H. Kubzdela, L. Richter, W. Jassem

AUTOMATYCZNE ROZPOZNAWANIE HASEŁ STERUJĄCYCH DLA POTRZEB OSÓB NIEPEŁNOSPRAWNYCH

17/1991

Warszawa 1991

ISSN 0208-5658

Praca wpłynęła do Redakcji dnia 5 lutego 1990 r.

IPPT PAN

Na prawach rękopisu

Instytut Podstawowych Problemów Techniki PAN
Nakład 130 egz. Ark.wyd.2 , Ark.druk. 2,5
Oddano do drukarni w maju 1991 r.
Nr zamówienia 1 60/91

Warszawska Drukarnia Naukowa, Warszawa,
ul.Śniadeckich 8

Henryk Kubzdela
Lutosława Richter
Wiktor Jassem
Zakład Fonetyki Akustycznej
IPPT PAN

AUTOMATYCZNE ROZPOZNAWANIE HASEŁ STERUJĄCYCH DLA POTRZEB OSÓB NIEPEŁNOSPRAWNYCH[1].

Streszczenie

Celem pracy było zbadanie efektywności rozpoznawania haseł sterujących wypowiadanych przez 10 osób (kobiety i mężczyzn). Zbiór obiektów rozpoznawanych objął 100 wypowiedzi (10 osób × 10 haseł), zaś zbiór wzorców 300 wypowiedzi (10 osób × 10 haseł × 3 powtórzenia). Zastosowano metodę rozpoznawania opartą na reprezentacji wypowiedzi wyrazu w formie obrazu binarnego. Dla każdej wypowiedzi sporządzono cztery wersje obrazu binarnego, wyznaczone według jednej z dwóch metod przekształcania wygładzonego widma amplitudowego w wektor binarny oraz jednej z dwóch wartości współczynnika maskowania. W wyniku przeprowadzonych porównań pomiędzy obrazami testowanymi i wzorcowymi uzyskano tablicę identyfikacji o wymiarach 30 × (100 × 4). Elementami tablicy są odległości poszczególnych obiektów od wybranych wzorców. Każdy wiersz tablicy stanowi 30-elementowy ciąg rosnący obejmujący 30 najbardziej podobnych wzorców dla danego obiektu. Oceniono stopień poprawności rozpoznawania w zależności od materiału językowego poddanego testowaniu, jak również stopień przydatności poszczególnych głosów jako wzorcowych oraz jako nadawców obiektów.

Wstęp.

Większość układów rozpoznających mowę zaadaptowanych jest na jeden głos (głos operatora) lub grupę dobranych głosów. Tę właściwość posiadały również w swej pierwotnej wersji modele automatycznego rozpoznawania wyrazów opracowane w ZFA IPPT PAN (Kubzdela [9], [10], [11]).

Wprowadzenie nowej wersji reprezentacji sygnału mowy w postaci obrazu binarnego wyznaczonego przy zastosowaniu tzw. funkcji wagi widma (Kubzdela [12]) zapewniło większą wszechstronność układu poprzez objęcie możliwością rozpoznawania szerszego kręgu głosów. Taki obraz bowiem w znacznym stopniu redukuje zróżnicowanie osobnicze. Stosując go

1 Praca wykonana w ramach CPBR 11.9.

w automatycznym rozpoznawaniu wypowiedzi wyrazów możliwe staje się posługiwanie się zbiorem wspólnych wzorców, reorezentujących różne głosy.

Celem niniejszej pracy było zbadanie efektywności rozpoznawania przez istniejący układ haseł sterujących wypowiadanych przez 10 osób (kobiety i mężczyzn). Uzyskane wyniki pozwolą ukierunkować dalsze prace nad skonstruowaniem użytkowych modeli rozpoznajników wypowiedzi wyrazów dla potrzeb osób niepełnosprawnych.

Aspekty zastosowawcze automatycznego rozpoznawania mowy.

1. Zalety i wady ARM.

W rozważaniach teoretycznych, jak i w pracach wdrożeniowych w zakresie łączności człowieka z maszyną przy użyciu głosu bierze się pod uwagę następujące okoliczności (por. np.Lea [13]):

1) Mowa jest najbardziej naturalną formą przekazywania informacji.

2) Przepływ informacji w sygnale mowy jest szybszy niż w innych modalnościach (w szczególności w porównaniu z zastosowaniem klawiatury maszyny do pisania lub klawiatury komputera).

3) Wykorzystuje się dodatkową możliwość sterowania, niezależnie od efektów motorycznych (rąk i nóg).

4) Ten sam nośnik informacji może być równocześnie wykorzystany z uwzględnieniem tak odbiorcy ludzkiego, jak i odbiornika nieożywionego.

5) Sygnał mowy może być nadawany w ciemności.

6) Sygnał mowy, jako zjawisko akustyczne, omija przeszkody w postaci przedmiotów wykonanych z materiałów odpornych na drgania w bardzo szerokim zakresie częstotliwości.

7) Powstaje nowy zakres rewalidacji i rehabilitacji medycznej w przypadku osób

a) niewidomych,

b) pozbawionych możliwości korzystania z efektorów motorycznych,

c) unieruchomionych czasowo.

8) Znaczne przyspieszenia (np. w rakietach) mają pomijalny

wpływ na wytwarzanie i przenoszenie sygnału mowy.

9) W niektórych przypadkach wymagających zabezpieczenia przed nieuprawnionym użytkownikiem zachodzi możliwość naturalnej identyfikacji nadawcy informacji.

10) Nadawanie informacji odbywa się z wykorzystaniem przetwornika o bardzo małych rozmiarach i małej masie (mikrofon) w porównaniu z klawiaturą.

11) Istnieje ewentualność nadawania informacji podczas ruchu nadawcy (np. w przypadku użycia spadochronu).

12) Powstaje możliwość zdalnego sterowania z wykorzystaniem abonenckiej sieci telefonicznej.

Przekazywanie informacji do maszyny (w ogólnym znaczeniu tego pojęcia, tj. z włączeniem także bardzo prostych urządzeń) za pomocą głosu ma określone wady i ograniczenia, do których należy zaliczyć następujące:

(A) Tak obecnie, jak i w dającej się przewidzieć przyszłości zakres informacji możliwej do przekazania urządzeniu "rozumiejącemu" mowę jest relatywnie wąski. Prace nad metodami rozpoznawania dowolnego tekstu w określonym języku nie przekroczyły jeszcze stadium teoretycznego.

(B) Sygnał mowy wykazuje bardzo znaczną zmienność pozasemantyczną, szczególnie w zakresie zróżnicowań międzyosobniczych. Sygnał ten zostaje, w celu automatycznego rozpoznawania, przetworzony na rozmaite, zależne od przyjętych założeń technicznych, parametry stanowiące funkcje czasowe. Jednakże nie wykryto dotychczas takich parametrów, które reprezentowałyby jedynie specyfikę głosu w odróżnieniu od takich, które przekazywałyby wyłącznie, albo prawie wyłącznie, informację lingwistyczną. Normalizacja parametrów akustycznych sygnału mowy względem cech indywidualnych głosu nadawcy lub - alternatywnie - zwielokrotnianie wzorców rozpoznawczych celem udostępnienia systemu dla różnych operatorów (ewentualnie dowolnego operatora), stanowią wciąż jeszcze znaczne utrudnienie w konkretnych rozwiązaniach.

(C) W normalnych warunkach aplikacyjnych (w odróżnieniu od laboratoryjnych) występują zakłócenia oraz (rzadziej) zniekształcenia, które zmniejszają poprawność automatycznego rozpoznawania. Częściowo trudność tą pokonuje się poprzez zastosowanie mikrofonów umieszczanych bardzo blisko ust lub

(rzadziej) laryngofonów.

2. Ogólna klasyfikacja systemów ARM.

Systemy ARM, tak laboratoryjne, jak i wdrożeniowe, sklasyfikować można z różnych punktów widzenia, np. metody analizy i przetwarzania sygnału oraz liczby i rodzajów ekstrahowanych parametrów. Należy poza tym zwrócić uwagę na następujące kryteria klasyfikacyjne:

(a) Dostosowanie do jednego lub bardzo ograniczonej liczby głosów w odróżnieniu od zastosowania do otwartego zbioru operatorów.

(b) Działanie systemu na pojedynczych, gramatycznie niezmiennych hasłach z ograniczonego słownika (*isolated word recognition*), na ciągach łącznie wymawianych, niezmiennych haseł (*connected word recognition*) oraz na mowie naturalnej (*continuous speech recognition*) ograniczonej tematycznie. Perspektywicznie rozpatruje się rozpoznawanie mowy spontanicznej (bez ograniczeń tematycznych i gramatycznych). W pierwszym przypadku - który zachodzi w referowanym w dalszych rozdziałach - doświadczeniu, sygnały sterujące mają postać oddzielnie wymówionych, pojedynczych wyrazów lub minimalnych składniowo fraz ("w górę", "do końca" itp.).

(c) Przyjęcie jako wstępnej (lub jedynej) jednostki rozpoznawczej: segmentu akustycznego (np. fragmentu sygnału o stałej rozciągłości czasowej, najczęściej 10 ms), segmentu fonetyczno-akustycznego, głoski (fonemu), półsylaby, sylaby, diady fonetycznej lub całego wyrazu (jednostek wyższego poziomu poza bardzo prostymi zestawieniami, por. wyż., nie stosuje się). W przypadku wykorzystania jednostek niższego rzędu rozpoznawanie jest dwu- lub więcejstopniowe (np. segment akustyczny → głoska → wyraz).

3. Bieżące i przyszłe wdrożenia.

Automatyczne rozpoznawanie mowy stosuje się - lub co najmniej przewiduje do rychłego zastosowania - generalnie ujmując - w trzech dziedzinach:

1) wprowadzanie danych,

2) sterowanie urządzeniami oraz

3) komunikacja (por.Ainsworth [1]).

