Henryk Kubzdela

UDOSKONALENIE REPREZENTACJI SYGNAŁU MOWY W FORMIE OBRAZÓW BINARNYCH

24/1987

WARSZAWA 1987

ISSN 0208-5658

Praca wpłynęła do Redakcji dnia 20 listopada 1986 r.

Na prawach rękopisu

Instytut Podstawowych Problemów Techniki PAN

Nakład 140 egz. Ark.wyd. 1,17 Ark.druk. 1,75
Oddano do drukarni w lipcu 1987 r.
Nr zamówienia 385/87.

Warszawska Drukarnia Naukowa, Warszawa,
ul. Śniadeckich 8

Henryk Kubzdela
Pracownia Fonetyki Akustycznej
IPPT PAN

UDOSKONALENIE REPREZENTACJI SYGNAŁU MOWY W FORMIE OBRAZÓW BINARNYCH.

1. Wstęp

Bardzo ważnym etapem w procesie automatycznego rozpoznawania mowy jest przetworzenie sygnału akustycznego w obraz. Liczy się w tym przekształceniu kilka istotnych czynników związanych zarówno z samym jego trybem, jak i efektem. Mimo, że dotychczas uczyniono ogromnie wiele w zakresie opisu parametrycznego sygnału mowy dla potrzeb automatycznego rozpoznawania mowy, to jednak problem ten jest jeszcze ciągle otwarty. W naszych polskich warunkach nie bez znaczenia jest czynnik opłacalności. W przyszłych modelach rozpoznających, które proponowane będą do szerszego zastosowania, w grę wejdą jedynie rozwiązania możliwie proste i równocześnie efektywne. Długo zapewne jeszcze zbyt drogie będą dla nas metody parametryzacji oparte o nowoczesną analizę sygnału mowy przy zastosowaniu predykcji liniowej, czy przekształcenia homomorficznego [3], [1].

Do wyniku przedstawionego poniżej doszedł autor zmierzając pierwotnie jedynie do udoskonalenia reprezentacji sygnału mowy w formie obrazów binarnych. Na tego typu reprezentacji oparto modele automatycznego rozpoznawania wyrazów przedstawione w pracy [7]. Praca niniejsza zawiera nową propozycję prostego opisu parametrycznego sygnału mowy. Zrodziła się ona w wyniku dotychczasowych własnych poszukiwań i doświadczeń w tej dziedzinie.

W poszukiwaniach tych kierowano się intencją sku-

pienia się na tej informacji zawartej w sygnale mowy, która wydaje się mieć priorytet w percepcji. Za punkt wyjścia przyjęto bardzo ogólnie sformułowaną zasadę o nasyceniu informacją w procesie percepcji elementarnego segmentu sygnału mowy.

2. Parametryzacja sygnału mowy na podstawie cech widmowych

Parametry stosowane w większości metod automatycznego rozpoznawania mowy zwykle odzwierciedlają różne cechy widmowe poszczególnych dźwięków mowy. Wynika to z oczywistego faktu, że cechy widmowe stanowią wykładnię cech dystynktywnych i na ich podstawie następuje identyfikacja poszczególnych dźwięków mowy przez człowieka. Parametryzacja ujmująca cechy widmowe w sposób globalny ma miejsce w metodzie predykacyjno-liniowej. We współczynnikach predykacyjnych $\{a_i\}$ zawarta jest informacja o całym widmie, tym pełniejsza, im większa jest liczba tych współczynników, czyli im wyższy jest stopień M wielomianu:

$$A(z) = \sum_{i=0}^{M} a_i z^{-i} \qquad (1)$$

wyrażającego charakterystykę filtru odwracającego sygnał mowy x(n) do postaci drgania pobudzającego e(n) . Wielomian A(z) spełnia więc następującą zależność:

$$A(z) \cdot X(z) = E(z), \qquad (2)$$

w której X(z) i E(z) reprezentują odpowiednio transformaty typu z sygnału mowy i źródła pobudzającego.

Prostsze metody parametryzacji sygnału mowy sięgają po wybrane parametry widmowe, takie jak: częstotliwości formantów, zakresy formantowe, momenty widmowe, energię sygnału w określonych pasmach częstotliwości i inne.

Niektóre z tych parametrów, jak np. częstotliwości formantów, w pewnych przypadkach nie dają się ekstrahować, w innych dokładność ich ekstrakcji jest mała. Momenty widmowe i

energia sygnału w określonych pasmach częstotliwości są uzależnione od preemfazy. Mankamentem parametru zdefiniowanego jako zakres formantowy jest jego znaczna płynność powodowana przez różne czynniki pozadystynktywne, takie jak poziom formantu czy częstotliwość podstawowa.

Fakt, iż każdy z wymienionych rodzajów parametryzacji w oparciu o cechy widmowe nie jest wolny od wad, skłania do poszukiwania parametrów odzwierciedlających inne cechy widmowe. Inspirację do takich poszukiwań potęguje chęć uzyskania sposobu parametryzacji prostszego od dotychczas znanych i umotywowanego przesłankami z wiedzy o czułości aparatu percepcyjnego człowieka na widmowe cechy dystynktywne dźwięków mowy.

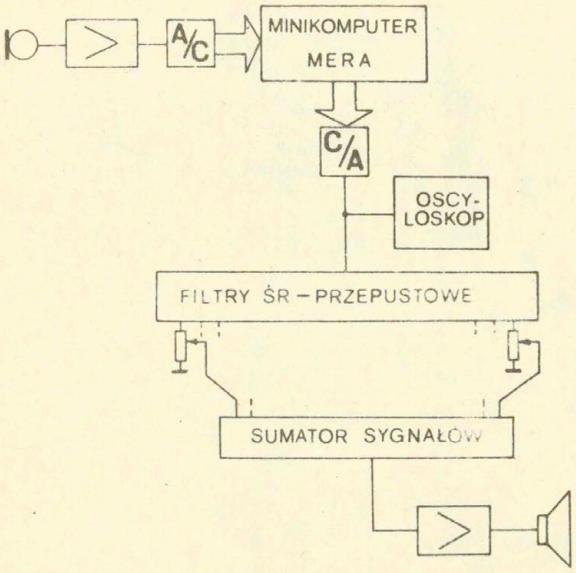
3. Rola cech widmowych w percepcji dźwięków mowy

