2.23 — akustyka mowy

J. Imiołczyk, H. Kubzdela

AUTOMATYCZNE ROZPOZNAWANIE GŁOSÓW NA PODSTAWIE SPEKTROGRAMÓW BINARNYCH

15/1987

WARSZAWA 1987

ISSN 0208-5658

Praca wpłynęła do Redakcji dnia 8 grudnia 1986 r.

Na prawach rękopisu

Instytut Podstawowych Problemów Techniki PAN

Nakład 160 egz. Ark.wyd. 1,17 Ark.druk. 1,75
Oddano do drukarni w kwietniu 1987 r.
Nr zamówienia 210/87.

Warszawska Drukarnia Naukowa, Warszawa,
ul. Śniadeckich 8

Janusz Imiołczyk
Henryk Kubzdela

Pracownia Fonetyki Akustycznej
IPPT PAN

AUTOMATYCZNE ROZPOZNAWANIE GŁOSÓW NA PODSTAWIE SPEKTROGRAMÓW BINARNYCH[1].

Streszczenie.

Opracowaną dla celów automatycznego rozpoznawania wyrazów metodę tworzenia i porównywania spektrogramów binarnych ([6]) wykorzystano - po wprowadzeniu kilku drobnych modyfikacji - do automatycznej identyfikacji zbioru głosów.

Materiał doświadczalny obejmował 11 wyrazów polskich, z których każdy został wypowiedziany przez 10 mówców (głosy męskie). Przedmiotem porównania były wszystkie spektrogramy reprezentujące ten sam wyraz. Wyznaczone w toku porównywania odległości stanowiły podstawę identyfikacji.

Uzyskane wyniki (ponad 99 % poprawnych identyfikacji) dowodzą, iż przedstawiona w pracy [6] metoda może być wykorzystana zarówno do automatycznego rozpoznawania wyrazów, jak i do automatycznego rozpoznawania głosów.

1. Wstęp.

Spektrogram dowolnej, nie zniekształconej wypowiedzi słownej stanowi odzwierciedlenie zawartych w niej fonetyczno-akustycznych cech lingwistycznych, zdeterminowanych przez skład głoskowy wypowiedzi, oraz cech osobniczych, wynikających z indywidualnego charakteru głosu mówcy. Precyzyjne wyodrębnienie tych dwóch typów cech nie jest możliwe ze względu na występujące między nimi ścisłe interakcje.

Spektrograficzne reprezentacje tej samej wypowiedzi języko-

1) Praca wykonana w ramach planu CPBR 11.9

wej w realizacji kilku różnych głosów cechuje na ogół dość znaczne podobieństwo, świadczące o dominującej roli informacji lingwistycznej w "kształtowaniu" takich reprezentacji. Fakt ten wykorzystuje się w układach automatycznego rozpoznawania mowy, w których spektrogramy wypowiedzi testowych (obiektów) porównywane są ze spektrogramami wypowiedzi wzorcowych poszczególnych słów, stanowiących jednostki słownika układu rozpoznającego. Metoda taka została m.in. zastosowana w opracowanym w efekcie kilkuletnich doświadczeń ([1] , [2] , [3] , [4] , [5]) układzie automatycznego rozpoznawania wyrazów ROWBIR 1 ([6]).

Każde z powtórzeń wypowiedzi o tej samej treści językowej przez ten sam głos w mniejszym lub większym stopniu różni się od pozostałych pod względem cech widmowych. Zróżnicowania takie, określane mianem wewnątrzosobniczych, mają w zasadzie charakter przypadkowy i są z reguły mniejsze od zróżnicowań międzyosobniczych, tj. takich, które występują między realizacjami tej samej wypowiedzi przez różne głosy. Mogą się one np. manifestować poprzez odmienne zlokalizowanie formantów w skali częstotliwości czy niejednakową szerokość pasm formantowych. Zachodzi jednak pytanie, czy są one na tyle powtarzalne i istotne, aby mogły być wykorzystane do efektywnego rozróżniania głosów.

Odpowiedź na postawione wyżej pytanie starano się uzyskać w niniejszej pracy. Bezpośrednim jej celem było stwierdzenie, czy zastosowana we wspomnianym już układzie automatycznego rozpoznawania wyrazów ROWBIR 1 metoda, oparta na porównywaniu spektrogramów wypowiedzi, okaże się również względnie niezawodna w automatycznej identyfikacji zbioru głosów.

2. Metoda tworzenia i porównywania spektrogramów binarnych.

Szczegółowy opis wykorzystanej tu metody przedstawiony został w pracy [6] . Z tego względu podane niżej informacje mają jedynie charakter ogólny i są opatrzone odpowiednimi odnośnikami do tej pracy[1)]. Ograniczenie to nie dotyczy oczywiście tych stosunkowo nielicznych przypadków, w których zaistniała

1) Poniższe dane odnoszą się tylko do tej wersji metody, która została wykorzystana w niniejszej pracy.

konieczność zmodyfikowania metody oryginalnej w celu jej lepszego dostosowania do potrzeb eksperymentu.

2.1. Podstawy techniczne.

Wykorzystany w niniejszej pracy zestaw aparaturowy (por. [6] , str. 44-52) obejmował minikomputer MERA 303 o pamięci operacyjnej 8 K bajtów wraz z modułem pamięci na dyskach elastycznych MDE-300 (4 x 256 K bajtów).
Ponadto, w skład zestawu wchodziło kilka wyspecjalizowanych urządzeń niestandardowych :

- 63-kanałowy analogowy analizator widma, obejmujący zakres od 120 do 8310 Hz,
- kanał funkcji analogowych KF-01,
- monitor graficzny MEMOSKOP (wielkość obrazu : 64 x 256 punktów),
- oscyloskop dwukanałowy z długą poświatą.

2.2. Wygładzanie widma cyfrowego.

Widmo sygnału mowy uzyskiwane na wyjściu analogowego analizatora widmowego zostaje przesłane do pamięci minikomputera. W celu wyeliminowania wpływu harmonicznego charakteru sygnału na obraz widma dokonuje się następnie jego programowego wygładzenia ([6], str. 58-60) poprzez uśrednienie w obrębie 5-elementowego okienka wagowego o wymiarach : 1/4, 3/4, 1, 3/4, 1/4, przesuwającego się wzdłuż osi częstotliwości. Obwiednię wygładzonego widma charakteryzuje bardziej regularny kształt, z wyraźnie zaznaczoną strukturą formantową.

2.3. Widmo binarne. Spektrogram binarny.

Wygładzone widmo cyfrowe zostaje poddane przekształceniu, którego efektem jest tzw. widmo binarne ([6], str. 60, 62-63) charakteryzujące się tym, że jego parametry przyjmują jedynie wartości zero-jedynkowe. Wartość 1 otrzymuje widmo binarne w tych miejscach, które odpowiadają wydatnym wypukłościom (formantom) obwiedni wygładzonego widma (zob. nierówność 5.7, str. 62), w pozostałych zaś - przyjmuje wartość 0.

Pojedyncze 63-parametryczne widmo binarne stanowi odzwierciedlenie częstotliwościowej struktury sygnału mowy w przedzia-

le czasowym wynoszącym ok. 23 ms. Ciąg widm binarnych tworzy spektrogram binarny analizowanej wypowiedzi.

