs/, /s/, /ʧɕ/, /ɕ/ (grupa 4, nagłos) dla czterech wariantów obrazów: 1 (WOW1), 2 (WOW3), 3 (PMW1), 4 (PMW3).



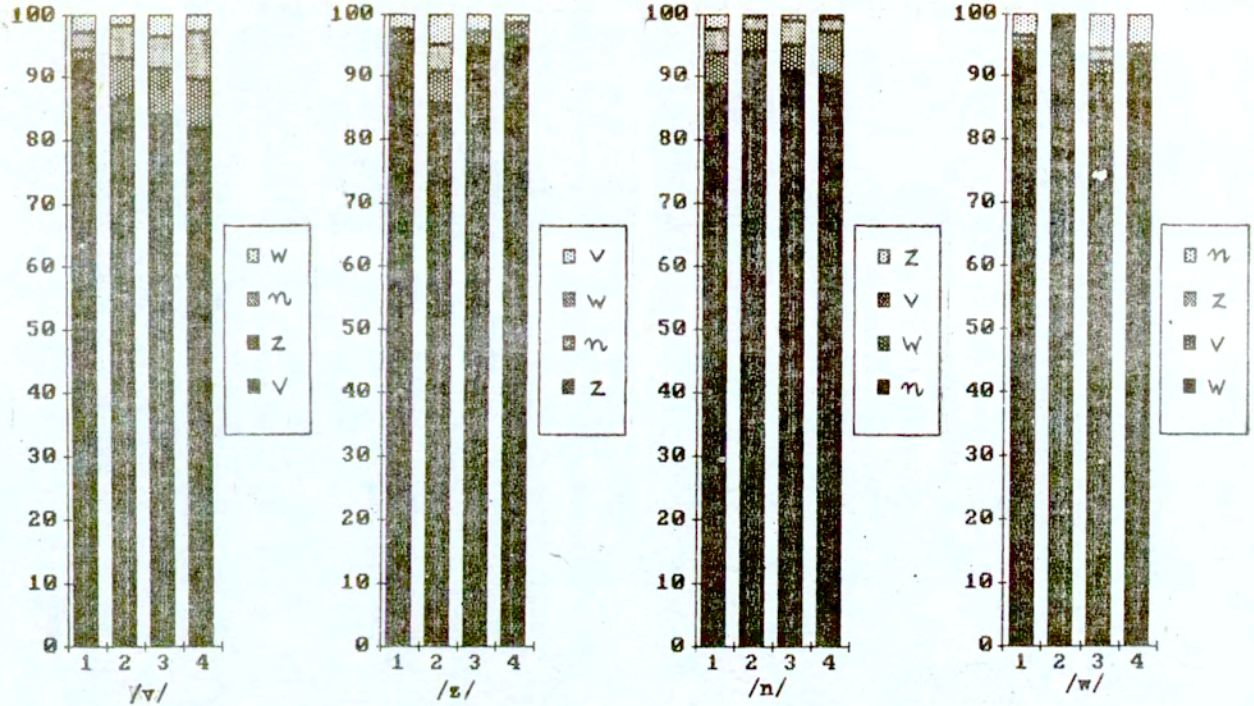
Ryc.8. Procentowy rozkład rozpoznań głosek: /r/, /l/, /m/, /ɲ/ (grupa 4, wygłos) dla czterech wariantów obrazów: 1 (WOW1), 2 (WOW3), 3 (PMW1), 4 (PMW3).