Dotychczas wdrożone, lub bliskie wdrożenia systemy ASR

znalazły zastosowanie w:

(A) pracy biurowej,

(B) przemyśle,

(C) transporcie,

(D) gospodarstwie domowym,

(E) medycynie,

(F) urządzeniach użyteczności publicznej,

(G) wojskowości oraz

(H) rehabilitacji i rewalidacji medycznej, jako pomoc dla osób niepełnosprawnych.

Niniejsza praca związana jest z ostatnią z wymienionych sfer zastosowań (tj. (H)). W tym zakresie dotychczasowe osiągnięcia dotyczą:

1) sterowania wózkiem inwalidzkim - [2], [3], [16];

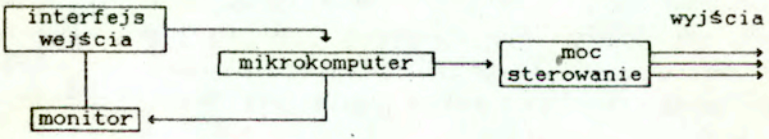
2) sterowania urządzeniami codziennego użytku - [3], [5] [7], [16];

3) wprowadzania tekstu do edytora - [4], [6], [14];

4) sterowania manipulatorem [15];

5) wprowadzania danych do komputera [8][2].

Tematycznie (ale nie z technicznego punktu widzenia) przedstawione w niniejszej pracy doświadczenia zbliżone są do pracy Dampera [5]. Chociaż jego system może, bez istotnych zmian, być użyty do innych zadań, został on specjalnie przewidziany dla celów sterowania wybranymi urządzeniami przez osoby całkowicie unieruchomione (na skutek choroby, operacji, wrodzonych lub nabytych stanów patologicznych itp.). System ten schematycznie przedstawia ryc.1.



Ryc.1. Schemat systemu ASR Dampera [5].

System Dampera [5] przewiduje kontrolne sprzężenie zwrotne przez monitor. Jeśli na monitorze pojawia się zapis poprawnego

[2] Większość wymienionych w tym miejscu danych bibliograficznych podajemy za [5].

rozpoznania, operator podaje wyraz *yes* powodujący uruchomienie/wyłączenie określonego urządzenia. Układ reaguje na wybranych siedem haseł: *heat, tv, radio, phone, cassette, bed, light*. Został on przetestowany, ale w pracy nie podano szczegółowych wyników. System jest dostrajany do głosu operatora, a szczególną wiarygodność wykazuje w rozpoznawaniu krytycznych wyrazów *yes* i *no*. Przedstawiona niżej praca jest koncepcyjną analogią pracy Dampera dostosowaną do języka polskiego. Słownik jest semantycznie podobny, natomiast rozwiązanie techniczne jest całkowicie odmienne i opiera się na koncepcjach jednego z współautorów (por. Kubzdela [9], [10], [11], [12]).

Materiał doświadczalny.

Zbiór haseł przygotowano z myślą o ich realnej przydatności w układach rozpoznających, służących osobom niepełnosprawnym. Objął on następujące wyrazy: *RADIO, DRZWI, WÓZEK, MAGNETOFON, TELEWIZOR, ŚWIATŁO, W GÓRĘ, W DÓŁ, TAK, NIE*. W doświadczeniu uwzględniono 10 głosów, przy czym połowę z nich stanowiły głosy kobiece, połowę męskie. Materiał językowy został utrwalony na taśmie magnetofonowej. Nagrania, wykonane w pomieszczeniu bezechowym, były powtarzane w odstępach kilkudniowych. Każda z osób wzięła udział w czterech sesjach nagraniowych, w rezultacie czego dysponowano czterema powtórzeniami każdego wyrazu. Trzy powtórzenia wchodziły w skład zbioru wzorców wypowiedzi, zaś czwarte powtórzenie w skład zbioru obiektów rozpoznawanych. Przy nagrywaniu wyrazów, odczytywanych z listy, utrzymywano równą głośność, intonację oraz umiarkowane tempo wypowiedzi. Pomiędzy kolejnymi wyrazami w zbiorze zachowywano 2-sekundowy odstęp. Przestrzegano starannej, lecz równocześnie naturalnej wymowy odczytywanych haseł.

Metoda rozpoznawania i opis doświadczenia.

Zastosowano metodę rozpoznawania opartą na przedstawieniu wypowiedzi wyrazu w formie obrazu binarnego. Obraz ten zostaje utworzony na podstawie kolejnych widm amplitudowych akustycznego sugnału wypowiedzi.

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

- pierwszej: WO+PM, będącej połączeniem metody wypukłości obwiedni widma i metody maskowania,
- drugiej: PM, będącej metodą maskowania.

Metody te przedstawiono szerzej w pracach: Kubzdela [9], [10], [11].

Poszczególne parametry wektora binarnego, odnoszące się do kolejnych punktów widma, otrzymują wartość 1 lub 0. W metodzie wypukłości obwiedni widma WO wartość ta jest wynikiem porównania lokalnych wypukłości obwiedni odpowiednio wygładzonego widma amplitudowego z wartościami progowymi, zależnymi od stopnia nachylenia tej obwiedni w poszczególnych punktach.

i-ty parametr b_i widma binarnego otrzymuje odpowiednią wartość binarną na podstawie następującego kryterium:

$$b_i = 1, \text{ jeśli } a_i - \frac{|a_{i+3} - a_{i-3}|}{2} \geq k_n \frac{a_{i+3} + a_{i-3}}{2}$$

$b_i = 0$, jeśli powyższy warunek nie jest spełniony.

a_{i-3}, a_i, a_{i+3} są rzędnymi wygładzonego widma amplitudowego w punktach i-3, i, i+3,

k_n jest współczynnikiem uzależniającym progową wartość wypukłości od nachylenia obwiedni widma.

Według metody maskowania PM o wartości parametru wektora binarnego decyduje położenie obwiedni widma względem tak zwanego poziomu maskowania, zależnego w danym punkcie widma od pola powierzchni pod obwiednią, w przedziale rozciągającym się z lewej strony tego punktu oraz od obranej wartości współczynnika maskowania. W tej metodzie

i-ty parametr b_i widma binarnego przyjmuje wartość zgodnie z następującym kryterium:

$$b_i = 1, \text{ jeśli } a_i > k_m \sum_{j=i}^{i-n} a_j \text{ dla } i>n, \text{ lub } a_i > 0 \text{ dla } i=n,$$

$b_i = 0$, jeśli powyższy warunek nie jest spełniony.

a_i oraz a_j oznaczają rzędne wygładzonego widma amplitudowego w

punktach i oraz j,

k_m jest współczynnikiem maskowania.

W pierwszej z dwóch metod binaryzacji zastosowanych w pracy użyte zostały równocześnie dwa powyższe kryteria, w drugiej natomiast jedynie drugie. Stosując dwie wartości współczynnika maskowania równe 1/16 i 1/32 wyznaczono za pomocą każdej z tych metod cztery różne obrazy binarne. Podczas wyznaczania wektora binarnego przeprowadzano jednocześnie redukcję liczby jego parametrów z 64 do 16, stosując zasadę podaną w pracy: Kubzdela [12].

Na ryc.2 zamieszczono w charakterze przykładu obrazy binarne wypowiedzi wyrazu "maius" (a) z pełną i (b) zredukowaną liczbą parametrów, wyznaczone metodą WO+PM przy czterech wartościach współczynnika maskowania równych 0, 1/32, 1/25, 1/16 oraz dla porównania spektrogram tej wypowiedzi (c) z czterostopniową skalą poziomów.

W zastosowanej 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 podobieństwa 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). Segmentem dwuelementowym nazwano dwie sąsiednie kolumny macierzy zero-jedynkowej, jaką stanowi obraz binarny wypowiedzi. Warunkiem istnienia wzajemnego podobieństwa dwóch segmentów jest spełnienie następującego kryterium:

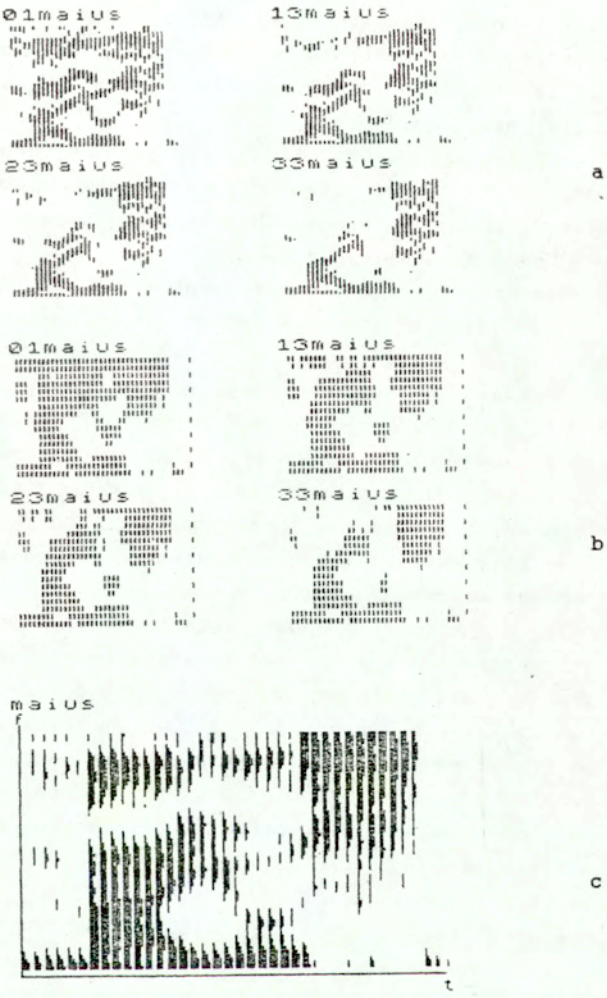
$$\frac{nz(1)}{\Sigma(1)_{st}} < k$$

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

(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



Ryc.2. Obrazy binarne (a,b) oraz 4-stopniowy spektrogram wypowiedzi wyrazu "maius". Przyjęte metody binaryzacji oraz wartości współczynnika maskowania oznaczono następująco:

01: WO, (WO+PM, W=0)
13: WO+PM, W=1/32
23: WO+PM, W=1/25
33: WO+PM, W=1/16

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.