Zagadnienie to pozostaje jeszcze ciągle nie w pełni wyjaśnione. Mowa jest zjawiskiem akustycznym bardzo złożonym. Złożoność ta polega na ogromnym bogactwie struktur widmowych, jakimi charakteryzuje się mowa w następujących po sobie krótkotrwałych segmentach czasowych, które określić można jako elementarne. Rozciągłość takich segmentów równa jest długości okresu drgania podstawowego na przestrzeni fragmentów dźwięcznych mowy, przyjmuje się natomiast dla niej stałą umowną wartość około 10 ms we fragmentach, w których źródłem pobudzania w procesie artykulacyjnym jest szum. Parametryzacja na podstawie cech widmowych dotyczy zwykle takiego elementarnego segmentu. Opis parametryczny ciągu tego rodzaju segmentów składających się na dowolną wypowiedź jest obrazem akustycznym tej wypowiedzi. Rozważania poniższe dotyczą segmentu rozumianego jak wyżej. Ocenę roli cech widmowych w percepcji dźwięków mowy komplikuje fakt, że trudno jest określić jednolite parametry widmowe obowiązujące w percepcji różnych dźwięków mowy. Szczegóły widmowe ważne dla percepcji pewnych dźwięków mowy są bez znaczenia dla percepcji innych dźwięków. Dźwięk mowy uznany powszechnie za jednostkę fonetyczną składa się z kilku segmentów elementarnych. Nawet

struktury kolejnych widm należących do jednego dźwięku nie są w pełni zgodne, chociaż w pewnych zakresach wykazują znaczne podobieństwo. Brak podobieństwa struktur widmowych sąsiednich segmentów próbują niektórzy badacze wykorzystać do segmentacji fonetycznej mowy [5]. Przypisuje się dużą rolę w percepcji dźwięków mowy tym cechom widmowym, które wyeksponowane zostały przez teorię wytwarzania mowy. Dotyczy to szczególnie częstotliwości formantów i zakresów formantowych głosek dźwięcznych, zwłaszcza samogłosek [4]. Parametry formantowe są jednak jednocześnie nośnikami cech osobniczych głosu i kłopotliwą rzeczą jest przeprowadzenie normalizacji sprowadzającej je dla różnych głosów do jednakowej skali. Fakt dość znacznych nieraz zróżnicowań międzyosobniczych parametrów formantowych w obrębie tej samej głoski poddaje w wątpliwość jednoznaczność opisu dźwięków mowy za pomocą tych parametrów. Wątpliwości te można by w znacznym stopniu wyjaśnić w drodze żmudnych badań polegających na generowaniu za pomocą syntezatora formantowego różnych odmian głoskowych i znajdowaniu współzależności pomiędzy wartościami cech widmowych a wynikami ocen percepcyjnych produktów syntezy.

Dla poszerzenia własnego poglądu o roli pewnych cech widmowych w percepcji mowy autor zmierza do przeprowadzenia odpowiednich eksperymentów polegających na ocenie percepcyjnej naturalnych dźwięków mowy, w których dokonana została eliminacja energii w wybranych pasmach częstotliwości. Krokiem w tym kierunku było zaadaptowanie wcześniej powstałego systemu analogowo-cyfrowego do podjęcia tego rodzaju eksperymentów. Zasadniczym elementem zmodyfikowanego systemu, którego schemat blokowy przedstawia rys. 1, jest blok analogowych filtrów środkowo-przepustowych zapożyczonych z wielokanałowego analizatora widma konstrukcji autora oraz sumator wyjść filtrów. Procedura wspomnianego eksperymentu za pomocą tego systemu jest następująca:

Zdigitalizowany przy pomocy przetwornika A/C sygnał mowy umieszcza się w pamięci operacyjnej minikomputera. Możliwe jest następnie wypisanie z komputera dowolnego fragmentu tego sygnału, a więc również fragmentu o rozciągłości jednej głoski i przekształcenie go ponownie do postaci analogowej.



Rys. 1. Schemat blokowy systemu do badań nad rolą cech widmowych w percepcji dźwięków mowy.

Wyizolowany fragment wypowiedzi wielogłoskowej zostaje przesłany przez wspomniany blok filtrów połączonych wejściami i wychodzących na sumator. W torach poszczególnych filtrów zamontowane są atenuatory pozwalające na regulację poziomu sygnału wyjściowego każdego filtru w granicach od 0 do 100%.

Przedstawiony system może posłużyć do określenia pasm sygnału mowy ważnych dla poprawnej identyfikacji tylko tych dźwięków mowy, których widmo na przestrzeni całego okresu trwania dźwięku nie ulega znaczącym zmianom. Przeprowadzenia

na szeroką skalę tego rodzaju eksperymentów badawczych nie planowano w ramach tej pracy. Wykonano jedynie próby sondażowe, które pozwoliły na stwierdzenie, iż pojedynczy dźwięk mowy daje się poprawnie identyfikować pomimo przeprowadzenia z nim szerokiej redukcji różnych zakresów widmowych. Utrata jego właściwej wartości fonematycznej następuje w wyniku eliminacji pewnych newralgicznych szczegółów widmowych. Dokładna znajomość tej współzależności pomogłaby właściwie sparametryzować sygnał mowy dla automatycznego rozpoznawania.

Analizując ważność różnych pasm widmowych dźwięku mowy dla zachowania jego wartości fonematycznej nasuwa się pogląd, iż z percepcyjnego punktu widzenia poszczególne pasma widmowe oddziaływują na siebie maskująco.

Istnieją priorytety tego maskowania, dzięki którym obecność lub brak energii w pewnych pasmach danego dźwięku mowy są dla jego poprawnej percepcji obojętne, natomiast sztucznie wywołany brak energii w innych pasmach tego dźwięku jest wyraźnie postrzegany uchem, a nawet powoduje utratę przez dźwięk swej wartości fonematycznej . Spojrzenie na rolę różnych cech widmowych w percepcji dźwięków mowy w oparciu o wyżej wyrażony pogląd być może dopomogłoby w ustaleniu właściwych zasad parametryzacji sygnału mowy.