2.4. Normalizacja czasowa.

Porównanie spektrogramów binarnych, stanowiące podstawę wykorzystywanej w układzie ROWBIR 1 metody, możliwe jest dopiero po ustaleniu, które z dwuwidmowych fragmentów jednego z dwóch porównywanych spektrogramów odpowiadają kolejnym fragmentom drugiego. Użycie w tym celu jedynie liniowej normalizacji czasowej nie dałoby zadowalających efektów ze względu na fakt, iż nawet w przypadku spektrogramów binarnych złożonych z identycznej liczby widm poszczególne elementy fonetyczne wypowiedzi mogą mieć inny rozkład czasowy. Istota przyjętego rozwiązania polega na poszukiwaniu zgodnych fragmentów w obszarze obejmującym jednakowej szerokości zakresy czasowe, stanowiące otoczenie krzywej quasi-liniowej (dyskretnej) normalizacji czasowej dwóch porównywanych spektrogramów ([6], wzór 6.1., str. 68).

2.5. Miara podobieństwa fragmentów.

Ustalenie zgodności poszczególnych fragmentów porównywanych spektrogramów binarnych musi się oczywiście dokonywać w oparciu o pewną miarę podobieństwa. Przyjęta w pracy miara wyraża się stosunkiem liczby jedynek występujących w obu porównywanych fragmentach na odpowiadających sobie pozycjach (tj. "pokrywających" się) do sumy wszystkich jedynek w obu tych fragmentach (zob.[6], wzór 6.2., str. 71). Miara ta może przyjmować wartości w granicach od 0 do 1, przy czym "0" oznacza zupełny brak podobieństwa, natomiast "1" - identyczność porównywanych fragmentów (w celu uproszczenia procedur obliczeniowych posługiwano się faktycznie miarą podobieństwa, która jest uzupełnieniem do 1 podobieństw wynikających ze wzoru 6.2).

2.6. Porównywanie spektrogramów binarnych.

Rezultatem porównania dwóch spektrogramów binarnych jest ciąg liczb wyrażających lokalne podobieństwa kolejnych fragmentów pierwszego spektrogramu (tzw. obiektu) i najbardziej podobnych do nich fragmentów drugiego z nich (traktowanego jako "wzorzec").

Należy w tym miejscu wspomnieć o dwu istotnych modyfikacjach, jakie wprowadzono do pierwotnej metody.

Ze względu na fakt, iż - zgodnie z założeniami wstępnymi - przedmiotem porównania w omawianej pracy były spektrogramy binarne pojedynczych wypowiedzi, faza adaptacji (wyznaczanie wzorcowego spektrogramu binarnego, zob. [6], str. 74-83) została praktycznie pominięta. Spektrogramy wszystkich realizacji danej wypowiedzi przez każdy z głosów traktowane były zarówno jako obiekty, jak i wzorce (zbiór obiektów był więc identyczny ze zbiorem wzorców).

Bezpośredni cel eksperymentu stanowiło nie rozpoznanie poszczególnych obiektów poprzez ich porównanie ze wszystkimi wzorcami (por. 6 , str. 90-91), lecz wyznaczanie - w toku tego porównania - odległości każdego z obiektów do poszczególnych wzorców. Odległości te wyrażał globalny wskaźnik podobieństwa, równy średniej arytmetycznej z podobieństw lokalnych.

Ze względu na tryb i przebieg, porównywanie można określić mianem szeregowego (por. 6 , str. 96). We wszystkich przypadkach obejmowało ono wszystkie spektrogramy reprezentujące tę samą wypowiedź językową (konkretnie : ten sam wyraz).

W charakteryzującej się stosunkowo niewielką pojemnością pamięci operacyjnej minikomputera MERA 303 mieściło się jednorazowo, oprócz obiektu, zaledwie 6 wzorców, z którymi obiekt ten miał być porównany. Pociągnęło to za sobą konieczność przechowywania wzorców w pamięci zewnętrznej i ich przesyłania do pamięci operacyjnej w grupach po 6, co oczywiście wydłużyło czas oczekiwania na wynik.

3. Identyfikacja głosów w oparciu o spektrogramy binarne[1)].

Doświadczenia z zakresu identyfikacji głosów przeprowadzono w dwóch etapach, z których pierwszy - o charakterze pilotażowym - miał stworzyć podstawy do wstępnej oceny efektywności przyjętej metody, drugi zaś obejmował właściwy eksperyment, w którym posłużono się obszerniejszym materiałem językowym i

1) Określenia "rozpoznawanie" i "identyfikacja" będą w dalszej części pracy używane dla uproszczenia, z zastrzeżeniem poczynionym w punkcie 2.6.

liczniejszą grupą głosów.

W eksperymencie wstępnym porównano ze sobą 24 spektrogramy binarne reprezentujące 4-krotne wymówienie przez 6 głosów (3 żeńskie i 3 męskie) wyrazu "sznurowadło" oraz - analogicznie - 24 spektrogramy wyrazu "Milanówek". We wszystkich przypadkach globalny wskaźnik podobieństwa między obiektem a 24-elementową grupą wzorców był najmniejszy (tzn. podobieństwo było największe) dla tych wzorców, które pochodziły od tego samego głosu co obiekt[1]. Wynik ten wskazywał na możliwość wykorzystania przyjętej metody do względnie niezawodnej identyfikacji zbioru głosów i stał się punktem wyjścia do przeprowadzenia szerzej zakrojonego eksperymentu.

3.1. Materiał językowy.

Przy doborze materiału językowego dla celów planowanego doświadczenia położono przede wszystkim nacisk na to, aby objął on głoski reprezentujące wszystkie fonemy języka polskiego. Ze względu na niewielką pojemność pamięci operacyjnej minikomputera MERA 303 zadecydowano ponadto, że wypowiedzi testowe powinny być stosunkowo krótkie, maksymalnie - 3-sylabowe. W skład zestawu doświadczalnego weszło ostatecznie 11 następujących wyrazów :

sadze, dżuma, hokej, dźwięki, błogość, flesz, źrenica, Giewont, zapis, rzeczy, bieda.

3.2. Nagrania.

W nagraniach materiału językowego, przeprowadzonych w pomieszczeniu bezechowym, posłużono się 10 głosami męskimi. Zadanie mówców polegało na czterokrotnym wymówieniu - w ok. 2-sekundowych odstępach - każdego z 11 wyrazów, zapisanych na dostarczonej liście. Nagrania poprzedzano każdorazowo instrukcją oraz krótkim treningiem, mającym na celu zminimalizowanie zróżnicowań w zakresie tempa mowy i przebiegu intonacji. Zwrócono także uwagę na to, aby poziom nagrań był jednakowy. W przypadku wystąpienia zauważalnych odstępstw od przyjętych "norm", nagranie powtarzano.

---

1) Najmniejsza (zerowa) odległość występowała przy porównaniu danego spektrogramu z nim samym.

Każda z osób nagrywała cały materiał w trakcie jednej sesji.

3.3. Tworzenie spektrogramów binarnych.

Kolejny etap pracy obejmował tworzenie spektrogramów binarnych wypowiedzi i ich zapisanie na dysku elastycznym. Zgodnie z założeniami wstępnymi wzięto przy tym pod uwagę jedynie trzy z czterech wymówień każdego wyrazu przez każdy z głosów (eliminowano z reguły czwarte, ostatnie wymówienie)[1]. Oznaczało to zredukowanie ogólnej liczby wypowiedzi z 440 do 330 (11 wyrazów x 3 wymówienia x 10 głosów).