Podczas znajdowania odpowiadających sobie segmentów w dwóch wzajemnie porównywanych obrazach binarnych brano pod uwagę istnienie dewiacji między rozkładami czasowymi elementów fonetycznych w wypowiedziach, których obrazy te dotyczyły. Czyniono to w następujący sposób:

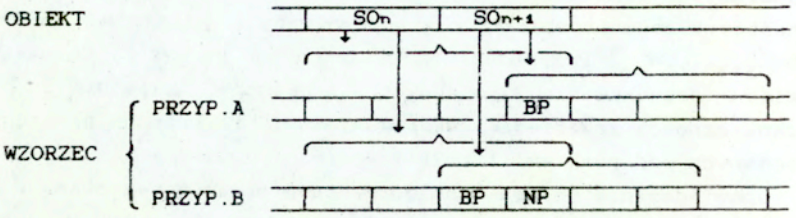
Dla kolejnego n-tego segmentu obiektu SO_n poszukiwano podobnego segmentu wzorca wśród jego czterech kolejnych segmentów $SW_{i+1},...SW_{i+4}$, począwszy od segmentu o numerze i+1. Dopuszczano, że segment taki może wystąpić po dwóch bezpośrednio go poprzedzających segmentach niepodobnych do SO_n. Spośród ewentualnie obecnych kilku segmentów podobnych *(P)* wybierany był segment najbardziej podobny *(BP)*. Zależnie od występujących okoliczności rozstrzygano, od którego segmentu wzorca począwszy będzie poszukiwany segment podobny do następnego segmentu obiektu SO_{n+1}.

1) Jeśli najbardziej podobny segment *(BP)* był czwartym w grupie (przypadek A na rys.3) lub jeśli zajmował miejsce niższe w grupie lecz segment po nim następujący nie wykazał podobieństwa z porównywanym segmentem obiektu (przypadek B na rys.3), wówczas dla następnego segmentu obiektu SO_{n+1} poszukiwany był podobny segment wzorca wśród czterech kolejnych segmentów, począwszy od segmentu *(BP)*. Przypadki te zilustrowano na rys.3. Segment *(BP)* może okazać się najbardziej podobny również do segmentu SO_{n+1}. Sytuacja taka może powtórzyć się wielokrotnie. Na podstawie analizy tego rodzaju przypadków ustalono, że może to niekiedy prowadzić do błędu w rozpoznawaniu. Dlatego wprowadzono ograniczenie liczby przypadków, w których dla kolejnych segmentów obiektu za najbardziej podobny uznany zostaje ten sam segment wzorca.
2) Jeśli nie zachodziły wymienione pod 1) okoliczności, odnoszące się do najbardziej podobnego segmentu wzorca, wówczas dla kolejnego segmentu obiektu poszukiwano podobnego segmentu wzorca wśród jego czterech kolejnych segmentów, poczynając od segmentu następującego po *(BP)*. Ten przypadek

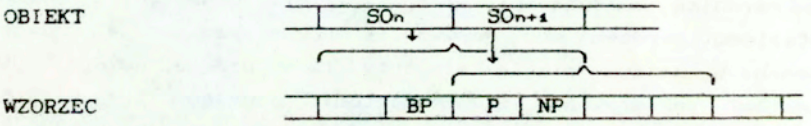
zilustrowano na rys.4.

3) Jeśli po raz pierwszy wśród segmentów $SW_{i+1},\ldots,SW_{i+4}$ nie znaleziono segmentu podobnego do segmentu obiektu, wówczas następny segment obiektu był porównywany z segmentami wzorca $SW_{i+2},\ldots,SW_{i+5}$. Przypadek ten zilustrowano na rys.5. Dalszy ciąg poszukiwania podobnych segmentów zależnie od okoliczności przebiegał według jednej z trzech procedur podanych pod punktami 1), 2) i 3).

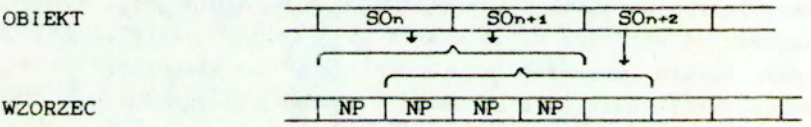
Doświadczenie polegało na porównaniu obrazów testowanych z obrazami wzorcowymi, dla każdej z czterech rozpatrywanych wersji obrazu binarnego oraz na uszeregowaniu uzyskanych wyników. Poszczególne wersje obrazu binarnego określane były przez metodę jego wyznaczania (WO+PM i PM) oraz przez wartość współczynnika maskowania (1/32 i 1/16). Oznaczenie O+PM zastąpiono skrótem WO. Porównanie przeprowadzono w trzech etapach. W każdym etapie elementy tego samego zbioru 100 obiektów porównywano z elementami jednego z trzech zbiorów liczących po 100 wzorców każdy. Zarówno obiekty, jak i każdy z trzech zbiorów wzorców pochodziły ze 100 wypowiedzi dziesięciu wyrazów wymówionych przez dziesięć głosów. Z wyników uzyskanych w każdym etapie dla każdego obiektu wybrano po 10 najmniejszych wartości i uszeregowano je w kolejności rosnącej począwszy od wartości najmniejszej. Dla każdego obiektu jednej wersji obrazu uzyskano tą drogą trzy 10-elementowe ciągi rosnące, które następnie złożono w jeden 30-elementowy ciąg rosnący. Ciągi takie, dotyczące wszystkich po kolei obiektów dla jednej wersji obrazów utworzyły tablicę o wymiarach 30×(100×4), której każdy wiersz składał się z czterech podwierszy. Elementami tej tablicy są odległości poszczególnych obiektów od wybranych wzorców oraz ich kody wyrazowe, głosowe i wzorcowe. Kod głosowy stanowiła pierwsza cyfra, a kod wyrazowy druga cyfra liczby dwucyfrowej. Kod wzorcowy zawarty był w numerze podwiersza tablicy. Podwiersz pierwszy dotyczył wyników z udziałem pierwszego zbioru wzorców. Podobnie podwiersze drugi i trzeci dotyczyły wyników z udziałem drugiego i trzeciego zbioru wzorców. Podwiersz czwarty zawierał liczby wyrażające odległość danego obiektu od najbliższych mu wzorców w poszczególnych zbiorach wzorców.



Rys.3.



Rys.4.



Rys.5.

Rys. 3 do 5. Zasady poszukiwania segmentu wzorca podobnego do kolejnego segmentu obiektu. Symbolicznie przedstawione wycinki obrazu obiektu i wzorca oznaczają porównywane segmenty. Klamra obejmuje grupę segmentów wzorca, wśród których poszukuje się segmentu najbardziej podobnego do danego segmentu obiektu, od którego poprowadzona jest strzałka.

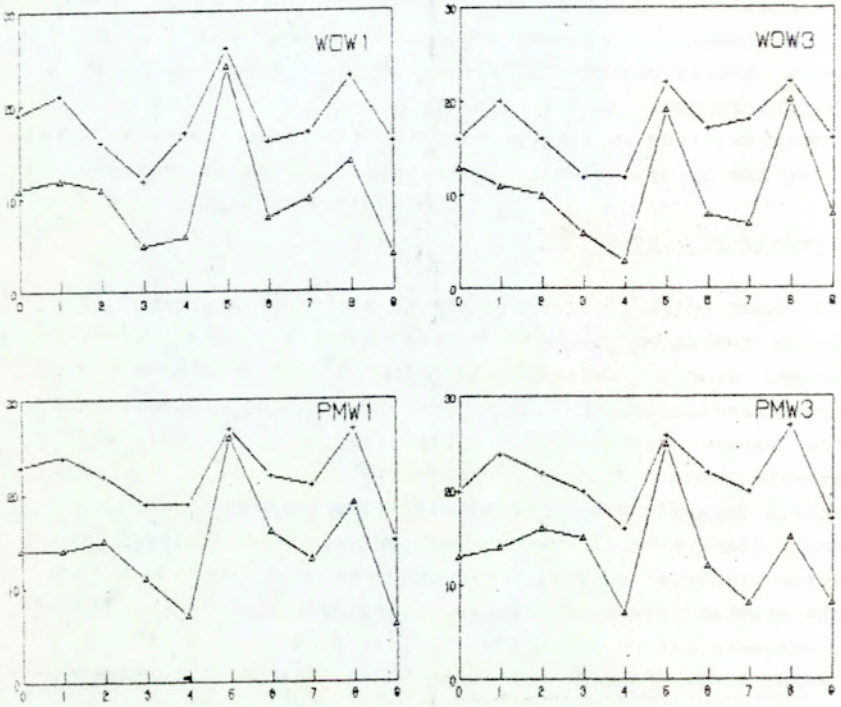
Ze względów technicznych trzeba było zrezygnować z wyznaczenia ciągu najmniejszych odległości w obrębie wzorca 300-elementowego i poprzestać na przeprowadzeniu porównań osobno w obrębie wzorców 100-elementowych, a następnie złożeniu wyniku zbiorczego. Wyniki dla tych obu procedur mogą być niejednakowe, lecz są zapewne na tyle wzajemnie zbliżone, że dają ten sam obraz zdolności rozpoznawczych badanego systemu.

Omówienie wyników.