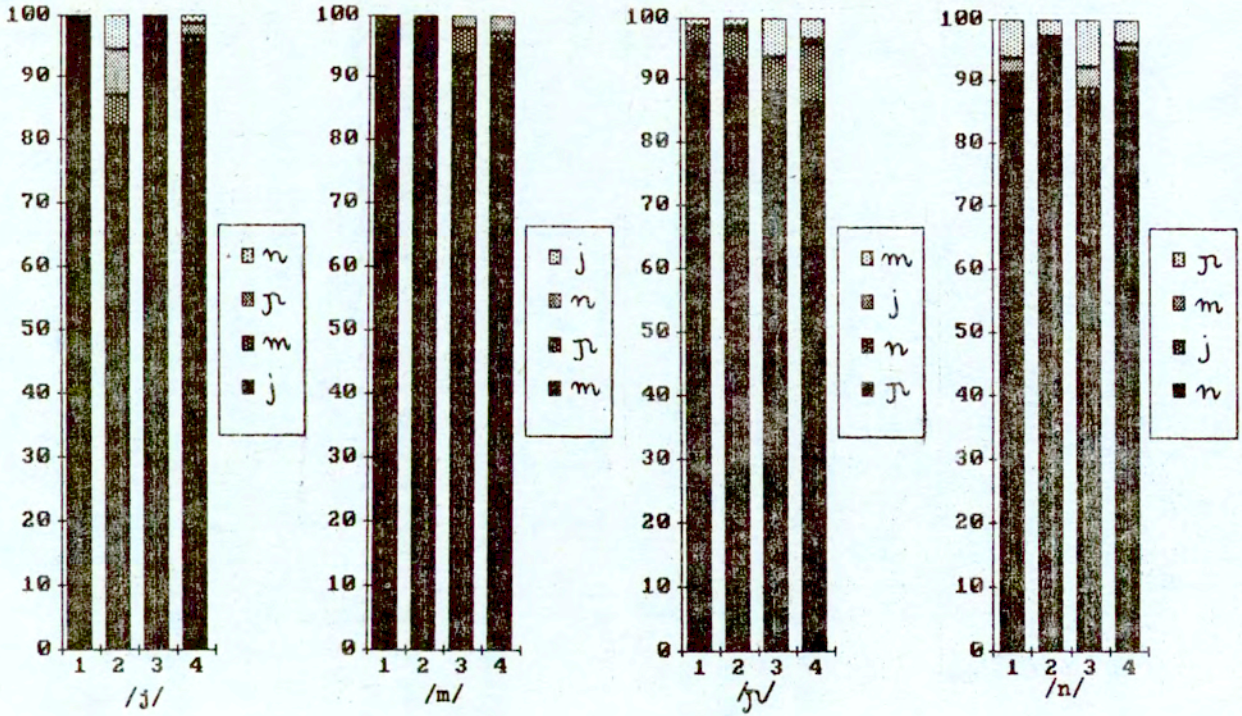
Ryc.9. Procentowy rozkład rozpoznań głosek: /ʑ/, /dʑ/, /ʃ/, /j/ (grupa 5, nagłos) dla czterech wariantów: 1 (WOW1), 2 (WOW3), 3 (PMW1), 4(PMW3).