4. Pojęcie nasycenia percepcyjnego

Widmo każdego dźwięku mowy posiada swój zakres, w którym przypadają cechy odzwierciedlające charakter dźwięku. Fakt ten znany jest z badań nad wyrazistością fonemów w różnych pasmach częstotliwości [2]. Początek tego zakresu dla wszystkich dźwięków periodycznych wyznaczony jest przez częstotliwość najniższej harmonicznej. Dla dźwięków szumowych przypada on wyżej i jest różny dla różnych rodzajów tego typu dźwięków mowy. Koniec tego zakresu jest także inny dla każdego typu dźwięku mowy. Nie pokrywa się on z końcem widma danego dźwięku. Wiadomo, że dla poprawnej identyfikacji samogłosek nie odgrywa roli część widma rozciągająca się powyżej drugiego formantu.

Załóżmy, że do oceny percepcyjnej podawany jest wielokrotnie, poprzez układ ograniczający od góry zakres widma, pewien dźwięk mowy. Na początku eksperymentu granica obcięcia ustawiona zostaje bardzo nisko, w pobliżu początku widma, po czym za każdym powtórzeniem dźwięku granicę tę podwyższa się. Poprawna identyfikacja tego dźwięku staje się możliwa dopiero od pewnej granicy obcięcia. Dalsze podnoszenie tej granicy nie ma już wpływu na wynik identyfikacji. Granica ta jest inna dla każdego dźwięku mowy. Można powiedzieć, iż nastąpiło coś w rodzaju nasycenia percepcyjnego, gdyż dawkowanie dalszych porcji informacji jest z punktu widzenia poprawnej identyfikacji dźwięku zbyteczne. Można twierdzić, że ta ilość informacji, która znalazła się poniżej granicy obcięcia, jest ważniejsza od reszty informacji. Zachodzi tu jak gdyby zdominowanie informacji zawartej w górnej części widma przez informację w dolnym zakresie.

Autor zainteresował się możliwością wykorzystania analogii do tego zjawiska przy opisie parametrycznym sygnału mowy. Znajomość wprost wspomnianej granicy nasycenia percepcyjnego dla poszczególnych dźwięków mowy nie miałaby dla tego opisu żadnego znaczenia. Parametr taki nie mógłby zostać użyty w automatycznym rozpoznawaniu mowy, gdyż jest on praktycznie niemierzalny. Granica, o której mowa, musiałaby wykazywać jakąś regularną współzależność ze zmierzalnymi cechami widmowymi sygnału mowy. Wyznaczenie takiej współzależności wymagałoby żmudnych badań i być może w wielu przypadkach byłoby niemożliwe.

5. Funkcja wagi widma

Zmierzając do parametryzacji sygnału mowy przy użyciu parametru pozostającego w analogii z wyżej omówioną granicą nasycenia percepcyjnego, autor wprowadził pojęcie wagi widma definiując ją następująco:

Wagą widma WW_i jest funkcja dyskretna wyrażająca proporcje dwóch sum, a mianowicie sumy rzędnych widma W_j pomniejszonych odpowiednio o sumy zmiennych przyrostów Δ_m,

oraz sumy rzędnych widma W_j pełnych. Matematycznym zapisem dej definicji jest wzor:

$$WW_i = \frac{\sum_{j=1}^{i}\left[W_j - \sum_{k=1}^{j} \Delta_{j-k}\right]}{\sum_{j=1}^{i} W_j} \qquad (3)$$

$\Delta_0 = 0$. i przyjmuje wartosci 1, ..., N, odpowiadające numeracji kolejnych rzędnych widmowych począwszy od najniższej. Ze wzrostem wartości indeksu licznik ilorazu we wzorze (3) przyrasta wolniej od mianownika, w wyniku czego wartość ilorazu maleje. Spadek wartości ilorazu (3) jest szybki, gdy widmo w dolnym zakresie częstotliwości jest pełne, tzn. że wszystkie jego składowe posiadają znaczące wartosci. Powolny spadek wartosci ilorazu (3) ma miejsce wówczas, gdy składowych znaczących w dolnym zakresie widma jest niewiele, lub gdy treść widmowa zaczyna się dopiero w górnym zakresie częstotliwosci.

Jak już powiedziano, składnikami sumy zewnętrznej w liczniku wyrażenia (3) są rzędne widma zmniejszone o sumy pewnych dodatnich zmiennych przyrostów Δ_m. Zakłada się, że wartosci poszczególnych przyrostów są odpowiednio proporcjonalne do poszczególnych rzędnych widmowych poprzedzających rzędną aktualnie dodawaną, czyli że:

$$\Delta_m = f(W_m). \qquad (4)$$

Zatem stopień pomniejszenia rzędnej zależy od wartości poprzednich rzędnych widmowych. Stanowi to analogię do przedstawionego wyżej zjawiska umniejszenia w procesie percepcji roli porcji informacji widmowej zlokalizowanej w miejscu i przez całą informację widmową zawartą w zakresie częstotliwości poniżej tego miejsca.

Brak jest precyzyjnych kryteriów, na podstawie których można by dokonać wyboru właściwego typu zależności (4). Nasuwają się jedynie wskazówki, aby ewentualnie przyrost Δ_m,

gdzie $m = j - k$, stawał się mniej zależny od rzędnej W_m w miarę wzrostu k, oraz aby zależność od W_m była zawsze liniowa. Pierwszej z tych dwóch wskazówek, ze względów czysto technicznych, chwilowo nie uwzględniono, uświadamiając sobie, że spadek zależności Δ_m od W_m ze wzrostem k powinien kształtować się niejednakowo w różnych zakresach widma. Powinien mianowicie, ze względu na liniową skalę częstotliwości analizatora, z którego pochodzą dane widmowe, stawać się ze wzrostem k wolniejszy w miarę przechodzenia w wyższy zakres widma. Spełnienie tego wymagania byłoby możliwe kosztem skomplikowania procedury wyznaczania funkcji wagi widma, w rezultacie czego funkcja ta przestałaby już być opłacalną w zastosowaniu do automatycznego rozpoznawania mowy.