Wartość odległości widniejąca na pierwszej pozycji w ciągu rosnącym, tworzącym dowolny wiersz tablicy, dotyczy obrazu binarnego wzorca, wykazującego największe podobieństwo do obrazu binarnego obiektu. Pojawienie się w tej pozycji kodu wzorca wyrazu testowanego, należącego do którejkolwiek z dziesięciu osób, oznacza rozpoznanie poprawne. W całym materiale doświadczalnym bardzo nieliczne okazały się przypadki błędnej identyfikacji wyrazu testowanego, wyrażającej się zajęciem pierwszej pozycji w wierszu przez kod obcego wzorca. Liczba niewłaściwie rozpoznanych obiektów, z ogólnej liczby 100, wyniosła dla wariantu WOW1: 3, dla WOW3: 2, dla PMW1: 4, dla PMW3: 1. Tym samym poprawność rozpoznawania testowanych haseł wynosi odpowiednio: wariant WOW1 - 97%, wariant WOW3 - 98%, PMW1 - 96%, PMW3 - 99%.

W idealnej sytuacji na wszystkich trzydziestu pozycjach dowolnego wiersza tablicy odległości powinny znaleźć się wzorce wypowiedzi stanowiących reprezentację wyrazu testowanego, jako najbardziej do niego podobne. Zgodnie z tym założeniem pojawienie się wzorca wyrazu innego, niż testowany, na pozycji od 2 do 30, należy również traktować jako błąd, chociaż o mniejszym znaczeniu, niż wystąpienie błędnej identyfikacji na pierwszej pozycji w wierszu. Przyjęto dwa kryteria poprawności dla tego rodzaju błędów: (1) liczba poprawnych rozpoznań do miejsca wystąpienia w wierszu po raz pierwszy kodu obcego wzorca, (2) ogólna liczba poprawnych rozpoznań w całym wierszu.

Ryc.6 podaje rozkład wyników rozpoznawania poszczególnych wyrazów z uwzględnieniem obu wymienionych kryteriów. Oddzielnie spo ządzono wykresy dla różnych wariantów obrazów. Maksymalna liczba poprawnych rozpoznań dla każdego wyrazu, traktowanego



Δ krzywa dolna – liczba poprawnych rozpoznań poprzedzających pierwszy błąd wśród uszeregowanych typowań, tworzących poszczególne wiersze tablicy identyfikacji.

+ krzywa górna – liczba poprawnych rozpoznań w całym wierszu tablicy

0 – radio
1 – drzwi
2 – wózek
3 – magnetofon
4 – telewizor
5 – światło
6 – w górę
7 – w dół
8 – tak
9 – nie

Rys.6. Wyniki rozpoznawania poszczególnych wyrazów testowanych z uwzględnieniem czterech wariantów obrazów i dwóch kryteriów poprawności. Numeracja na osi poziomej odpowiada liczbowemu oznaczeniu poszczególnych wyrazów.

jako jednostkowy obiekt i określonego kryterium wynosi 30 i odnosi się do sytuacji, gdy wszystkie 30 wartości w wierszu należy do wzorców reprezentujących wyraz testowany. Wartości umieszczone na wykresie stanowią średnie arytmetyczne obliczone na podstawie wyników rozpoznawania danego wyrazu wymówionego przez 10 osób. Rozkłady wyników uzyskanych z zastosowaniem dwóch kryteriów oceny wykazują wspólne tendencje; punkty ekstremalne widoczne na wykresach odnoszą się na ogół do tych samych wyrazów. Stwierdzona zgodność dotyczy przyjętych kryteriów oceny poprawności dla wszystkich czterech zastosowanych wariantów obrazów binarnych. Najmniejszą liczbę błędnych identyfikacji w wierszu stwierdzono w przypadku rozpoznawania wyrazów: *ŚWIATŁO*, *TAK*. Największą liczbę błędów popełniono przy rozpoznawaniu wyrazów: *NIE*, *TELEWIZOR*, *MAGNETOFON*. Znamienne jest, iż błędne rozpoznania odnotowane w kilku przypadkach na pierwszej pozycji w wierszu (p. str.15), które stanowią podstawę oceny zdolności rozpoznawczych systemu, dotyczyły wyłącznie tych samych trzech wyrazów: *NIE*, *TELEWIZOR*, *MAGNETOFON*.

Sporządzono matryce błędnych rozpoznań, odnotowanych na którejkolwiek pozycji we wszystkich wierszach tablicy. Czterem częściom tablicy, odnoszących się do określonych wersji obrazów binarnych, odpowiadają cztery matryce błędów (tab.1). Najbardziej efektywna, ze względu na najmniejszą liczbę błędnych rozpoznań w całej tablicy, okazała się wersja PMW1. Globalna liczba stwierdzonych błędów wyniosła odpowiednio: dla wariantu WOW1: 39%, dla WOW3:, 42%, dla PMW1: 27%, dla PMW3: 29%. Tendencje zauważalne w rozkładzie wyników powtarzają się we wszystkich czterech matrycach. Najczęściej pojawiający się błąd polega na rozpoznaniu wyrazu *TELEWIZOR* jako *RADIO* (dla czterech wersji obrazów w kolejności jak wyżej: 30%, 36%, 34%, 42% przypadków). Uderza jednak fakt, że błędnych identyfikacji polegających na odebraniu wyrazu *RADIO* jako *TELEWIZOR* odnotowano bez porównania mniej (w kolejności jak wyżej: 21%, 11%, 12%, 10%). Nie można więc mówić, że oba te wyrazy wzajemnie się mieszają w procesie rozpoznawania, gdyż istotną rolę odgrywa tu fakt, który z wyrazów jest obiektem rozpoznawanym, zaś który wzorcem, do którego obraz obiektu się przyrównuje.

Tab.1. Matryce błędów popełnionych przy rozpoznawaniu wyrazów, dla poszczególnych wariantów obrazów binarnych (w %).

A) WOW1

Wyraz testowany	Identyfikacja błędna										
	0	1	2	3	4	5	6	7	8	9	Σ
0	-	0,7	0,3	8,7	21,3	-	0,7	3,7	0,7	1,0	37,1
1	-	-	1,7	0,7	5,0	16,0	-	-	-	5,7	29,1
2	2,7	3,7	-	6,7	1,3	8,7	8,3	0,3	5,7	9,3	46,7
3	23,7	2,7	0,3	-	5,7	2,7	5,0	12,7	4,0	3,7	60,5
4	30,3	8,0	1,0	1,3	-	0,3	0,3	1,0	-	1,0	43,2
5	0,3	4,0	-	0,3	-	-	0,7	3,3	3,3	2,0	13,9
6	4,7	2,3	10,7	2,0	1,0	2,0	-	9,0	3,0	10,7	45,4
7	2,3	2,7	5,7	1,7	0,7	1,0	28,0	-	0,7	1,0	42,8
8	14,0	-	-	5,3	2,3	0,3	-	-	-	2,0	23,9
9	11,7	11,7	3,7	2,7	9,7	3,0	1,0	0,7	4,3	-	48,5
										$\bar{x}$ =	39,1

0 - *radio*, 1 - *drzwi*, 2 - *wózek*, 3 - *magnetofon*, 4 - *telewizor*, 5 - *światło*, 6 - *w górę*, 7 - *w dół*, 8 - *tak*, 9 - *nie*.

B) WOW3

Wyraz testowany	Identyfikacja błędna										
	0	1	2	3	4	5	6	7	8	9	Σ
0	-	-	2,0	13,3	11,0	0,7	4,7	9,0	0,7	0,7	42,1
1	0,3	-	1,3	0,3	4,0	20,7	0,3	1,7	-	5,7	34,3
2	1,0	-	-	11,0	1,3	3,3	15,0	2,7	6,0	3,7	44,0
3	20,7	-	0,7	-	1,3	1,7	1,6	17,3	0,7	1,0	59,4
4	36,0	4,7	4,7	4,0	-	0,7	4,3	3,3	0,7	1,3	59,7
5	3,3	2,7	-	2,3	-	-	6,3	9,0	1,7	0,7	26,0
6	5,7	-	5,7	2,7	0,7	-	-	24,0	24,0	1,7	42,2
7	0,7	-	5,7	1,7	0,7	0,3	27,7	-	-	3,3	40,1
8	15,7	-	-	8,0	3,3	-	-	-	-	-	27,0
9	12,3	8,7	5,7	6,0	4,7	1,0	2,7	1,3	4,7	-	47,1
										$\bar{x}$ =	42,2

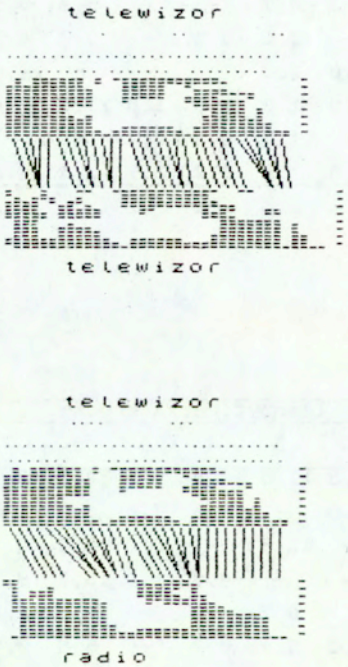
C) PMW1

Wyraz testowany	Identyfikacja błędna										
	0	1	2	3	4	5	6	7	8	9	Σ
0	–	0,3	0,3	5,3	12,3	0,3	0,3	2,3	–	9,7	21,8
1	0,7	–	1,3	0,3	5,7	7,0	–	0,3	–	6,0	21,3
2	0,3	–	–	7,7	–	1,7	6,3	0,3	2,7	7,7	26,7
3	11,0	–	0,3	–	4,7	0,7	9,0	11,0	–	1,3	38,0
4	34,0	–	0,3	0,7	–	–	0,3	0,3	–	0,3	35,9
5	0,7	0,3	–	3,0	0,3	–	0,7	4,7	–	–	9,7
6	2,0	–	5,0	2,0	–	–	–	14,0	0,3	2,7	26,0
7	3,7	1,3	3,0	1,0	1,0	1,0	19,0	–	–	1,3	31,3
8	8,0	–	–	1,7	0,3	0,3	–	–	–	–	10,3
9	11,3	11,7	8,3	1,0	8,7	0,7	2,0	4,0	1,0	–	48,7
											$\bar{x}$ = 27,0