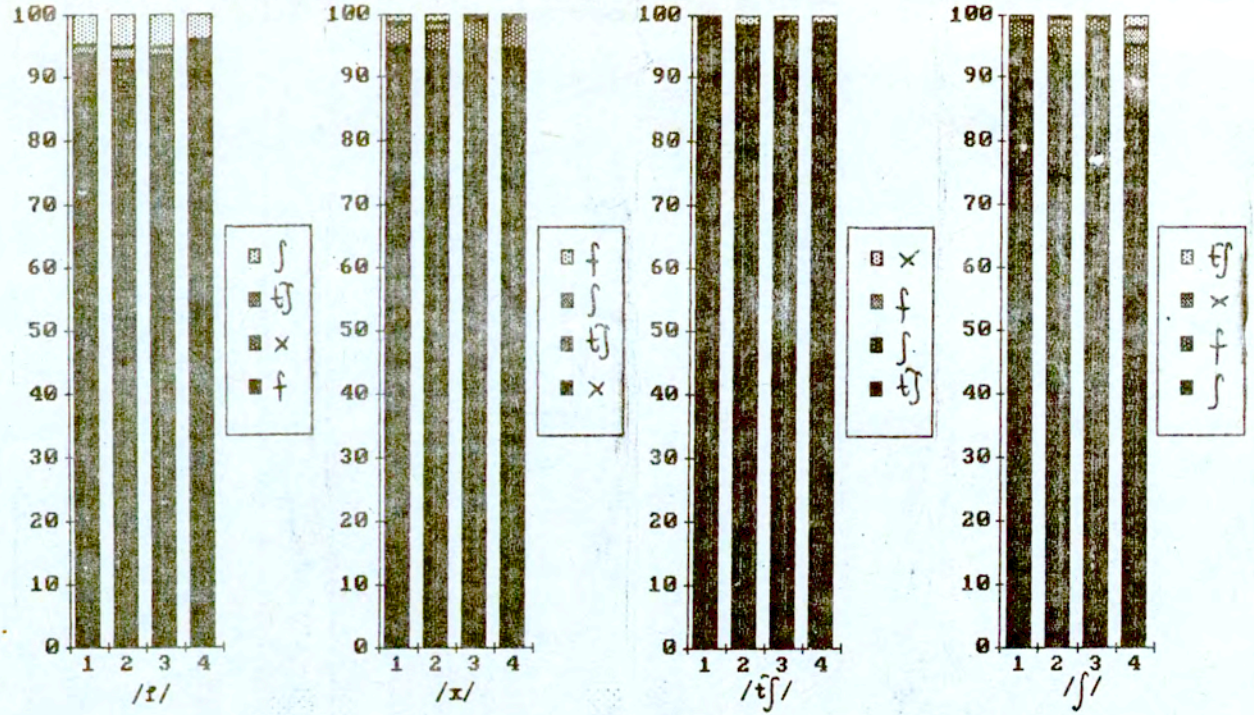
Ryc.10. Procentowy rozkład rozpoznań głosek: /k/, /ʧ/, /f/, /x/ (grupa 5, wygłos) dla czterech wariantów obrazów: 1 (WOW1), 2 (WOW3), 3 (PMW1), 4 (PMW3).



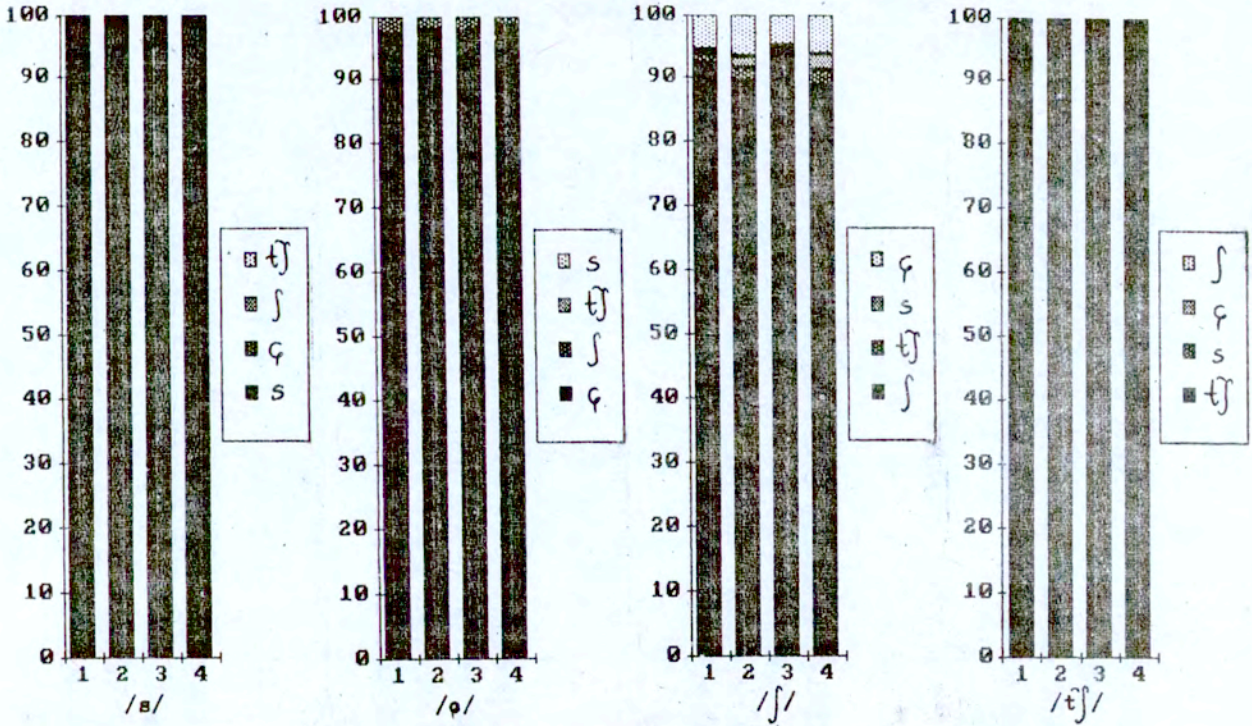
Ryc. 11. Procentowy rozkład rozpoznań głosek: /v/, /z/, /n/, /w/ (grupa 6, nagłos) dla czterech wariantów obrazów: 1 (WOW1), 2 (WOW3), 3 (PMW1), 4 (PMW3).