W efekcie analizy widmowej odtworzonej z taśmy wypowiedzi oraz serii przekształceń pochodzących z tej analizy widm (por. pkt. 2.2 i 2.3) uzyskiwano w pamięci minikomputera spektrogram binarny. Specjalny program umożliwiał wyświetlenie spektrogramu na ekranie monitora i jego weryfikację[2]. Z możliwości tej korzystano jednak tylko w odniesieniu do niektórych spektrogramów, głównie w przypadkach, gdy poprawność realizacji odtwarzanej wypowiedzi budziła zastrzeżenia (np. ze względu na substytucję głosek, brak płynności wypowiedzi itp.)[3]. Uznając, że stosowanie innych kryteriów niż tu wymienione stanowiłoby manipulację materiałem, spektrogramy pozostałych wypowiedzi przesyłano do pamięci zewnętrznej z pominięciem ich wizualizacji.

Kolejność w jakiej zapisywano spektrogramy poszczególnych wypowiedzi na dysku elastycznym była zgodna z kolejnością występowania tych wypowiedzi na taśmie magnetofonowej. Obszar pamięci przeznaczony na zapisanie jednego spektrogramu obejmował 640 bajtów, co umożliwiało zgromadzenie całego materiału na jednym dysku (330 x 640 bajtów ≈ 212 Kbajtów).

3.4. Tworzenie testowych zbiorów spektrogramów binarnych.

W celu uporządkowania zgromadzonego na dysku I materiału pierwotnego w sposób właściwy z punktu widzenia potrzeb ekspery-

---

1) Por. niżej.

2) Istniała również możliwość wydruku bieżącego spektrogramu.

3) Zamiast niepoprawnie zrealizowanej wypowiedzi uwzględniano czwarte, "rezerwowe" wymówienie.

mentu dokonano jego przepisania na dwa nowe dyski (dysk II i dysk III). Na pierwszym z nich spektrogramy występowały w 30-elementowych grupach, z których każda obejmowała wszystkie (po 3) wymówienia przez wszystkie głosy jednego (tego samego) wyrazu. Sposób uporządkowania zilustrowano poniżej :

Nr kol.	Wyraz : SADZE
1	Głos 1, wymówienie 1
2	- - - - 2
3	- - - - 3
4	Głos 2, wymówienie 1
5	- - - - 2
6	- - - - 3
7	Głos 3, wymówienie 1
...	... ...
30	Głos 10, wymówienie 3
	Wyraz : DŹUMA
31	Głos 1, wymówienie 1
...	... ...

Podobnie jak na dysku I, obszar przeznaczony na zapisanie jednego spektrogramu binarnego obejmował 640 bajtów. W trakcie przepisywania spektrogramów istniała możliwość ich wizualizacji na ekranie monitora, co zmniejszało ryzyko popełnienia błędów.

Kolejność występowania spektrogramów na dysku III była identyczna jak na dysku II. Zasadnicza różnica polegała na zapisaniu ich bezpośrednio po sobie, tzn. w ściśle przylegających do siebie obszarach o niejednakowej objętości, zależnej od rozciągłości czasowej wypowiedzi. Spektrogramy zapisano wraz z ich etykietami w postaci kodów wyrazu, głosu i numeru wymówienia (przykładowo : etykieta "1DZW1" oznacza wyraz o numerze 1 na liście [SADZE], głos DZW, wymówienie 1, zaś etykieta "6JB3" - wyraz o numerze 6 [FLESZ], głos JB, wymówienie 3).

Spektrogramy zapisane na dysku II i na dysku III nazwano odpowiednio "zbiorem obiektów" i "zbiorem wzorców"[1].

3.5. Wyznaczanie odległości między obiektami i wzorcami.

Zasadnicza część doświadczenia obejmowała porównanie ze sobą - metodą "każdy z każdym" x 2 (tj. np. A z B i B z A) - wszystkich spektrogramów binarnych reprezentujących ten sam wyraz. Ponieważ łączna liczba różnych realizacji tego samego wyrazu wynosiła 30 (10 głosów x 3 wymówienia), dokonano w obrębie każdego wyrazu - ogółem 900 porównań, uzyskując w efekcie 900 odrębnych wartości globalnego wskaźnika odległości (dalej : GWO, por. pkt. 2.5 i 2.6)[2].

W toku porównywania do pamięci operacyjnej minikomputera przepisywany był najpierw z dysku II jeden z 30 spektrogramów-obiektów, a następnie - z dysku III - grupa 6 spektrogramów - wzorców, z którymi obiekt ten miał być porównany (por. pkt. 2.6). W efekcie porównania 6 wartości GWO wyprowadzanych było na drukarkę, po czym następowało przesłanie do pamięci operacyjnej kolejnej grupy 6 wzorców. Po wydrukowaniu ostatniej, trzydziestej wartości GWO do pamięci pobierany był kolejny obiekt.

Wydruk odległości miał postać matrycy przedstawionej w Tablicy 1. Występująca na przecięciu kolumny i wiersza wartość GWO (wyrażona ósemkowo) określa stopień podobieństwa między spektrogramami binarnymi dwóch wypowiedzi, zidentyfikowanych w nagłówku tej kolumny i tego wiersza. Podobieństwo jest tym większe, im niższa jest wartość GWO (zerowe wartości GWO na przekątnej głównej matrycy stanowią wynik porównania danego spektrogramu z nim samym).

W celu zwiększenia czytelności matrycy podzielono ją na kwadraty, z których każdy zawiera 9 wartości GWO, określających

---

1) Należy pamiętać, że określenia te są jedynie umowne, gdyż w istocie oba zbiory były identyczne (por. pkt. 2.6).

2) Wartości globalnego wskaźnika odległości uzyskiwane w wyniku dwukrotnego porównania ze sobą dwóch tych samych spektrogramów (np. 1PD1 z 1JB3 i 1JB3 z 1PD1) różniły się nieco od siebie ze względu na fakt, iż w poprzedzającym porównanie procesie normalizacji czasowej wzorzec (drugi element pary) był we wszystkich przypadkach "dopasowany" do obiektu, uznano za konieczne uwzględnienie obu tych wartości.