D) PMW3

Wyraz testowany	Identyfikacja błędna										
	0	1	2	3	4	5	6	7	8	9	Σ
0	–	–	2,3	11,0	10,0	–	4,0	3,3	–	1,7	32,3
1	0,3	–	1,0	–	5,3	7,3	0,3	1,0	–	6,3	21,5
2	1,0	–	–	12,3	–	0,3	8,0	0,7	1,3	3,0	26,6
3	6,7	–	–	–	4,0	1,0	13,7	8,3	–	–	33,7
4	42,3	–	1,7	0,7	–	–	2,3	1,3	–	–	48,3
5	1,3	–	0,3	5,7	0,3	–	5,0	2,0	–	–	14,6
6	0,7	–	2,3	1,7	1,7	–	–	17,7	–	2,0	26,1
7	1,7	–	2,0	3,3	0,3	0,3	21,0	–	–	5,0	33,6
8	7,3	–	1,3	2,0	1,0	–	–	–	–	–	11,6
9	7,7	4,3	10,0	2,7	5,3	0,3	6,7	5,0	0,7	–	42,7
											$\bar{x}$ = 29,1

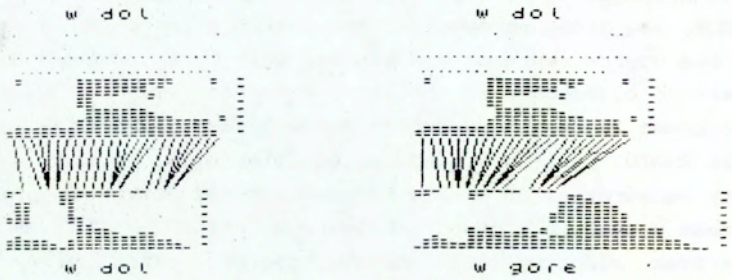
Znaczna liczba błędnych identyfikacji wyrazu *TELEWIZOR* jako *RADIO* stanowi wynik dość nieoczekiwany, ze względu na odmienną strukturę fonetyczną oraz długość obu haseł. Zamieszczona na ryc.7 ilustracja pozwala uzasadnić tego rodzaju decyzje podjęte przez układ.



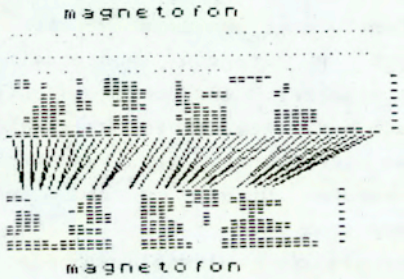
Ryc.7. Porównanie obrazów binarnych wypowiedzi: TELEWIZOR - TELEWIZOR oraz TELEWIZOR - RADIO.

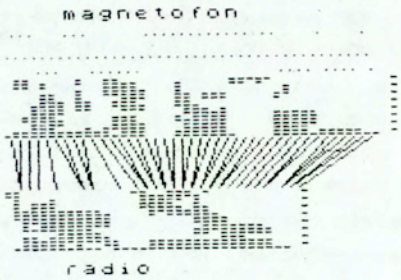
Rycina zawiera dwie pary obrazów binarnych, z których pierwsza przedstawia porównanie obiektu *TELEWIZOR* z wzorcem *TELEWIZOR*, zaś druga porównanie tego samego obiektu z wzorcem *RADIO*. Oba wzorce reprezentują ten sam głos i oba uzyskały taką samą wartość odległości od obiektu, wynoszącą 21, co pozwala identyfikować testowaną wypowiedź zarówno jako *TELEWIZOR*, jak też jako *RADIO*. Linie poprowadzone od obiektu do wzorca łączą segmenty najbardziej do siebie podobne spośród przyrównywanych, wyznaczone w oparciu o zasady opisane na str.13 - 15. Górny rząd kropek widoczny nad obrazem obiektu odnosi się do segmentów obiektu, które nie znalazły odpowiadających sobie segmentów we wzorcu, spełniających warunek podobieństwa w oparciu o przyjęte kryterium (p. str.11). Dolny rząd kropek odnosi się do segmentów, które spełniły warunek podobieństwa. Para *TELEWIZOR - TELEWIZOR* wykazuje brak podobieństwa jedynie w odniesieniu do kilku segmentów. Para *TELEWIZOR - RADIO* pomimo różnic w strukturze fonetycznej wykazuje wyraźne podobieństwo obrazów binarnych już w trakcie pobieżnej obserwacji. Rozmieszczenie kropek widocznych nad obrazem obiektu pozwala stwierdzić, iż w ramach przyjętego progu określającego granice podobieństwa, znaczne fragmenty obu wypowiedzi zostały zakwalifikowane jako podobne. Dotyczy to fragmentów /ele/ oraz /a/, /i/ oraz /j/, /or/ oraz /o/.

Następnym z kolei często pojawiającym się błędem była identyfikacja hasła *W DÓŁ* jako *W GÓRĘ* (w kolejności wariantów obrazów jak wyżej 28%, 28%, 19%, 21% przypadków). Z tego rodzaju błędem można było się liczyć ze względu na zbliżoną strukturę fonetyczną obu haseł. Na ryc.8 zamieszczono dwie pary obrazów binarnych *W DÓŁ - W DÓŁ* oraz *W DÓŁ - W GÓRĘ*, w których występuje ten sam obiekt, zaś uwzględnione wzorce reprezentują ten sam głos. W obu przypadkach wartość odległości pomiędzy obiektem i wzorcem wynosi 19. Podobieństwo obrazów binarnych w parze *W DÓŁ - W DÓŁ* jest bardzo wyraźne. Zaledwie cztery segmenty nie spełniły warunku podobieństwa w oparciu o zastosowane kryterium. Równie nieliczne okazały się segmenty niepodobne w parze *W DÓŁ - W GÓRĘ*, jednakże w przypadku tej pary uderza ograniczenie obszaru poszukiwań odpowiadających sobie segmentów obiektu i wzorca do początkowego fragmentu wzorca. Druga część obrazu, stanowiąca



Ryc.8. Porównanie obrazów binarnych wypowiedzi: W DÓŁ - W DÓŁ oraz W DÓŁ - W GÓRĘ.





Ryc.9. Porównanie obrazów binarnych wypowiedzi: MAGNETOFON - MAGNETOFON - MAGNETOFON oraz MAGNETOFON - RADIO.

reprezentację fragmentu /rẽ/, nie została objęta poszukiwaniem, co nie przeszkodziło uzyskać wysoki stopień podobieństwa w obrębie pary. Wyjaśnienie tego faktu dostarcza prezentowana ilustracja. Głoska /w/ wymawiana w wygłosie wyrazu *W DÓŁ* zlała się z poprzedzającą samogłoską /u/ dając w efekcie taki sam obraz, jak samogłoska /u/ w wyrazie *W GÓRĘ*, jedynie nieco w stosunku do niego wydłużony. Dzięki temu etap poszukiwania zakończył się we wzorcu na samogłosce /u/, bez możliwości kontynuowania go na pozostałej części wzorca[3]. Ograniczenie porównywania obrazów do ich bardzo podobnych fragmentów doprowadziło do identyfikacji obiektu jako tożsamego z wzorcem.

Błędy zachodzące w odwrotnym kierunku, a polegające na odebraniu hasła *W GÓRĘ* jako hasła *W DÓŁ* występowały również bardzo licznie (9%, 24%, 14%, 18% przypadków w kolejności jak wyżej). W odniesieniu do tej pary można więc mówić o wzajemnym mieszaniu wyrazów w procesie automatycznego rozpoznawania, uwarunkowanym ich podobieństwem fonetycznym.

Częste błędy dotyczą rozpoznania wyrazu *MAGNETOFON* jako *RADIO*, ale tylko w odniesieniu do metody WO (24%, 21% przypadków). *RADIO* jako *MAGNETOFON* odbierano znacznie rzadziej (9% i 13% przypadków). Na ryc.9 zamieszczono dwie pary obrazów binarnych,w których występuje ten sam obiekt: *MAGNETOFON*. Wzorce: *MAGNETOFON* i *RADIO* zostały wymówione przez różne głosy, przy czym wzorzec *RADIO* okazał się być bardziej podobnym do obiektu (wartość odległości wynosi 55), niż wzorzec *MAGNETOFON* (wartość odległości 63). Obie pary zawierają dużą liczbę segmentów, które nie spełniają warunku podobieństwa w oparciu o przyjęte kryterium, przy czym większa ich liczba wystąpiła w parze: *MAGNETOFON - MAGNETOFON*, niż *MAGNETOFON - RADIO*, co w tym przypadku przesądza o błędnej identyfikacji.

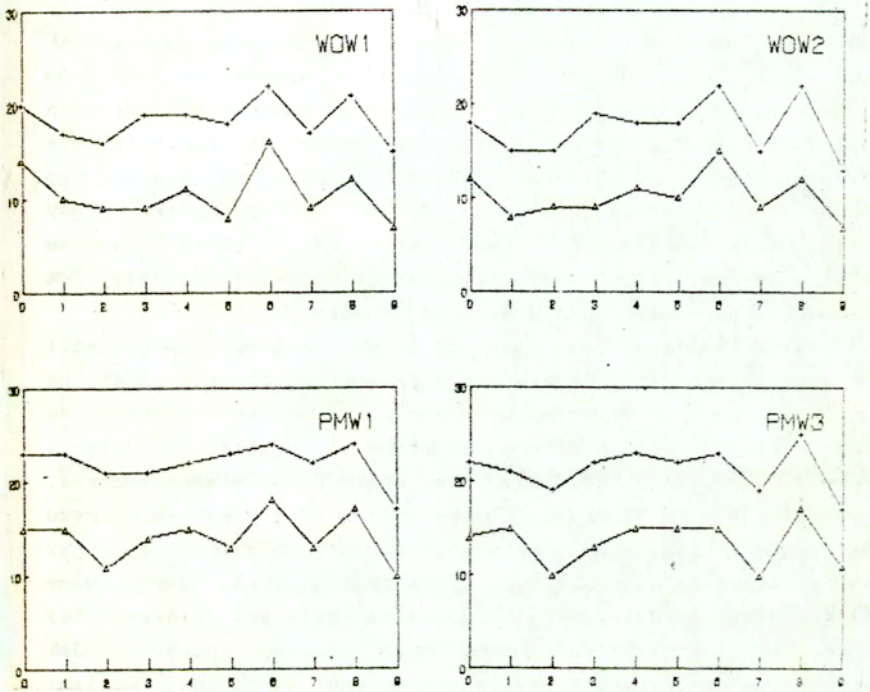
Jak wykazują dane z tab.1, wśród stwierdzonych błędów nie występowały w ogóle niektóre substytucje. Np. przy zastosowaniu dowolnej wersji obrazów nie zdarzyło się, aby hasło *TAK* zostało zidentyfikowane na którejkolwiek pozycji w wierszu jako hasło *W GÓRĘ*, zaś para *TAK* i *DRZWI* nie była nigdy mieszana zarówno wówczas, gdy obiektem rozpoznawanym był wyraz *TAK*, a wzorcem

[3] Dla obecnych doświadczeń pominięto procedurę oceny podobieństwa obrazów na podstawie długości reszty wzorca, nie objętej porównaniem z obiektem.