Przyjmując zatem jedynie liniową zależność przyrostu Δ_m od rzędnej widma W_m, czyli

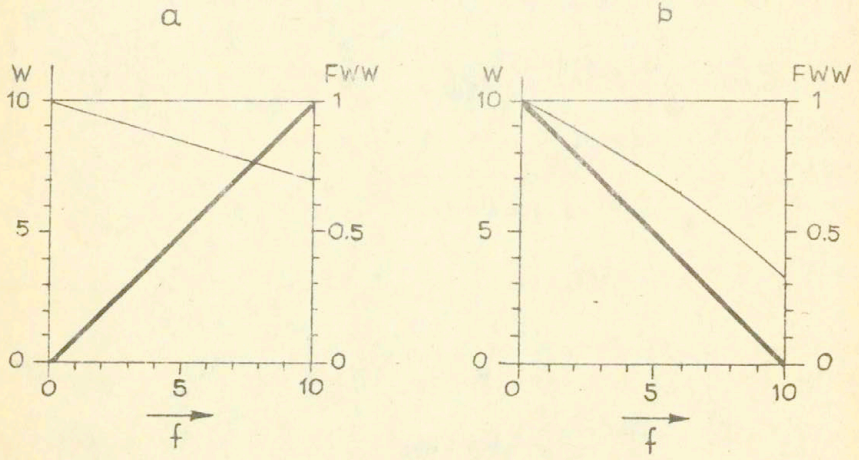
$$\Delta_m = q \cdot W_m$$

wzór (3) przekształcono do postaci:

$$WW_i = 1 - \frac{\overbrace{A_{i-1} + \overbrace{a_{i-1} + q \cdot W_{i-1}}^{a_i}}^{A_i}}{\underbrace{S_{i-1} + W_i}_{S_i}} \qquad (5)$$

sugerującej obliczanie wartości licznika i mianownika występującego w nim ilorazu w sposób rekursywny dla $i = 1, \ldots, N$. Warunki początkowe wynoszą: $A_0 = 0$, $a_0 = 0$, $W_0 = 0$, $S_0 = 0$.

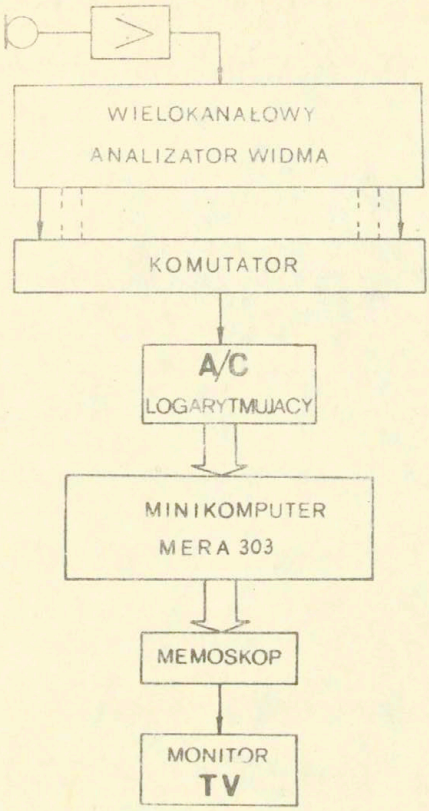
Na rys. 2 przedstawiono dla przykładu przebiegi funkcji wagi widma dla dwóch przypadków, a mianowicie, gdy widmo jest liniowo rosnące i malejące.



Rys. 2. Przebiegi funkcji wagi widma dla widma: a/ liniowo rosnącego, b/ liniowo malejącego.

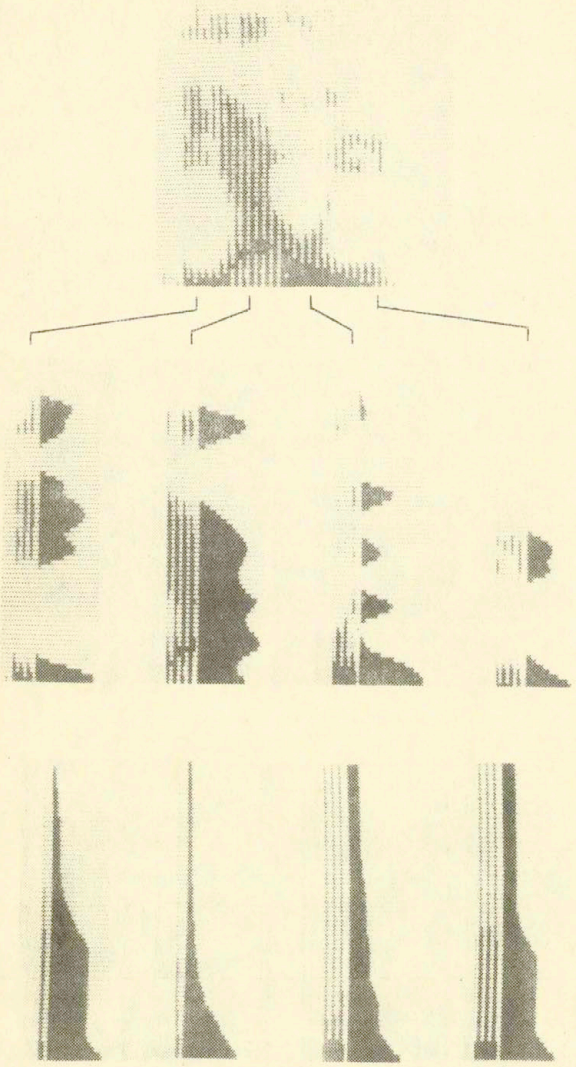
6. Funkcja wagi sygnału mowy

Do wyznaczenia funkcji wagi widma posłużono się systemem analogowo-cyfrowym, którego uproszczony schemat blokowy przedstawiono na rys. 3. Wyznaczenia widma sygnału mowy dokonywał analogowy analizator częstotliwości. Wynik analizy wpisany w systemie „on line" do minikomputera przechodził wpierw proces wygładzenia metodą podaną wcześniej przez autora (6). Uzyskano w ten sposób w zapisie cyfrowym w pamięci operacyj-