S A D Z E

	DZW			JI			TM			PDZ			WF			JB			PD			RC			HK			TW		
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
DZW1	0	24	25	33	35	32	43	37	40	34	36	36	45	42	42	44	37	42	44	41	37	36	35	36	44	43	43	36	42	37
DZW2	23	0	21	33	33	32	41	37	36	34	36	34	46	42	42	43	41	41	44	41	37	37	35	37	44	43	43	42	41	40
DZW3	26	21	0	31	33	32	44	37	36	36	37	34	47	43	43	43	40	41	43	41	35	40	36	40	44	45	43	42	41	36
JI1	34	31	30	0	21	20	36	32	33	26	27	32	42	36	37	42	36	42	40	42	40	36	36	37	45	44	45	37	41	36
JI2	35	32	32	20	0	16	34	34	35	27	31	33	41	35	40	43	40	42	38	40	37	40	37	40	46	43	44	35	37	35
JI3	33	31	31	17	16	0	34	31	33	25	27	31	41	36	37	42	36	41	35	40	35	37	35	37	46	44	45	34	36	33
TM1	41	37	42	34	35	33	0	23	25	30	31	41	41	34	34	45	42	41	35	41	43	43	40	43	43	45	43	34	44	43
TM2	34	35	35	30	32	31	23	0	20	27	30	33	41	35	36	45	40	42	34	36	36	35	34	35	43	42	43	33	36	35
TM3	37	33	34	31	33	32	25	20	0	30	27	32	42	36	37	44	41	41	35	37	34	35	32	35	42	40	41	33	37	35
PDZ1	35	33	35	25	26	24	30	30	27	0	16	24	34	27	33	43	42	43	36	42	37	36	35	35	43	41	42	34	36	34
PDZ2	42	37	40	33	33	32	33	33	30	22	0	25	41	35	36	45	43	44	42	46	42	42	40	41	45	44	45	36	41	40
PDZ3	37	36	34	32	33	30	41	33	32	24	23	0	41	40	46	42	41	45	45	47	41	42	41	40	44	42	44	42	42	36
WF1	43	44	46	41	42	41	42	42	43	34	36	41	0	24	26	46	43	44	40	44	47	42	42	42	51	47	47	41	43	41
WF2	41	41	43	36	35	36	35	34	34	31	32	40	24	0	23	44	42	42	33	37	40	40	40	42	46	42	43	37	41	40
WF3	41	43	42	35	37	37	35	36	36	32	33	45	25	24	0	44	42	41	34	37	43	40	40	42	47	45	43	36	43	41
JB1	44	43	43	42	43	42	44	44	44	44	43	41	46	43	43	0	27	26	41	42	41	40	40	37	45	41	44	44	45	43
JB2	40	40	40	36	40	37	42	37	41	42	40	40	44	41	42	26	0	26	40	41	37	35	34	34	43	41	44	43	41	42
JB3	43	40	42	42	43	43	41	42	41	44	41	45	46	43	41	25	25	0	41	40	37	36	40	40	44	43	46	43	43	43
PD1	43	43	40	36	32	35	33	32	33	33	36	42	37	32	33	40	40	36	0	23	24	35	34	37	50	46	47	32	36	34
PD2	40	41	36	40	37	37	41	35	36	40	42	44	43	36	35	40	40	36	24	0	24	36	31	34	45	43	45	33	35	34
PD3	36	36	34	37	37	36	42	36	35	37	40	40	45	40	43	40	36	36	27	26	0	35	36	33	46	40	45	36	40	35
RC1	36	40	40	40	41	40	43	35	35	37	37	41	43	40	41	42	35	40	37	37	36	0	22	25	37	41	37	40	37	35
RC2	35	35	36	36	37	36	41	34	34	35	35	37	41	41	37	40	35	37	35	34	37	21	0	22	41	40	41	34	33	32
RC3	36	40	41	36	40	37	42	35	35	36	37	37	44	43	42	41	37	42	40	37	35	25	23	0	42	40	41	41	36	34
HK1	45	44	43	45	46	47	44	45	43	44	44	45	51	46	46	45	42	44	52	46	50	37	40	40	0	32	26	43	44	43
HK2	43	43	44	44	45	46	47	45	43	43	43	43	46	43	46	42	42	44	44	44	44	41	42	40	32	0	27	43	43	42
HK3	44	43	44	46	45	46	44	46	43	43	43	44	46	43	43	45	45	45	50	45	47	41	43	41	26	30	0	44	40	40
TW1	37	41	41	37	34	34	35	34	34	34	34	42	41	37	37	43	42	42	33	35	36	36	34	37	43	42	43	0	22	22
TW2	41	40	37	41	40	36	44	36	35	40	36	37	44	40	42	43	40	43	37	37	37	35	33	35	43	42	40	22	0	20
TW3	37	37	36	35	34	32	43	34	36	35	35	35	41	37	41	42	41	42	35	34	36	35	32	32	44	40	37	22	20	0

Tablica 1: Wartosci GWO (wyrazone osemkowo) dla spektrogramow 30 wypowiedzi wyrazu SADZE

stopień podobieństwa spektrogramów trzech realizacji danego wyrazu przez dwa głosy.

3.6. Wyniki i ich omówienie.

Wartości występujące w poszczególnych wierszach Tablicy 1 określają podobieństwo między obiektem, którego etykieta podana jest na początku wiersza, i 30 wzorcami. Do każdego z obiektów najbardziej podobny jest jeden z dwóch wzorców pochodzących od tego samego głosu co obiekt[1]. Poza jednym przypadkiem (obiekt PDZ1 - wzorzec JI3), także drugi z wzorców należących do tego samego głosu co obiekt jest bardziej podobny do tego obiektu niż wszystkie pozostałe wzorce (odległość między PDZ1 i PDZ3 wynosi 24g i jest taka sama jak odległość między PDZ1 i JI3). Wynika stąd, iż zróżnicowanie wewnątrzosobnicze, zaznaczające się przy porównaniu ze sobą spektrogramów binarnych kilku wymówień tego samego wyrazu przez ten sam głos są konsekwentnie mniejsze od zróżnicowań międzyosobniczych. Wniosek ten potwierdzają rezultaty uzyskane dla pozostałych 10 wyrazów. W przypadku 8 z nich wszystkie obiekty charakteryzuje największe podobieństwo do jednego z dwóch wzorców reprezentujących ten sam głos. Jedynie w wyrazach "dżuma" i "Giewont" dwa obiekty (po jednym w każdym z tych wyrazów) okazały się bardziej podobne do wzorców należących do innych głosów. Dotyczyło to obiektów DŻUMA TM2 (najmniejsza odległość - do DŻUMA JI1) oraz GIEWONT TM3 (najmniejsza odległość - do GIEWONT JB2). W obu tych przypadkach powodem znacznej odległości obiektu od dwóch pozostałych wzorców głosu TM był brak dźwięcznego zwarcia na początku wypowiedzi[2].

Z ogólnej liczby 330 obiektów, 328 cechowało większe podobieństwo do wzorca reprezentującego ten sam głos niż do

---

1) Wartość zerowa nie jest oczywiście brana pod uwagę.

2) Materiał ilustracyjny w postaci wydruków spektrogramów binarnych zamieszczono w Dodatku na końcu pracy.

któregoś z pozostałych wzorców. Oznacza to, iż gdyby w automatycznej identyfikacji głosów przyjąć jako kryterium wartość GWO, poprawność rozpoznawania wynosiłaby ponad 99 % [1]. Jak należy oczekiwać, poprawność ta byłaby jeszcze wyższa w przypadku uwzględnienia większej liczby repetycji każdego z wyrazów przez ten sam głos.

Każda z 900 wartości podanych w Tablicy 1 dostarcza jedynie informacji o odległości między pewnym (pojedynczym) obiektem i pewnym (pojedynczym) wzorcem. W celu uzyskania informacji bardziej ogólnych zsumowano - po uprzednim sprowadzeniu do postaci dziesiętnej - wszystkie (18) wartości GWO odnoszące się do każdej pary głosów (a więc zawarte w kwadratach położonych symetrycznie względem przekątnej wyznaczonej przez wartości zerowe), zastępując je ich średnią arytmetyczną[2]. Otrzymano w rezultacie matrycę trójkątną, w której na przecięciu każdego wiersza i każdej kolumny występowała pojedyncza wartość, wyrażająca (średnią) odległość między dwoma głosami, zidentyfikowanymi w nagłówku tego wiersza i tej kolumny. Matryca tego typu dla wyrazu "dżuma" przedstawiona jest w Tablicy 2. Podobnie jak w pozostałych 10 matrycach, występujące w niej wartości są najmniejsze w obrębie tego samego głosu (dotyczy to także głosu TM, mimo iż drugie wymówienie wyrazu "dżuma" przez ten głos nie było dla niego "typowe").

Średnie wartości odległości umożliwiają uszeregowanie głosów według ich podobieństwa oraz przybliżone określenie stopnia ich "zindywidualizowania".

Pełniejszy obraz zróżnicowań między głosami daje matryca zamieszczona w Tablicy 3. Każda z występujących w niej wartości stanowi średnią arytmetyczną z sumy odległości między dwoma danymi głosami w całym materiale (tj. we wszystkich 11 matrycach

---

1) Z opracowania odpowiedniego (stosunkowo prostego) programu zrezygnowano, koncentrując się na istotnej z punktu widzenia potrzeb automatycznego rozpoznawania mowy kwestii odległości między głosami.