DRZWI, jak i na odwrót: obiektem *DRZWI*, a wzorcem *TAK*.

Liczba poprawnych identyfikacji stwierdzonych w tablicy odległości wykazuje, jak z powyższego wynika, wyraźne uzależnienie od materiału językowego - poszczególne hasła uzyskiwały różny stopień poprawności rozpoznawania. Dla oceny efektywności układu istotne znaczenie posiada również określenie ewentualnego wpływu zróżnicowania osobniczego. Wyniki rozpoznawania wyrazów dostarczonych przez poszczególnych mówców zamieszczono na ryc.10. Na każdym z czterech wykresów odnoszących się do różnych wariantów obrazów binarnych naniesiono liczbę poprawnych rozpoznań, wyznaczoną w oparciu o dwa kryteria oceny (p. str.15). Maksymalna liczba poprawnych rozpoznań każdego wyrazu, reprezentującego określony głos, wynosi 30 (39 wartości w wierszu). Wartości umieszczone na wykresach stanowią średnie arytmetyczne odnoszące się do wszystkich dziesięciu wyrazów wymówionych przez daną osobę. Generalnie należy stwierdzić, że wyniki rozpoznawania w zależności od głosu wymawiającego testowane hasła, wykazują mniejsze zróżnicowanie, niż w zależności od materiału językowego. Wniosek ten posiada ważne znaczenie dla oceny praktycznej przydatności badanego systemu. Spośród dziesięciu analizowanych głosów nieznaczną przewagę nad pozostałymi wykazują jedynie osoby: MK oraz PS. Wymawiane przez nie hasła poddane testowaniu uzyskiwały stosunkowo najwięcej poprawnych rozpoznań w całej tablicy. Z kolei najmniej poprawnych identyfikacji pojawiało się, gdy obiekty rozpoznawane pochodziły od osoby UM. Rozpiętość pomiędzy największą oraz najmniejszą liczbą poprawnych rozpoznań dla różnych głosów wyniosła w zależności od przyjętego kryterium oraz zastosowanej wersji obrazów od 20 do 30%.

Wbrew oczekiwaniom , rozpoznawanie haseł nie zawsze przebiega najlepiej w oparciu o wzorce pochodzące od osoby wypowiadającej wyraz testowany. Na pierwszej pozycji w wierszu (największe podobieństwo do obiektu) równie często pojawiała się wypowiedź należąca do osoby, od której pochodził rozpoznawany aktualnie obiekt, jak wypowiedź należąca do którejkolwiek z dziewięciu pozostałych osób. W wielu przypadkach rozpoznawanie przebiegało więc lepiej na cudzych wzorcach, co wskazuje na znaczne uniezależnienie układu od



Δ krzywa dolna – liczba poprawnych rozpoznań poprzedzających pierwszy błąd wśród uszeregowanych typowań, tworzących poszczególne wiersze tablicy identyfikacji

\+ krzywa górna – liczba poprawnych rozpoznań w całym wierszu tablicy

0 – JI m(ęski)	5 – HK m
1 – LR ż(eński)	6 – MK ż
2 – WI m	7 – MS m
3 – GD ż	8 – PS ż
4 – PD m	9 – UM ż

Rys.10. Wyniki rozpoznawania wyrazów dostarczonych przez poszczególne głosy z uwzględnieniem czterech wariantów obrazów i dwóch kryteriów poprawności. Numeracja na osi poziomej odpowiada liczbowemu oznaczeniu poszczególnych głosów.

głosu operatora. Najwyższy stopień podobieństwa pomiędzy obiektem i wzorcami, reprezentującymi ten sam głos występował najczęściej u osoby JI - 83% wśród wszystkich obrazów binarnych, które znalazły się na pierwszej pozycji w wierszach odpowiadających obiektom JI, stanowią wzorce JI (łącznie dla czterech wariantów obrazów). Dla głosów MK i PS stwierdzono po 75% takich przypadków, dla HK - 73%. Niektóre głosy zdecydowanie preferowały obce wzorce. Np. wśród wzorców najbardziej zbliżonych do rozpoznawanych obiektów zaledwie 28% stanowią własne repetycje w przypadku osoby UM.

Tab.2 podaje dla każdego głosu, którego wypowiedzi podlegały testowaniu, łączną liczbę poprawnych rozpoznań na dowolnej pozycji wiersza w oparciu o wzorce należące do poszczególnych osób. Liczby umieszczone na przekątnej dotyczą poprawnych rozpoznań uzyskanych w oparciu o własne wzorce. Maksymalna liczba dla każdej osoby wynosi 30 - w każdym wierszu tablicy identyfikacji teoretycznie powinny pojawić się trzy wzorce należące do osoby, która dostarczyła rozpoznawany obiekt, zaś każdej osobie odpowiada dziesięć wierszy tej tablicy przy uwzględnieniu określonego wariantu obrazów. Jak wynika z danych, tylko w jednym przypadku (głos HK, wariant PMW3) znalazły się w określonej części tablicy wszystkie wzorce reprezentujące głos, którego wypowiedzi testowano. W pozostałych przypadkach nie wszystkie własne wzorce mieściły się w liczbie trzydziestu najbardziej zbliżonych do obiektu obrazów binarnych. Liczby wyróżnione w tabelach posiadają wartości wyższe lub równe liczbie leżącej na przekątnej. Dotyczą przypadków, gdy liczba poprawnych rozpoznań uwzględnionych w tablicy następowała równie często lub częściej w oparciu o obce wzorce, niż w oparciu o wzorce reprezentujące ten sam głos, co obiekt. Sytuacja taka najczęściej zachodziła w przypadku głosu PD.

Wyniki eksperymentu świadczą o tym, że wykorzystanie wspólnego dla dziesięciu osób zbioru wzorców w niczym nie obniżyło zdolności rozpoznawczych testowanego układu, gdyż często identyfikacja przebiegała lepiej w oparciu o wzorce pochodzące od innego głosu, niż obiekt.

Średnie wartości liczby poprawnych rozpoznań zamieszczone w ostatnim wierszu tabeli (tab.2) pozwalają zorientować się w

Tab.2. Liczba poprawnych rozpoznań w oparciu o wzorce należące do różnych głosów.

A) WOW1

Głos testowany	Głos wzorcowy										
	JI	LR	WI	GD	PD	HK	MK	MS	PS	UM	x
JI	29	18	24	21	17	25	17	22	13	17	20,3
LR	20	27	22	18	13	13	14	12	12	17	15,8
WI	19	17	26	14	16	16	6	18	6	17	15,5
GD	21	19	27	25	12	20	17	15	12	18	18,6
PD	27	17	26	13	20	24	13	22	12	16	19,0
HK	22	20	22	17	16	27	14	22	6	11	17,7
MK	19	27	26	24	11	23	26	19	22	26	22,3
MS	26	20	26	11	16	21	11	24	1	11	16,7
PS	20	22	21	22	10	23	22	15	27	25	20,7
UM	16	20	19	16	11	14	12	11	9	24	15,2
$\bar{x}$	21,9	20,7	23,9	18,1	14,2	20,6	15,2	16,0	12,0	18,2	18,3

Głosy męskie: JI, WI, PD, HK, MS.

Głosy kobiece: LR, GD, MK, PS, UM.

B) WOW3

Głos testowany	Głos wzorcowy										
	JI	LR	WI	GD	PD	HK	MK	MS	PS	UM	$\bar{x}$
JI	27	19	28	10	18	16	9	27	9	13	17,7
LR	17	28	15	14	10	12	12	18	8	17	15,1
WI	20	14	27	7	19	13	7	24	5	12	14,8
GD	17	23	19	24	10	20	18	19	17	21	18,8
PD	26	17	28	13	21	21	9	26	5	14	18,0
HK	20	20	23	10	21	26	13	23	6	16	17,8
MK	15	29	21	21	11	26	27	19	21	25	21,5
MS	24	18	22	3	21	11	9	27	3	10	14,8
PS	18	28	20	26	10	24	26	20	28	24	22,4
UM	9	20	13	9	10	9	10	17	6	23	12,6
$\bar{x}$	19,3	21,6	21,6	13,7	15,1	17,9	14,0	22,0	10,8	17,5	17,4

C) PMW1

Głos testowany	Głos wzorcowy										$\bar{x}$
	JI	LR	WI	GD	PD	HK	MK	MS	PS	UM	
JI	29	20	26	21	19	24	20	27	25	19	23,0
LR	20	29	22	23	13	23	23	22	23	27	22,5
WI	23	18	28	20	17	21	14	27	16	23	20,7
GD	21	19	23	26	12	25	21	23	15	25	21,0
PD	26	16	22	19	21	25	22	25	21	24	22,1
HK	22	27	22	20	20	28	25	28	15	25	23,2
MK	26	24	26	21	17	24	28	19	25	29	23,9
MS	28	18	28	18	16	23	21	27	15	23	21,7
PS	27	25	23	19	20	28	24	22	29	23	24,0
UM	18	19	19	17	11	14	14	18	11	28	16,9
$\bar{x}$	24,0	21,5	23,9	20,4	16,6	23,5	21,2	23,8	19,5	24,6	21,9