Rys. 3. Uproszczony schemat blokowy systemu analogowo-cyfrowego, za pomocą którego wyznaczono wartości funkcji wagi widma.

nej minikomputera ciąg widm pochodzących z wypowiedzi określonego wyrazu. Memoskop wchodzący w skład wspomnianego systemu umożliwiał obejrzenie oddzielnie wybranych widm w wersji dokładnej lub wszystkich widm równocześnie w wersji uproszczonej, tzn. skwantowanych z zastosowaniem czterech poziomów kwantyzacji amplitudowej. Następnie dla każdego widma wyznaczona została funkcja wagi. Również przy pomocy memoskopu można było obejrzeć w wersji dokładnej przebieg funkcji wagi dowolnie wybranych widm. Ciąg uproszczonych obrazów fun-

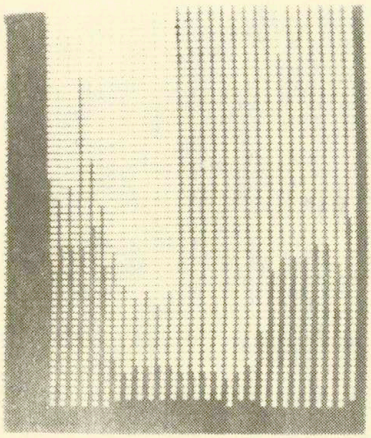


Rys. 4. 4-poziomowy spektrogram wyrazu „biały" oraz cztery wybrane widma wraz z wykresami ich funkcji wagi.

kcji wagi odnoszących się do poszczególnych widm całej wypowiedzi można było obejrzeć równocześnie. Zastosowano w tym przypadku także 4-poziomową kwantyzację wartości funkcji wagi widma.

Na rys. 4 przedstawiono spektrogram z kwantyzacją 4-poziomową wyrazu „biały" oraz obrazy czterech wybranych widm tego spektrogramu wraz z wykresami funkcji wagi widma.

Na rys. 5 zamieszczono obraz wykonany techniką podobną do tej, jaką otrzymano 4-poziomowy spektrogram z rys. 4, ale przedstawiający zamiast ciągu widm, jak ma to miejsce w przypadku spektrogramu, ciąg przebiegów funkcji ich wagi. Warunki techniczne nie pozwalały na uzyskanie tego typu obrazu z kwantyzacją lepszą . Obraz taki jest jedną z wielu możliwych form reprezentacji sygnału mowy, raczej dotychczas nie spotykaną.



Rys. 5. Obraz uproszczonch przebiegów funkcji wagi dla poszczególnych widm wyrazu „biały".

7. Parametryzacja sygnału mowy na podstawie funkcji wagi widma dla automatycznego rozpoznawania mowy

Nowy sposób parametryzacji będzie konkurencyjny wówczas, gdy cechować go będzie oczywista i korzystna przewaga nad dotychczas znanymi sposobami. Funkcja wagi widma takich cech wprost nie posiada. Zastąpienie parametrów widmowych wartościami funkcji wagi widma nie miałoby sensu. Opis parametryczny sygnału mowy dla jej automatycznego rozpoznawania powiniem być prosty, a tego postulatu nie spełnia sposób wykorzystujący wprost jako parametry wartości funkcji wagi widma.

Autor zdawał sobie z tego faktu sprawę wprowadzając pojęcie wagi widma. Niekorzystną cechą funkcji wagi widma jest np. to, że ma ona postać ilorazu. Wyznaczenie wartości tego typu funkcji wymaga operacji dzielenia, które w realizacji komputerowej jest czasochłonne.

Spodziewano się natomiast, że jako parametr przydatny w automatycznym rozpoznawaniu mowy, odegrać może rolę pewien charakterystyczny punkt funkcji wagi widma, na wzór rozpatrywanego w tej pracy pojęcia granicy nasycenia percepcyjnego. Punkt ten stanowiłby górną granicę zakresu widma, ważnego dla identyfikacji fonematycznej mowy. Przypisana by mu była określona wartość g funkcji wagi widma WW. Dla ważnego zakresu widma obowiązywałby zatem warunek:

$$WW > g \qquad (6)$$

Ponieważ funkcja WW ma postać ilorazu $\frac{L}{M}$, dla wyznaczenia granicy g nie trzeba wyliczać jego wartości, lecz jedynie określić od jakiego i /patrz wzór (5)/ spełniona zostaje nierówność:

$$L > g \cdot M \qquad (7)$$

Wartość g należy wybrać taką spośród wielu możliwych, aby maksymalnie różnicowała odmiennie segmenty mowy pod względem granicy ważnego zakresu widma, i aby gwarantowała powtarzalność tego rodzaju zróżnicowań.

Do opisu parametrycznego sygnału mowy dla automatycznego rozpoznawania mowy należałoby także wykorzystać dane o tych podzakresach ważnego zakresu widma, w których funkcja wagi widma utrzymuje stałą wartość.

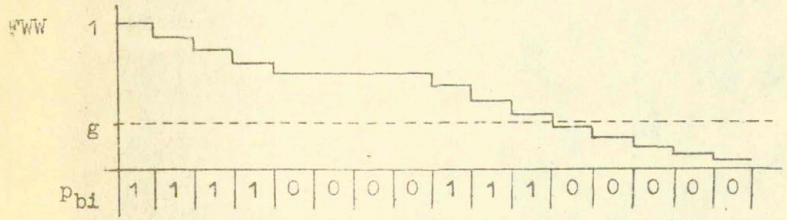
Wszystkie wyżej zaproponowane parametry związane z funkcją wagi widma można kodować w formie binarnej. Przynależność punktu widma do zakresu, w którym spełniona jest nierówność (6)można zakodować przez przypisanie temu punktowi wartości "1". Punktom widma leżącym poza tym zakresem nadać należy wówczas wartość Ø. Wartość Ø otrzymują również punkty widma przypadające w podzakresach ważnego zakresu widma, w których funkcja wagi widma jest niezmienna.