2) Średnią arytmetyczną zastąpiono również 9 wartości zawartych w obrębie każdego z kwadratów leżących na przekątnej "zerowej"; przy jej obliczaniu nie brano pod uwagę trzech wartości zerowych.

DZUMA

	DZW	JI	TM	PDZ	WF	JB	PD	RC	HK	TW
DZW	14.2	35.4	35.5	38.4	40.7	32.7	41.5	37.9	36.6	38.4
JI		17.3	32.2	28.5	26.6	33.7	29.7	27.9	32.8	25.7
TM			24.7	34.8	31.0	32.4	35.4	33.0	33.7	31.2
PDZ				13.8	23.7	39.7	29.6	28.3	28.6	24.4
WF					11.5	32.0	26.8	27.6	29.3	22.6
JB						17.0	36.3	38.9	37.9	34.7
PD							19.3	29.6	30.4	27.6
RC								14.3	27.1	25.6
HK									16.0	27.4
TW										15.3

Tablica 2: Srednie odleglosci miedzy glosami dla wyrazu DZUMA

	DZW	JI	TM	PDZ	WF	JB	PD	RC	HK	TW
DZW		30.9	33.6	34.4	35.0	32.1	34.8	35.0	37.5	35.0
JI			31.5	29.4	32.1	32.2	30.6	30.6	34.2	30.5
TM				31.2	32.6	33.1	31.9	32.7	36.1	30.4
PDZ					30.0	35.1	31.0	31.1	31.5	29.5
WF						33.1	30.7	32.4	35.7	30.1
JB							31.9	33.7	36.4	32.8
PD								29.9	34.0	28.4
RC									31.6	29.6
HK										32.9
TW										

Tablica 3: Srednie odleglosci miedzy glosami dla wszystkich wyrazow

"wyrazowych"). Analiza tej matrycy oraz matryc dla poszczególnych wyrazów nasuwa kilka wniosków istotnych zarówno z punktu widzenia identyfikacji głosów, jak i automatycznego rozpoznawania mowy. Wynika z niej przede wszystkim, iż w rozważanej grupie niektóre głosy charakteryzują się dość znacznym podobieństwem do niemal wszystkich pozostałych, inne natomiast wykazują wyraźne zindywidualizowanie. Szczególnie dobitnie i konsekwentnie zindywidualizowanie to zaznacza się w przypadku głosów HK i DZW, dla których przedstawione w Tablicy 4 średnie z odległości od wszystkich pozostałych głosów są największe.

Głos	Średnia	Odchyl. stand.	Głos	Średnia	Odchyl. stand.
TW	31.0	3.9	WF	32.3	3.2
JI	31.3	1.7	TM	32.6	2.4
PD	31.4	3.4	JB	33.4	2.0
PDZ	31.5	3.6	DZW	34.3	3.2
RC	31.8	2.9	HK	34.4	4.0

Tablica 4: Zestawienie średnich z odległości każdego głosu od wszystkich pozostałych oraz odchyleń standardowych

Do grupy głosów o najniższych średnich zaliczają się TW, JI, PD oraz PDZ. Należy jednak zwrócić uwagę, że trzy z nich (TW, PD, PDZ) - na co wskazują dość znaczne wartości odchylenia od odpowiadających im średnich - są do niektórych głosów zdecydowanie niepodobne. Przykładowo, niska wartość średnia dla głosu TW jest wynikiem jego znacznego podobieństwa do głosów PD, PDZ, RC, WF, TM i JI, natomiast stosunkowo wysoka wartość odchylenia standardowego - wynikiem braku podobieństwa do trzech pozostałych (DZW, HK, JB).

Z porównania wartości średnich i odchyleń w Tablicy 4 wynika, że głosem, który charakteryzuje się największym podobieństwem do wszystkich pozostałych jest JI. Odległości dzielące go od innych głosów są niewielkie (niska średnia) i dosyć do siebie zbliżone (niska wartość odchylenia).

3.7. Podsumowanie.

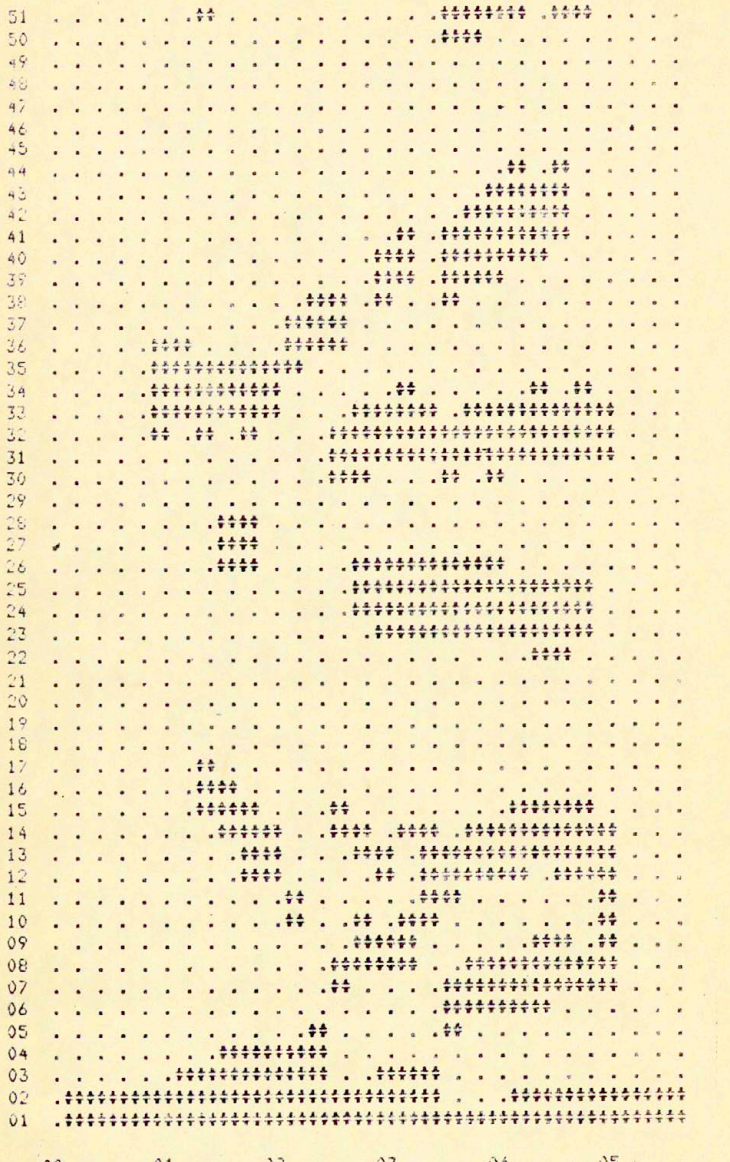
Uzyskane w pracy wyniki świadczą o skuteczności przyjętej metody identyfikacji głosów. Na ogólną liczbę 330 obiektów zaledwie 2 (ok. 0,6 %) okazały się bardziej podobne do wzorców pochodzących od innych (niż one) głosów, przy czym oba przypadki dotyczyły tego samego głosu i miały identyczne podłoże. Wynika stąd, iż zróżnicowania wewnątrzosobnicze są konsekwentnie mniejsze od zróżnicowań międzyosobniczych głosów. Zaznacza się to szczególnie wyraźnie przy porównaniu średnich odległości między spektrogramami binarnymi reprezentującymi ten sam głos z analogicznymi średnimi dla pozostałych głosów (Tablica 2).