D) PMW3

Głos testowany	Głos wzorcowy										$\bar{x}$
	JI	LR	WI	GD	PD	HK	MK	MS	PS	UM	
JI	29	22	27	18	20	22	17	26	17	21	21,9
LR	15	27	23	20	14	19	20	20	21	26	20,5
WI	24	18	29	15	17	19	13	22	12	21	19,0
GD	16	24	24	27	15	22	25	22	25	23	22,3
PD	28	18	25	19	26	25	23	25	14	24	22,7
HK	24	25	26	20	18	30	16	26	15	23	22,3
MK	21	24	20	25	16	24	28	22	25	26	23,1
MS	21	20	26	16	17	20	13	29	10	18	19,0
PS	25	26	26	25	14	28	28	20	29	26	24,7
UM	17	19	16	17	12	14	13	21	15	26	17,0
$\bar{x}$	22,0	22,3	24,2	20,2	16,9	22,3	19,6	23,3	18,3	23,4	21,3

przydatności określonych głosów, jako dostarczycieli wzorców. Jedynie dwa głosy: PD i PS wykazują zdecydowane odstępstwo od pozostałych - w całym materiale najmniej było przypadków tego rodzaju, że poprawna identyfikacja przebiegała w oparciu o wzorce pochodzące od tych osób. Prawdopodobnie przyczynę tego stanowi fakt, że wypowiedzi dostarczone przez PD i PS posiadają nieco wolniejsze tempo w porównaniu z wzorcami pochodzącymi od innych osób. Natomiast pozostałe głosy nie wykazują wyraźnych zróżnicowań jako głosy wzorcowe, zwłaszcza w odniesieniu do metody PM - poprawne rozpoznania uzyskiwano w takim samym stopniu w oparciu o wzorce należące do różnych głosów.

Dobre wyniki uzyskane przy rozpoznawaniu obiektów pochodzących od określonego głosu nie zawsze oznaczają, że wzorce dostarczone przez ten głos decydują o poprawnej identyfikacji w sytuacji, gdy wykorzystuje się wspólny dla różnych mówców zbiór wzorców. Przykładem tego jest głos PS, który wykazał się wysokim stopniem poprawności rozpoznawania w oparciu o reprezentujące go wzorce.

Na podstawie danych z tab.2 można wyłonić najbardziej uniwersalne spośród testowanych dziesięciu głosów, tzn. takie, które pozwoliły uzyskać nalepsze wyniki rozpoznawania zarówno jako dostarczyciele wzorców, jak i obiektów. W tym celu w tab.3 zestawiono uśrednione wartości leżące w brzegowej kolumnie i wierszu tab.2, uszeregowane w kolejności rosnącej.

Kolejność głosów w tab.3, rozpatrywanych oddzielnie jako głosy wzorcowe oraz głosy dostarczające testowanych obiektów, nie jest jednakowa. Można jednak uznać, że najbardziej wszechstronnym głosem, który uzyskiwał relatywnie najlepsze wyniki zarówno jako wzorcowy, jak dostarczający obiekty, okazał się głos JI, po nim głos HK.

Jak wynika z danych z tab.2, płeć osoby dostarczającej wzorce miała znaczenie dla wyników rozpoznawania wyrazów wymawianych przez głosy męskie. Obiekty reprezentujące którykolwiek głos męski uzyskiwały największą liczbę poprawnych identyfikacji w całej tablicy odległości w oparciu o wzorce męskie. Natomiast w odniesieniu do głosów kobiecych nie stwierdzono, aby zachodziła preferencja wzorców żeńskich.

Analiza wyników zamieszczonych w tab.2 pozwala uchwycić związki, jakie zachodzą pomiędzy niektórymi głosami. W

Tab.3. Średnia liczba poprawnych rozpoznań w całej tablicy identyfikacji z uwzględnieniem obiektów oraz wzorców dostarczonych przez każdy głos.

WOW1				WOW3			
Głos wzorcowy		Głos testowany		Głos wzorcowy		Głos testowany	
WI	23,9	MK	22,3	MS	22,0	PS	22,4
JI	21,9	PS	20,7	LR	21,6	MK	21,5
LR	20,7	JI	20,3	WI	21,6	GD	18,8
HK	20,6	PD	19,0	JI	19,3	PD	18,0
UM	18,2	GD	18,6	HK	17,9	HK	17,8
GD	18,1	HK	17,7	UM	17,5	JI	17,7
MS	18,0	MS	16,7	PD	15,1	LR	15,1
MK	15,2	LR	15,8	MK	14,0	WI	14,8
PD	14,2	WI	15,5	GD	13,7	MS	14,8
PS	12,0	UM	15,2	PS	10,8	UM	12,6

PMW1				PMW3			
Głos wzorcowy		Głos testowany		Głos wzorcowy		Głos testowany	
UM	24,6	PS	24,0	WI	24,2	PS	24,7
JI	24,0	MK	23,9	UM	23,4	MK	23,1
WI	23,9	HK	23,2	MS	23,3	PD	22,7
MS	23,9	JI	23,0	LR	22,3	GD	22,3
HK	23,5	LR	22,5	HK	22,3	HK	22,3
LR	21,5	PD	22,1	JI	22,0	JI	21,9
MK	21,2	MS	21,7	GD	20,2	LR	20,5
GD	20,4	GD	21,0	MK	19,6	WI	19,0
PS	19,5	WI	20,7	PS	18,3	MS	19,0
PD	16,6	UM	16,9	PD	16,9	UM	17,0

przypadkach, gdy obiekty dostarczane przez jeden głos są najlepiej identyfikowane na wzorcach drugiego głosu i na odwrót, obiekty pochodzące od drugiego głosu preferują wzorce pierwszego głosu, można mówić o bliskości tych głosów, wyrażającej się wysokim stopniem podobieństwa ich obrazów binarnych. Tego rodzaju sytuacja zachodzi dla grupy głosów męskich: JI, WI, HK, MS. Poza obrębem grupy znalazł się głos

męski: PD, od którego pochodzące obiekty wykazują co prawda bliższy związek z wzorcami męskimi, niż z kobiecymi, jednakże reprezentujące go wzorce nie są preferowane przez pozostałe głosy męskie.

Spośród głosów kobiecych można wyłonić jedynie parę LR - UM, która wykazuje wzajemną bliskość dostarczając sobie nawzajem najlepiej identyfikowanych wzorców.

Uwzględnienie w niniejszym doświadczeniu czterech wersji obrazów binarnych pozwoliło ocenić ich przydatność dla celów rozpoznawania haseł w oparciu o wspólny dla kilku osób zbiór wzorców. Jak wynika z danych zamieszczonych w tab.1 i 2, metoda PM okazała się zdecydowanie lepsza od WO, dostarczając większą liczbę poprawnych identyfikacji we wszystkich pozycjach tabeli odległości. Różnice pomiędzy wynikami uzyskanymi za pomocą tych dwóch metod utrzymują się konsekwentnie dla wartości średnich globalnych w każdej z tabel, jak i średnich dla poszczególnych obiektów (tab.1) oraz głosów (tab.2). Porównanie odpowiadających sobie wartości w kolumnach brzegowych tabel WOW1 - PMW1 oraz WOW3 - PMW3 wypada zawsze na korzyść PM.

Wpływ wartości współczynnika maskowania (W1 lub W3) na wyniki rozpoznawania, nie jest jednoznaczny. Co prawda średnie globalne w tabelach wykazują mniejszą liczbę błędów dla wariantów W1 niż W3, jednak zróżnicowanie to nie jest konsekwentnie utrzymywane w odniesieniu do średnich dla poszczególnych obiektów (tab.1) i głosów (tab.2). Reasumując, rozpoznawanie haseł przebiegało w sposób najbardziej efektywny w oparciu o wariant PMW1 obrazów binarnych.

Analiza statystyczna

Poddano statystycznej analizie wariancyjnej - ANOVA - wyniki testu biorącego pod uwagę poprawność rozpoznawania w 30 najmniejszych odległościach pomiędzy rozpoznawanym obiektem a wzorcami.

Przebieg wykresów (górnych) na ryc. 6 i 10 pozwala wysunąć hipotezę, że rozpoznawanie w odniesieniu tak do poszczególnych osób, jak i poszczególnych wyrazów nie jest zależne od stosowanych wariantów obrazów binarnych. Zakłada się, że mamy do czynienia z trzema czynnikami: G - Głosy, W - Wyrazy oraz B

- warianty obrazów Binarnych, przy liczbie poziomów każdego czynnika: G - 10, W - 10 oraz B - 4. Pierwsza hipoteza sprowadza się zatem do zbadania, czy zachodzi interakcja G B oraz W B. Odnośna dwuczynnikowa ANOVA dała następujące wyniki:

G B:

Źródło zmienności	SS	df	MS	F	p(f)
Głosy (G)	1935.3	9	215.0	8.22	<0.001
Warianty (B)	1473.4	3	491.1	18.78	<0.001
Interakcja G B	289.3	27	10.7	0.41	0.997
Wyjaśniona	3698.1	39	94.8	3.63	<0.001
Resztowa	9414.3	360	26.2		
Łączna	13112.4	399	32.8		

W B:

Źródło zmienności	SS	df	MS	F	p(F)
Wyrazy (W)	4151.5	9	461.3	24.3	<0.001
Warianty (B)	1473.4	3	491.1	25.8	<0.001
Interakcja W B	642.1	27	23.7	1.25	0.185
Wyjaśniona	6267.1	39	160.7	8.45	<0.001
Resztowa	6845.3	360	19.0		
Łączna	13112.4	399	32.8		

W obu przypadkach interakcja jest nieistotna statystycznie, przy czym, jak widać, dla G B zachodzi zupełny brak interakcji, zaś dla W B interakcja jest znacznie powyżej najbardziej tolerancyjnego poziomu istotności $\rho = 0.05$. Nie ma zatem podstaw do odrzucenia hipotezy zerowej o braku interakcji tak G B, jak i G B.