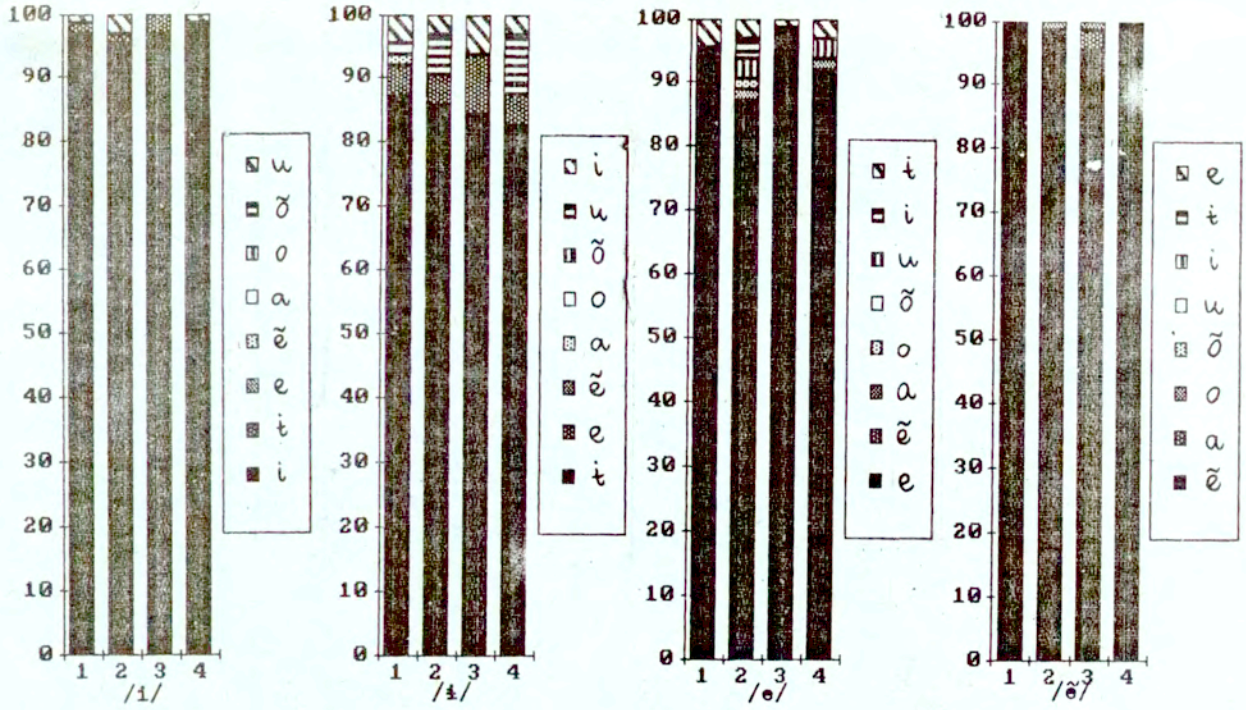
Ryc.12. Procentowy rozkład rozpoznań głosek: /j/, /m/, /ɲ/, /n/ (grupa 6, wygłos) dla czterech wariantów obrazów: 1 (WOW1), 2 (WOW3), 3 (PMW1), 4 (PMW3).