Jak można przypuszczać, zwiększenie niezawodności zaproponowanej metody identyfikacji głosów byłoby możliwe przy uwzględnieniu większej liczby repetycji każdego z wyrazów przez ten sam głos, bądź przy posłużeniu się wzorcami "wypadkowymi", utworzonymi w oparciu o kilkakrotne wymówienie tego samego wyrazu przez dany głos.

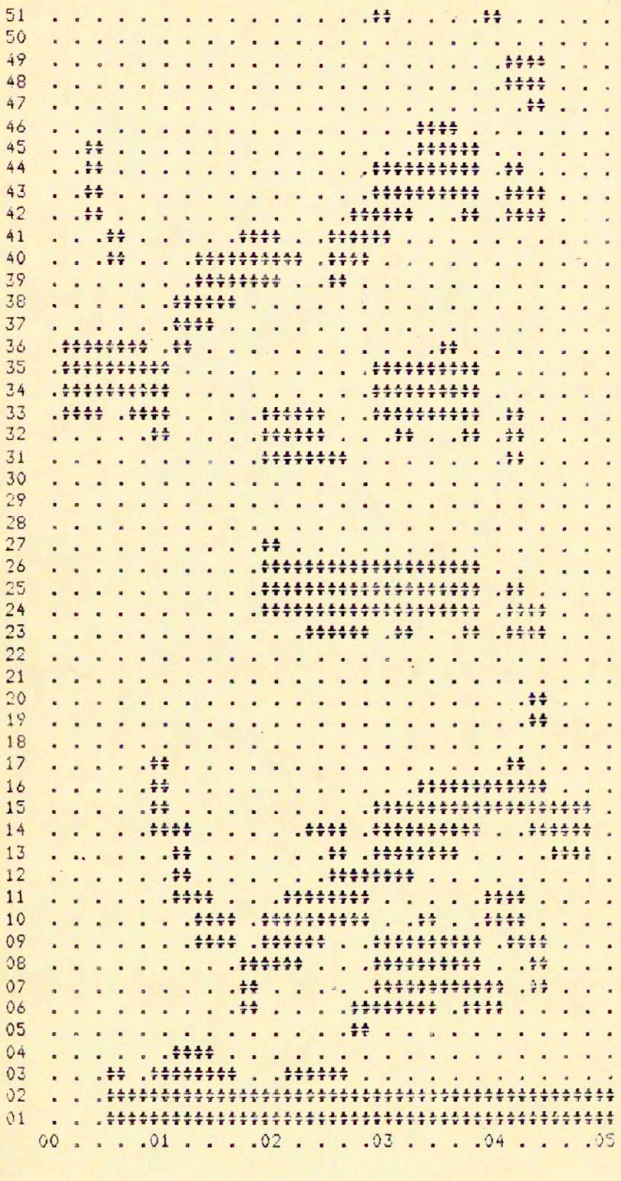
Zgodnie z oczekiwaniami, wykorzystane w doświadczeniu głosy różniły się od siebie pod względem stopnia zindywidualizowania/podobieństwa do innych, przy czym dwa z nich (HK i DZW) okazały się niepodobne do niemal wszystkich pozostałych. "Środkowe" miejsce w grupie przypadło pod tym względem głosowi JI, którego średnie podobieństwo do wszystkich pozostałych było największe. Można przypuszczać, iż uwzględnienie w układzie automatycznego rozpoznawania mowy wzorców pochodzących od tego głosu umożliwiłoby względnie efektywne sterowanie układem większej liczbie operatorów.