Przeprowadzono jednoczynnikową ANOVA dla czynnika B otrzymując następujące wyniki:

Źródło zmienności	SS	df	MS	F	p(F)
między grupami	1473.4	3	491.1	16.71	<0.001
wewnątrz grup	11639.0	396	29.4		
łączna	13112.4	399			

Średnie (w kolejności rosnącej):

WOW3	17.36
WOW1	18.28
PMW3	21.25
PMW1	21.90

Stosując kryterium HSD (statystycznie wiarygodnego odstępu), na poziomie α =0.05 otrzymuje się tablicę:

	WOW3	WOW1	PMW1	PMW3
WOW3				
WOW1				
PMW1	*	*		
PMW3	*	*		

W powyższej tablicy * oznacza różnicę istotną na poziomie α = 0.05 dla grup na przecięciu rzędu z kolumną. Jak widać, warianty WOW3 i WOW1 nie różnią się między sobą istotnie, podobnie PMW1 i PMW3. Natomiast zachodzi istotna na poziomie α = 0.05 różnica między każdą z pierwszych dwu oraz każdą z pozostałych dwu grup.

Dla czynnika W otrzymujemy w ANOVA jednoczynnikowej następujące wyniki:

Źródło zmienności	SS	df	MS	F	p(F)
między grupami	4151.5	9	461.2	20.08	<0.001
wewnątrz grup	8960.9	390	23.0		
całkowita	13112.4				

Średnie (w kolejności rosnącej) oraz istotne na poziomie α=0.05 różnice według kryterium HSD przedstawia następująca tabela:

		MA	TE	NI	WD	WO	WG	RA	DR	TA	SW
MA	15.65										
TE	15.95										
NI	16.00										
WD	18.92										
WO	19.20	*									
WG	19.53	*	*	*							
RA	20.00	*	*	*							
DR	22.03	*	*	*							
TA	24.53	*	*	*	*	*	*	*			
SW	25.18	*	*	*	*	*	*	*			

Zastosowano następujące skróty: MA(*gnetofon*), TE(*lewizor*), NI(*e*), WD = *W DÓŁ*, WO = *WÓZEK*, WG = *W GÓRĘ*, RA(*dio*), DR(*zwi*), TA(*k*), SW = *ŚWIATŁO*.

Podobnie jak powyżej, gwiazdka na przecięciu określonego wiersza z określoną kolumną oznacza różnicę w średnich istotną na poziomie α=0.05. Według tutaj uwzględnianego kryterium, najlepiej rozpoznaje się wyraz *ŚWIATŁO*, a najgorzej wyraz *MAGNETOFON*.

ANOVA jednoczynnikowa dla czynnika G dała następujące wyniki:

Źródło zmienności	SS	df	MS	F	p(F)
między grupami	1935.3	9	215.0	7.50	<0.001
wewnątrz grup	11177.0	390	28.6		
całkowita	13112.4				

Wartości średnie (w kolejności rosnącej) oraz różnice między nimi istotne na poziomie α=0.05 według kryterium HSD przedstawia poniższa tabela:

		UM	WI	MS	LR	GD	PD	HK	JI	MK
UMk	15.43									
WIm	17.53									
MSm	18.05									
LRk	18.73									
GDk	20.18	*								
PDm	20.25	*								
HKm	20.45	*								
JIm	20.73	*								
MKk	22.70	*	*	*	*					
PSk	22.95	*	*	*	*					

(Litera k po inicjałach oznacza głos kobiecy, zaś m - męski).

Jak widać z powyższych zestawień i wyników analizy wariancji, zróżnicowania pomiędzy głosami były wyraźnie mniejsze niż między wyrazami. Niemniej najlepsze wyniki uzyskano dla głosu PS, a najgorsze dla głosu UM. Nie widać ogólnego zróżnicowania w wynikach między głosami męskimi a kobiecymi.

Przeprowadzono również dwuczynnikową ANOVA dla czynników G i W.

Źródło zmienności	SS	df	MS	F	p(F)
Wyrazy	4151.52	9	461.28	36.31	<0.001
Głosy	1935.32	9	215.04	16.92	<0.001
Interakcja G W	3213.80	18	39.68	3.12	<0.001
Wyjaśniona	9300.65	99	93.95	7.39	<0.001
Resztowa	3811.78	300	12.71		
Całkowita	13112.40	399	32.86		

Wobec bardzo istotnej interakcji pomiędzy czynnikami G i W uzasadnione jest przytoczenie średnich blokowych:

głosy / wyrazy	JI	LR	WI	GD	PD	HR	MK	MS	PS	UM
RA	18.3	21.3	18.8	22.5	21.3	20.3	25.0	15.3	21.8	15.8
DR	18.3	23.8	20.8	22.0	19.8	22.0	25.8	19.5	23.0	25.5
WO	19.3	14.8	19.5	19.3	20.0	20.8	25.3	19.8	22.5	11.0
MA	14.0	11.8	15.3	11.5	20.8	17.3	17.8	14.8	25.3	8.3
TE	17.0	9.3	13.0	17.8	18.5	16.3	20.5	11.8	20.0	15.5
SW	28.0	23.8	18.5	25.5	27.3	26.5	29.0	23.0	28.8	21.5
WG	22.5	18.3	21.0	22.3	21.0	19.5	22.5	19.3	24.0	5.0
WD	22.0	18.8	14.5	19.3	20.5	14.3	22.0	21.0	22.3	14.8
TA	25.3	26.8	25.5	24.8	24.3	24.8	23.0	25.3	23.3	22.5
NI	22.8	19.0	8.5	17.0	11.3	21.0	16.3	11.0	18.8	14.5

Stwierdzoną statystycznie interakcję między głosami i wyrazami unaocznia powyższa tablica, w której można łatwo stwierdzić, na przykład w obrębie określonego głosu, najlepiej i najgorzej rozpoznawany wyraz (w znaczeniu rozpatrywango kryterium), albo w obrębie określonego wyrazu - głos wymawiający go najbardziej lub najmniej dystynktywnie (w sensie przyjętego tutaj kryterium).

Uwagi końcowe.

Przeprowadzone doświadczenie wykazało, iż testowany system charakteryzuje znaczne uniezależnienie od głosu nadawcy. Wykorzystanie wspólnego zbioru wzorców dla grupy osób złożonej z kobiet i mężczyzn nie ograniczyło zdolności rozpoznawczych układu. Wyniki zamieszczone w tablicy identyfikacji wskazują na

znaczną elastyczność w zakresie korzystania z obcych wzorców. Częstokroć największe podobieństwo do obiektu, a więc najmniejszą wartość odległości pomiędzy obrazami binarnymi tegoż obiektu i wybranego wzorca wykazywał wzorzec pochodzący od innego głosu, niż obiekt. Nawet preferowanie męskich wzorców w przypadku rozpoznawania obiektów nadanych przez głosy męskie, nie ograniczało w wyraźny sposób możliwości identyfikacji tych obiektów w oparciu o wzorce kobiece.

Przeprowadzona analiza błędnych identyfikacji umożliwia dotarcie do niektórych źródeł błędów, co pozwoli ukierunkować dalsze prace nad udoskonaleniem systemu.

Bibliografia.

[1]. AINWORTH, W. A., Speech Recognition by Machine, P. Peregrinus Ltd., London 1988.

[2]. CLARK, J.A. & ROEMER, R.B., Voice controlled wheelchair, Archives of Physical Medicine and Rehabilitation, 58, 169-175, 1977.

[3]. COHEN, A. & GRAUPE, D., Speech recognition and control systems for the severely disabled, Journal of Biomedical Engineering 2, 99-107, 1980.

[4]. CREASEY, G.H., Voice controlled systems for the handicapped, Proceedings, World Congress on Medical Physics and Engineering, Hamburg, September 1982, Paper 6.04, 1982.

[5]. DAMPER, R.I., Voice-input aids for the physically disabled, International Journal of Man-Machine Studies, 21, 541-553, 1984.

[6]. DAMPER, R.I., DABBAGH, H.H., Voice-input word-processor aid for the physically disabled. Paper presented at Institution of Electronics and Radio Engineering Conference, Microelectronics in Health Care, Leeds, June 1983.

[7]. DAMPER, R.I. & PURSWANI, V.N., Voice input environment control for the physically disabled, Proceedings, World Congress on Medical Physics and Biomedical Engineering, Hamburg, September 1982, Paper 6.07, 1982.

[8]. GLENN, J.W., HILLER, H.K. & BROWMAN, M.T., Voice terminal may offer opportunities for employment to the disabled, American Journal of Occupational Therapy, 30, 309-313, 1976.

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

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

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

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

[13]. LEA, W.A., The value of speech recognition systems, Trends in Speech Recognition (W.A. Lea, ed.), Prentice Hall, Englewood Cliffs, New Jersey, 9-18, 1980.

[14]. ODOR, J.P. & SHARP, S., How microcomputers can help severely disabled people, Preceedings, Disability and Technology in the '80s Brighton, March 1981, 105-111, 1981.

[15]. WICKE, R., ENGLEHARDT, R.A., AWARD, R., LEIFER, L & VAN DER LOOS, M., Evaluation of a speaker-dependent voice recognition unit as an input device for control of a robotic arm, Paper presented at 6th Annual Conference on Rehabilitation Engineering, San Diego, June 1983.

[16]. YOUDIN, M., SELL, G.H., REICH, T., CLAGNAZ, M., LOUIE, H. & KOLWICZ, R., A voice controlled powered wheelchair and environmental control for the severely disabled, Medical Progress through Technology, 7, 139-143, 1980.