Przedstawioną zasadę binaryzacji opisu parametrycznego opartego o pojęcie funkcji wagi widma wyrazić można następującym zapisem:

$$\left.\begin{array}{l} b_i = 1 \text{ , jeśli } WW_i > g \wedge WW_i > WW_{i+1} \\ \\ b_i = 0 \text{ , jeśli } (WW_i > q \wedge WW_i = WW_{i+1}) \\ \qquad \vee \; WW_i \leqslant g \end{array}\right\} \text{ , } \qquad (8)$$

gdzie b_i symbolizuje binarną wartość przypisaną i-temu punktowi widma, i zmienia się od 1 do N , gdzie N jest liczbą punktów widma.

Na rys. 6 zilustrowano zasadę binaryzacji opisu parametrycznego sygnału mowy - na podstawie funkcji wagi widma.



Rys. 6. Przykład binaryzacji opisu parametrycznego wykorzystującej wybrane cechy funkcji wagi widma.

8. Wybór wartości q i g.

Praca obecna ma na celu udoskonalenie opisu parametrycznego sygnału mowy w formie obrazów binarnych w porównaniu z dotychczas stosowaną przez autora reprezentacją wypowiedzi wyrazu przez jego spektrogram binarny. Ponieważ obraz binarny utworzony na podstawie funkcji wagi widma zależy od dwóch czynników, a mianowicie od założonego współczynnika q określającego zależność przyrostu tłumiącego Δ_m od rzędnej widma oraz od wartości progu g, możliwy jest w założonych granicach dobór dla obu tych parametrów wartości optymalnych z punktu widzenia zbliżenia się do wytyczonego celu. Udoskonalenie reprezentacji sygnału mowy służyć ma rozszerzeniu zakresu rozpoznawanych elementów sygnału mowy w tym sensie, że zapewnić ma większą niezależność poprawności rozpoznawania od głosu operatora.

Przeprowadzono następujące badania, które z jednej strony wykazały, iż operując w automatycznym rozpoznawaniu wyrazów obrazem binarnym wypowiedzi utworzonym na podstawie wartości funkcji wagi widma możliwe będzie osiągnięcie większej niezależności wyniku rozpoznawania od cech głosu, a z drugiej strony pozwoliły dobrać optymalne wartości q i g z grupy rozpatrywanych.

Na podstawie funkcji wagi widma wyznaczono obrazy binarne wypowiedzi wyrazów "jemioła" i "julia" wypowiedziane przez jeden głos męski i jeden żeński. Pierwszy z tych wyrazów wypowiedziany był przez każdy głos dwukrotnie. Przy dotychczas stosowanej przez autora reprezentacji sygnału mowy rozpoznanie wypowiedzi danego wyrazu przez głos żeński na podstawie wzorca głosu męskiego było wykluczone. Dlatego jako materiału testowego użyto takiego zestawu głosów. Obrazy spektrograficzne wyrazów "jemioła" i "julia" charakteryzuje duża ilość ugięć formantowych oraz obecność wysokich formantów, szczególnie różnicujących głosy. Obraz binarny tworzono według zasady zdefiniowanej wzorami (8). Funkcja wagi odnosiła się do widma 63-punktowego. Przykłady otrzymanego w ten sposób obrazu binarnego wyrazu "biały" wypowiedzianego przez głos męski dla kilku wartości współczynnika q i progu g oraz spektrogram binarny tej wy-

powiedzi utworzony według zasad wcześniej opracowanych [7] podano na rys.7.

Dla 6-ciu par wartości q i g wyznaczono odległości pomiędzy obrazami:

- wypowiedzi tego samego wyrazu przez ten sam głos (zs),
- wypowiedzi tego samego wyrazu przez dwa głosy (zp),
- wypowiedzi różnych wyrazów przez ten sam głos (iws) i przez dwa głosy (iwp).

Ponacto te same warianty odległości wyznaczono dla tradycyjnych spektrogramów binarnych każdej z rozpatrywanych wypowiedzi.

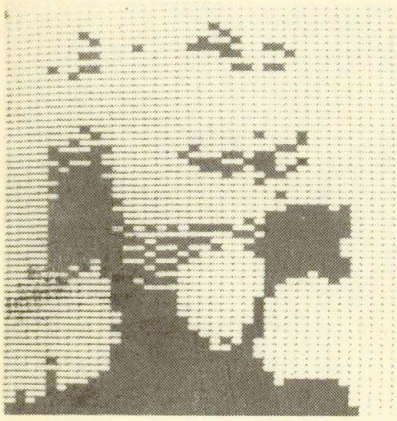
Jako odległość pomiędzy obrazami przyjęto średnie podobieństwo lokalne dwóch porównywanych obrazów binarnych. Użyto takiej samej miary podobieństwa, jaką posługiwano się w pracy [7]. Symbole rozpatrywanych odległości obrazów różnych wypowiedzi wyrazów zdefiniowano też w formie tablicy 1a. Prócz wymienionych już oznaczeń zs , zp , iws i iwp użyto w niej dodatkowo następujących symboli: WR1 i WR2 na oznaczenie wyrazów "jemioła" i "julia", WP1 i WP2 na oznaczenie pierwszej i drugiej wypowiedzi wyrazu "jemioła", Ż i M na oznaczenie głosów żeńskiego i męskiego. W tablicy 1b podano średnie z wartości odległości obrazów wypowiedzi dotyczących wariantów zakreślonych w tablicy 1a wspólnym konturem. Odległości dotyczące wariantów zs i iws uśredniono osobno dla każdego głosu.