D O D A T E K

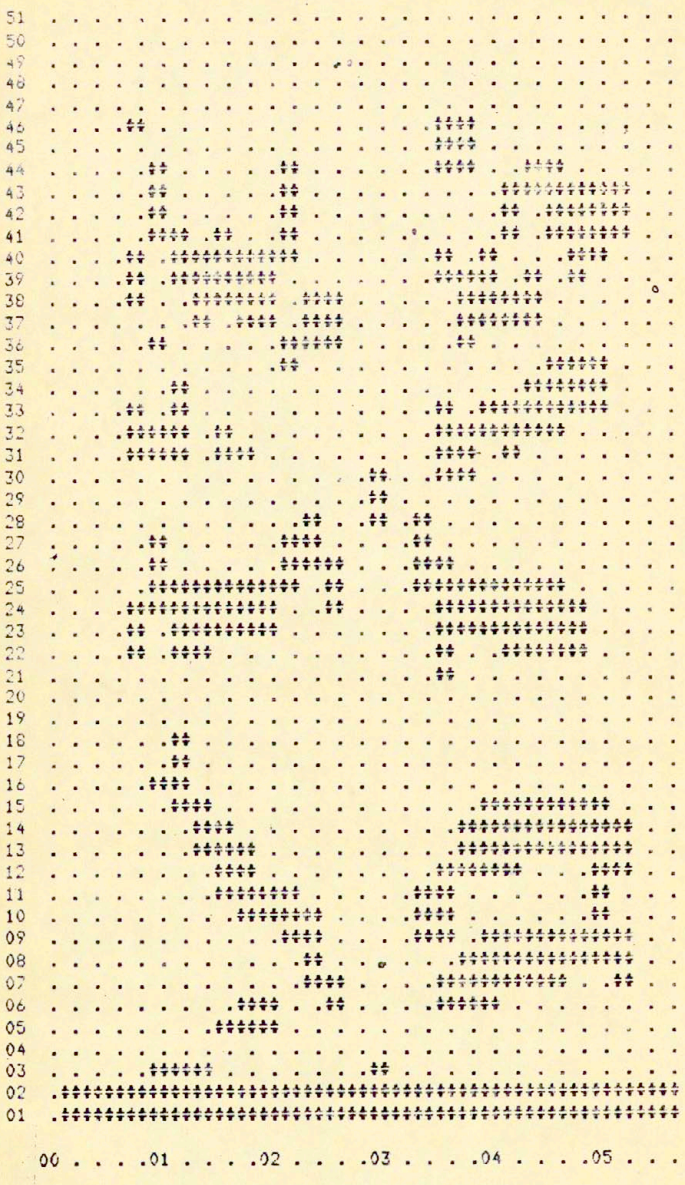
SPEKTROGRAMY BINARNE WYBRANYCH WYPOWIEDZI



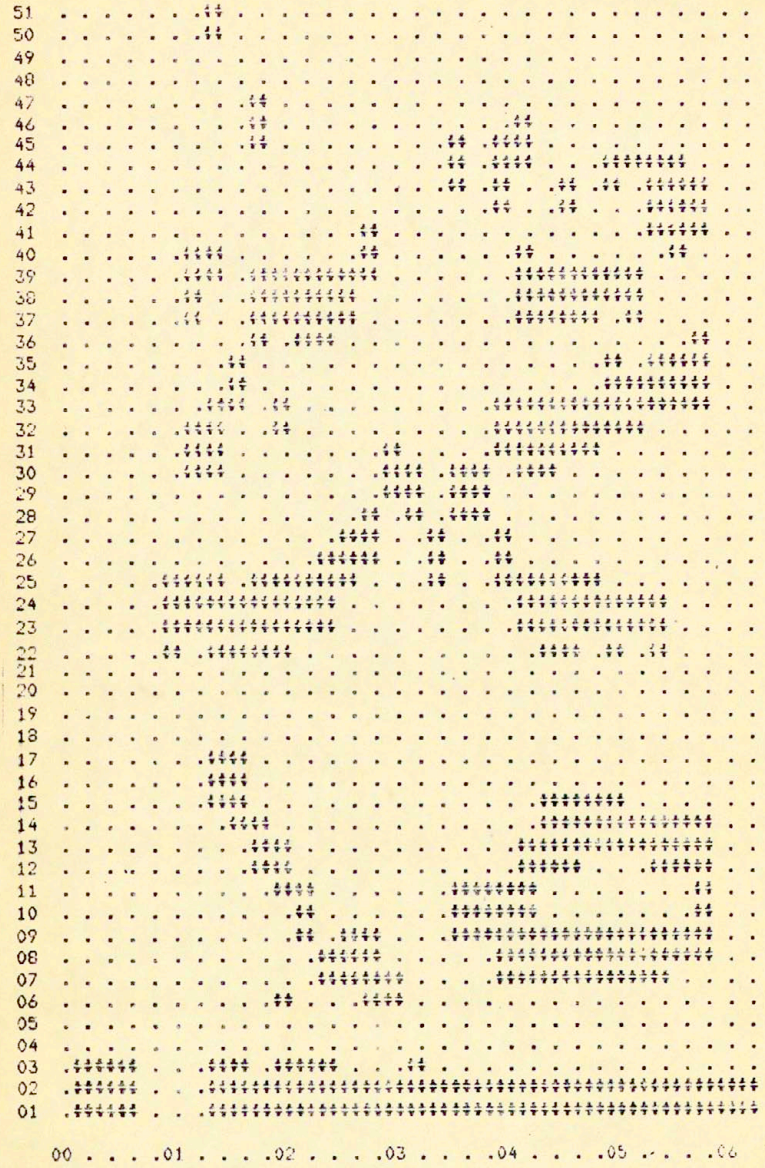
Spektrogram wypowiedzi DZUMA TM1



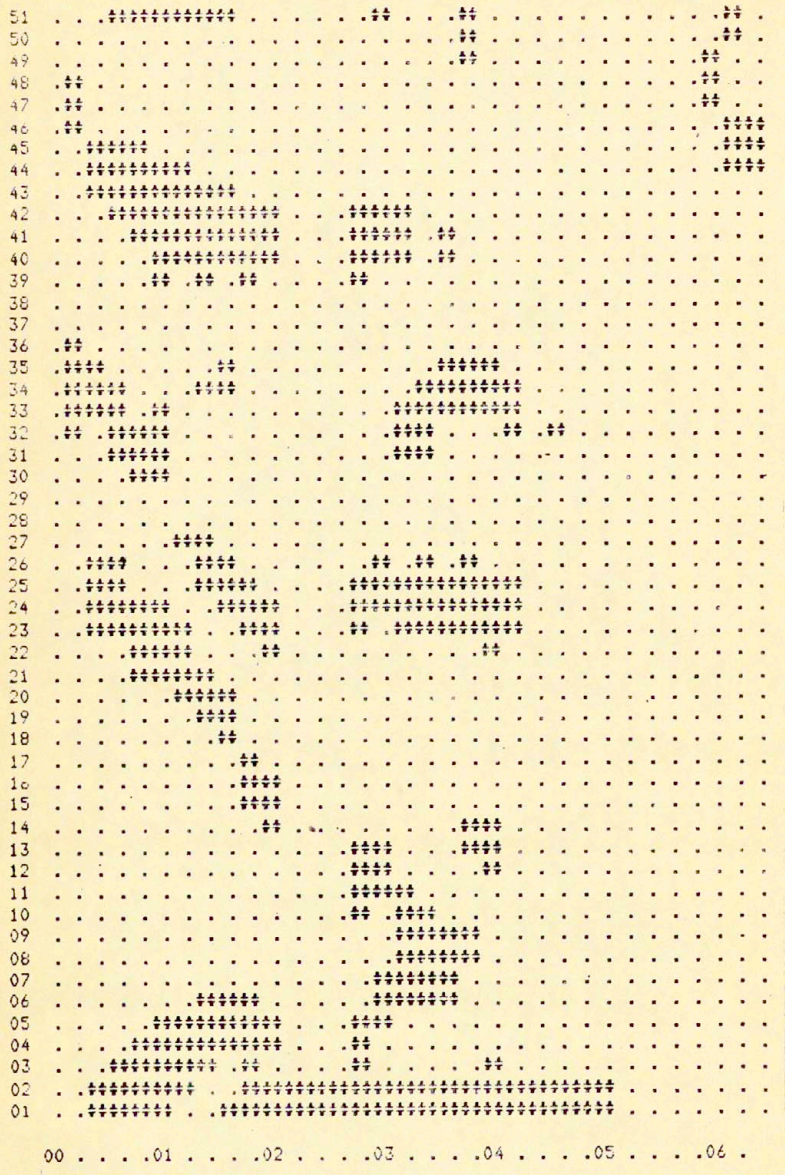
Spektrogram wypowiedzi DŹUMA TM2



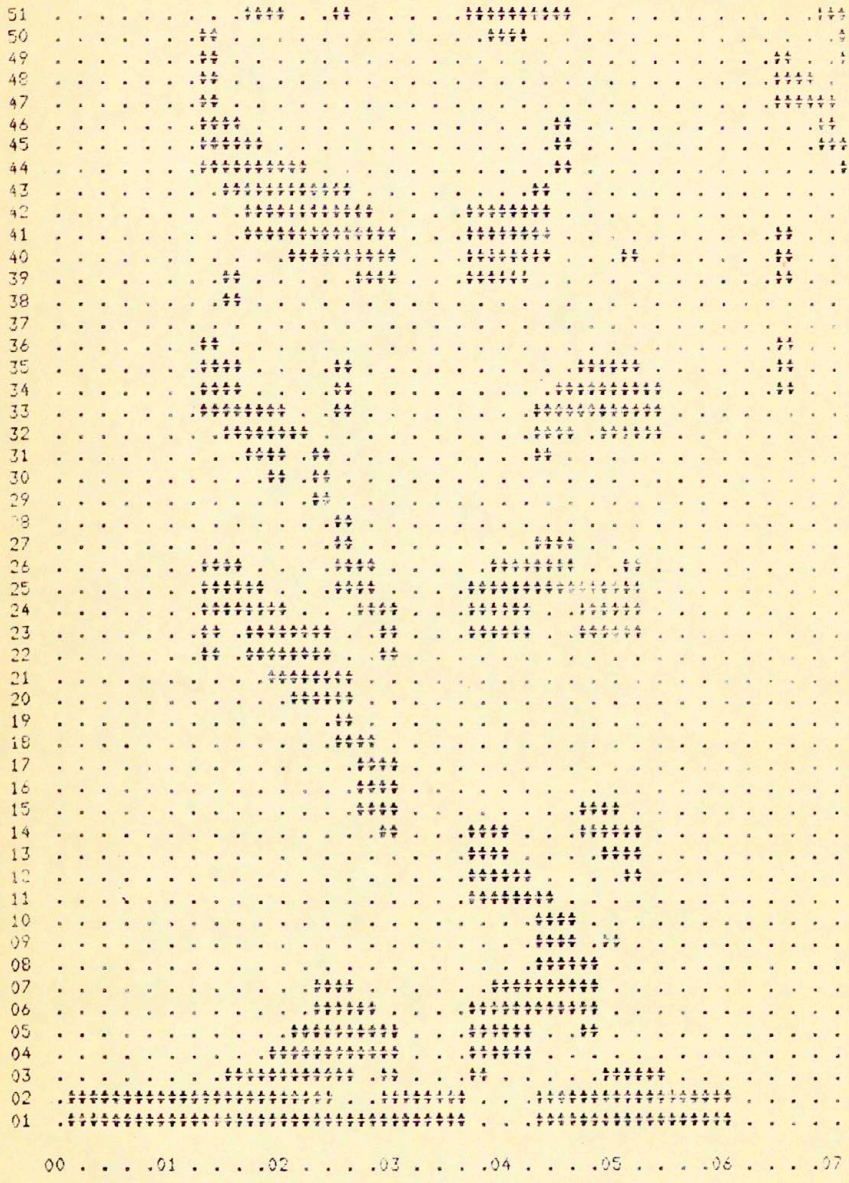
Spektrogram wypowiedzi DŻUMA JI1



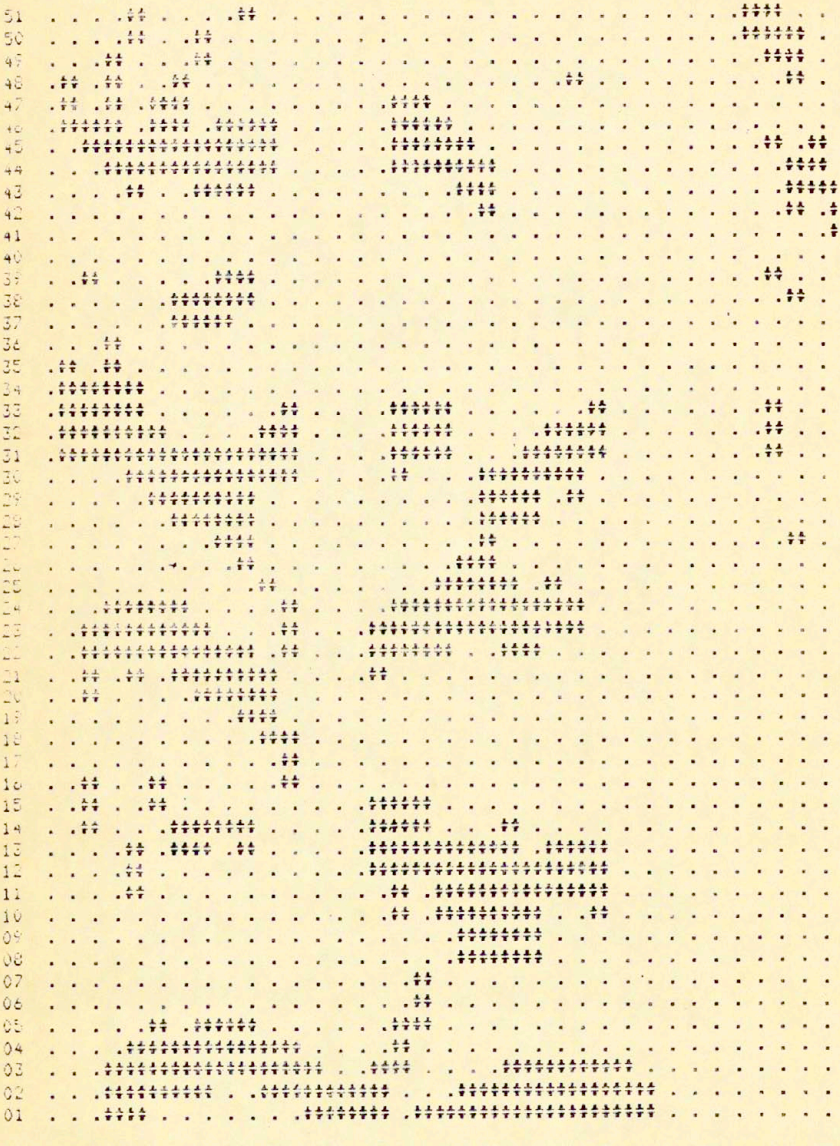
Spektrogram wypowiedzi DŻUMA JI2



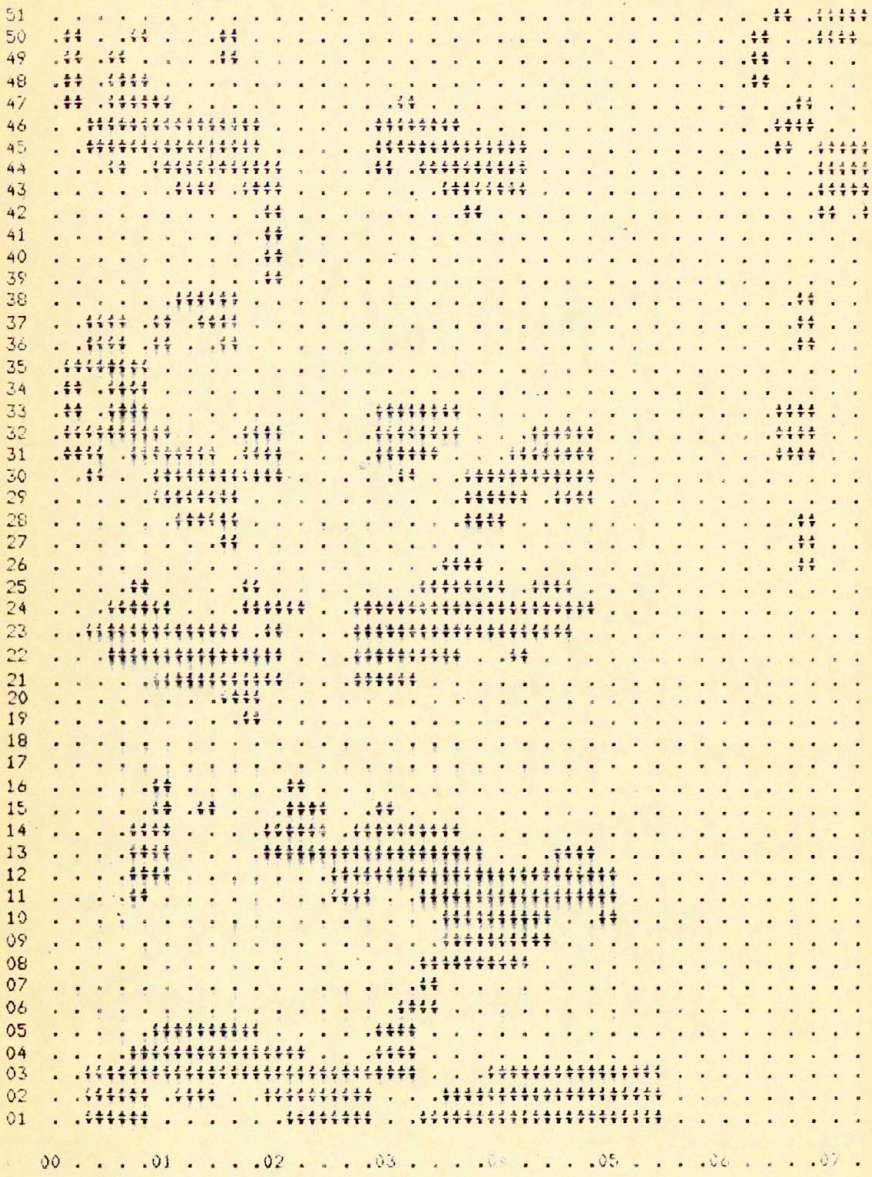
Spektrogram wypowiedzi GIEWONT TM2



Spektrogram wypowiedzi GIEWONT TM3



Spektrogram wypowiedzi GIEWONT JB2



Spektrogram wypowiedzi GIEWONT JB1

BIBLIOGRAFIA

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

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

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

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

[5] KUBZDELA, H., Próby automatycznego rozpoznawania wyrazów wymawianych przez różne głosy w oparciu o grupowe zbiory wzorcowych spektrogramów binarnych, Prace IPPT 47/1983, Warszawa.

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