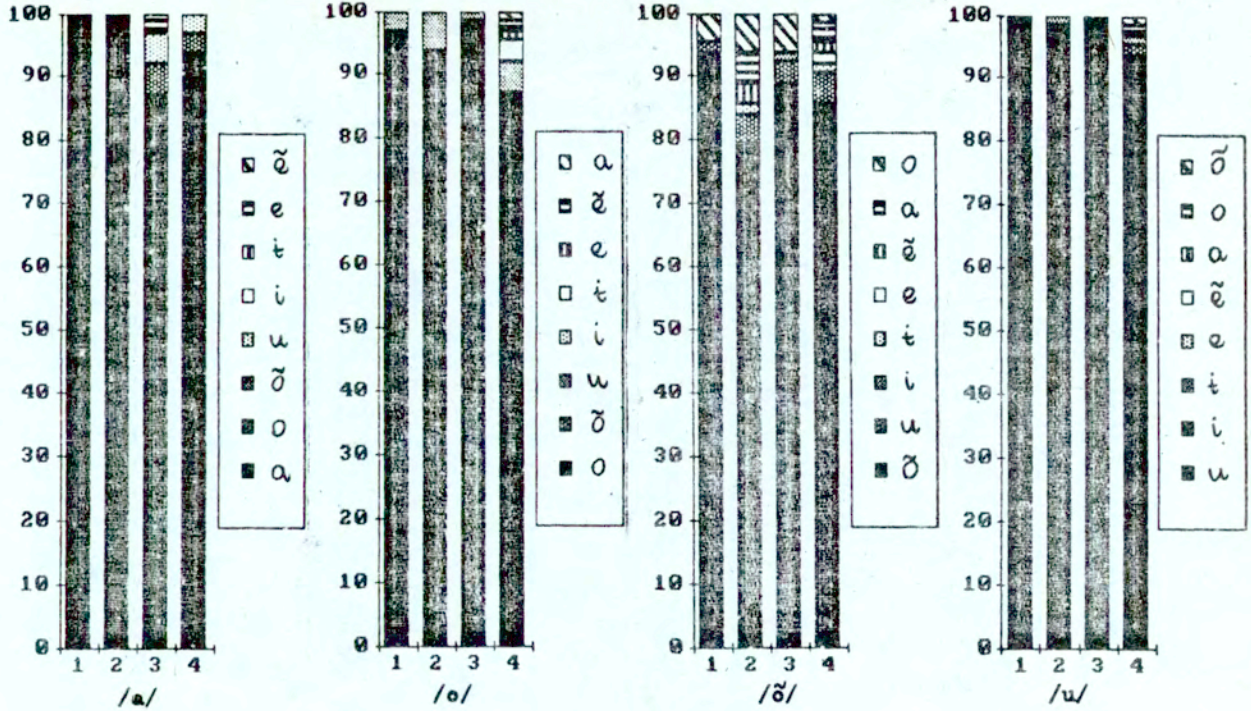
Ryc.13. Procentowy rozkład rozpoznań głosek: /f/, /x/, /tʃ/, /ʃ/ (grupa 7, nagłos) dla czterech wariantów obrazów: 1 (WOW1), 2 (WOW3), 3 (PMW1), 4 (PMW4).



Ryc.14. Procentowy rozkład rozpoznań głosek: /s/, /ɸ/, /ʃ/, /tʃ/ (grupa 7, wygłos) dla czterech wariantów obrazów: 1 (WOW1), 2 (WOW3), 3 (PMW1), 4 (PMW3).



Ryc. 15. Procentowy rozkład rozpoznań głosek: /i/, /ɨ/, /e/, /ẽ/, (grupa 8) dla czterech wariantów obrazów: 1 (WOW1), 2 (WOW3), 3 (PMW1), 4 (PMW3).



Ryc.16. Procentowy rozkład rozpoznań głosek: /a/, /o/, /õ/, /u/ (grupa 8) dla czterech wariantów obrazów: 1 (WOW1), 2 (WOW3), 3 (PMW1), 4 (PMW3).