Jako podstawę wyboru optymalnych wartości q i g przyjęto wartość ilorazu:

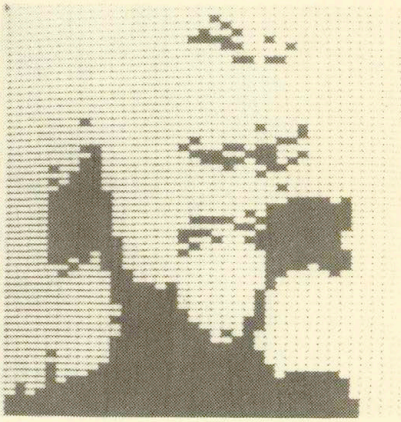
$$\frac{iws_{max}}{zp} ,$$

czyli stosunek maksymalnej spośród dwóch średnich odległości iws między obrazami wypowiedzi różnych wyrazów przez ten sam głos i średniej zp z odległości między obrazami wypowiedzi tych samych wyrazów przez różne głosy.

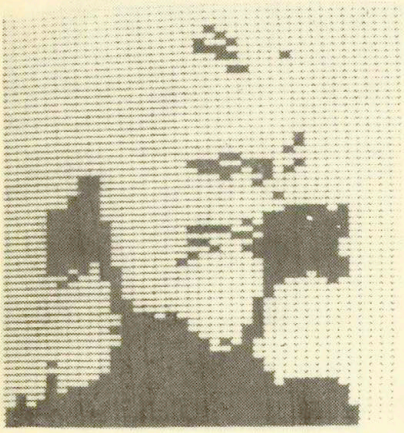
Posługując się takim kryterium uznano, iż spośród rozpatrywanych wartości q i g optymalną jest para: q=1/8 , g=0.30.



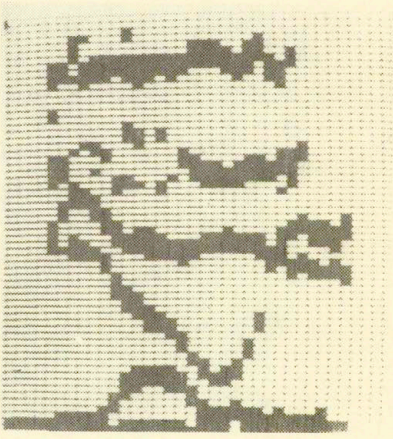
q = 1/8 , g = 0.17



q = 1/8 , g = 0.25



q = 1/8 , g = 0.30



SB

Rys. 7 Obrazy binarne uzyskane na podstawie funkcji wagi widma dla rozpatrywanych wartości q i g oraz spektrogram binarny wypowiedzi wyrazu "biały" głosem męskim.

Tablica 1a.
Znaczenia symboli odległości obrazów, dla których wyliczono średnie przedstawione w tablicy 1.

			Ż			M		
			WR1		WR2	WR1		WR2
			WP1	WP2		WP1	WP2	
Ż	WR1	WP1		zs	iws	zp	zp	iwp
		WP2	zs		iws	zp	zp	iwp
	WR2		iws	iws		iwp	iwp	
M	WR2	WP1	zp	zp	iwp		zs	iws
		WP2	zp	zp	iwp	zs		iws
	WR2		iwp	iwp		iws	iws	

Tablica 1b.
Odległości obrazów binarnych wypowiedzi testowych wyrazów "jemioła" i "julia" dla wybranych wartości q i g oraz analogiczne odległości tradycyjnych spektrogramów binarnych tychże wypowiedzi.

Lp.	Wariant		zs		zp	iws		iwp	iws_{max}/zp
	q	g	M	Ż		M	Ż		
1	1/16	0.42	7	17	15.1	20	24.25	24.62	1.60
2	1/16	0.37	7.5	17	15.25	24.75	24.75	24.17	1.62
3	1/16	0.30	7.5	17	15.85	21.25	25	24.75	1.57
4	1/8	0.30	6	12	12.6	18.75	21.5	22.12	1.7
5	1/8	0.25	6	12.5	13.35	18.7	21.75	23	1.62
6	1/8	0.17	7	13.5	14.25	19.5	23	23.75	1.61
7	SB		14.5	31.5	38.8	29.5	34.25	39.12	0.88

9. Wyniki testów rozpoznawania wypowiedzi wyrazów z zastosowaniem nowego rodzaju reprezentacji sygnału mowy i wnioski.

Celem przedstawionej w tej pracy próby reprezentacji wypowiedzi wyrazu w formie obrazu binarnego opartego o pojęcie funkcji wagi widma było uzyskanie sposobu parametryzacji pozwalającego na poszerzenie zakresu rozpoznawanych elementów mowy.

Przez poszerzenie rozumiano objęcie szerszego grona głosów możliwością automatycznego rozpoznawania wypowiedzi w oparciu o wspólny zbiór wzorców. Jak wiadomo większość układów rozpoznających wypowiedzi wyrazów adaptowanych jest na jeden głos lub grupę dobranych głosów. Taką cechę posiada też opracowany przez autora model automatycznego rozpoznawania wyrazów w oparciu o spektrogramy binarne.

Jak dalece przedstawiony nowy sposób parametryzacji sygnału mowy pozwala na poprawne rozpoznawanie wyrazów wypowiadanych przez głos, od którego nie pochodzą wzorce, ilustrują wyniki następującego eksperymentu. Wybrano 10 wyrazów, które stanowią minireprezentację języka polskiego, gdyż zawierają reprezentantów wszystkich grup fonemowych tego języka.

Są nimi:

1	sadze	6	źrenica
2	dżuma	7	giewont
3	hokej	8	zapis
4	dźwięki	9	rzeczy
5	flesz	10	baśnie

Wyrazy te wymówiły 4 głosy oznaczone przez: M1, M2, Z1, Z2. Dwa z tych głosów M1, M2 były męskie, a pozostałe dwa żeńskie. Posługując się modelem automatycznego rozpoznawania wyrazów operującym reprezentacją wyrazu w formie obrazu binarnego zmierzono odległości pomiędzy obrazami binarnymi wypowiedzi wszystkich dziesięciu wyrazów jednym głosem (umownie wyrazy te określono jako wzorce) a obrazami binarnymi wypowiedzi tych samych wyrazów przez pozostałe 3 głosy. Pomiary wykonano dla dwóch typów reprezentacji wypowiedzi wyrazu w formie obrazu binarnego,

a mianowicie dla obrazów wyznaczonych na podstawie funkcji wagi widma (ObB) z wartościami q = 1/8 i g = 0.30 uznanymi za optymalne w zbadanym zakresie oraz obrazów zwanych spektrogramami binarnymi (SB). Wyraz, którego wzorzec uzyskał najmniejszą odległość od obrazu testowanej wypowiedzi stanowił wynik rozpoznania.

W tablicach 2 i 2a oraz 3 i 3a zamieszczono uzyskane wyniki rozpoznawania. Tablice 2 i 2a dotyczą przypadku reprezentacji wypowiedzi wyrazu w formie obrazu binarnego uzyskanego z funkcji wagi widma, a tablica 3 i 3a odnosi się do przypadku stosowania spektrogramu binarnego jako formy reprezentacji wypowiedzi wyrazu. Liczby w tablicach 2 i 3 oznaczają ilości błędnych rozpoznań. Liczby w tej samej kolumnie, dotyczą jednego zbioru wypowiedzi wzorcowych głosem oznaczonym w nagłówku kolumny. Liczby w tym samym wierszu dotyczą natomiast wypowiedzi testowych jednym głosem porównywanych z wypowiedziami wzorcowymi pochodzącymi od różnych głosów. Podobnie liczby w tablicach 2a i 3a wyrażają procentowe poprawności rozpoznawania.

Łączna liczba błędów rozpoznawania dla przypadku reprezentacji wypowiedzi w formie nowych obrazów binarnych (przypadek ten oznaczono umownie przez NObB) wynosi:

13 na 120 wypowiedzi testowych przez 3 głosy,
głos 4-ty dostarczył wzorców .

Na taki wynik dominujący wpływ mają wypowiedzi głosu M1. Aż 11 błędów dotyczy tego głosu. Bez uwzględnienia tego głosu łączna liczba błędów wynosi:

2 na 60 wypowiedzi testowych przez 2 głosy,
głos 3-ci dostarczył wzorców .

Dla przypadku reprezentacji wypowiedzi w formie dotychczas stosowanych spektrogramów binarnych (przypadek ten oznaczono umownie przez DSB) błędy są znacznie liczniejsze:

36 na 120 wypowiedzi testowych przez 3 głosy,
głos 4-ty dostarczył wzorców .

Także w tym przypadku wynik globalny jest obciążony dużą liczbą błędów spowodowanych przez głos M1. Bez uwzględnienia tego gło-

NObB	Głos wypowiedzi wzorcowych M1	M2	Z1	Z2
Głos wypowiedzi testowych M1		4	4	0
M2	1		0	0
Z1	1	0		0
Z2	1	1	1	

Tablica 2.

Liczby błędnych identyfikacji w testach rozpoznawania z zastosowaniem nowego sposobu parametryzacji.

NObB	Głos wypowiedzi wzorcowych M1	M2	Z1	Z2
Głos wypowiedzi testowych M1		60	60	100
M2	90		100	100
Z1	90	100		100
Z2	90	90	90	

Tablica 2a.

Procentowa poprawność rozpoznawania uzyskana w testach z zastosowaniem nowego sposobu parametryzacji.

DSB	Głos wypowiedzi wzorcowych M1	M2	Z1	Z2
Głos wypowiedzi testowych M1		5	4	4
M2	3		1	5
Z1	2	1		0
Z2	2	7	2	

Tablica 3.

Liczby błędnych identyfikacji w testach rozpoznawania z zastosowaniem reprezentacji w formie spektrogramu binarnego.

DSB	Głos wypowiedzi wzorcowych M1	M2	Z1	Z2
Głos wypowiedzi testowych M1		50	60	60
M2	70		90	50
Z1	80	90		100
Z2	80	30	80	

Tablica 3a.

Procentowa poprawność rozpoznawania uzyskana w testach z zastosowaniem reprezentacji w formie spektrogramów binarnych.

su wynik ulega wprawdzie poprawie, gdyż całkowita liczba błędów wynosi:

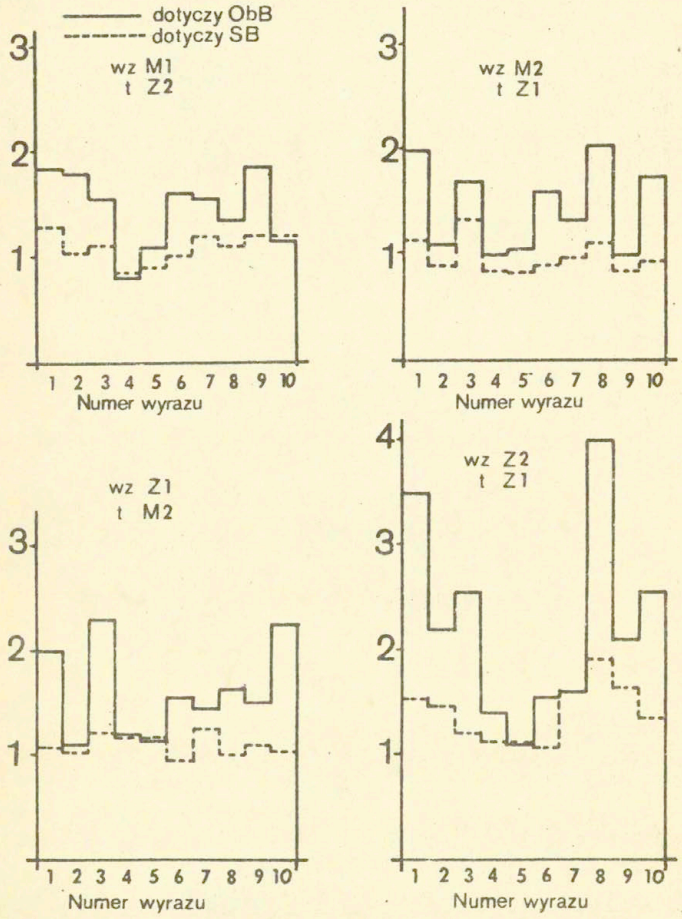
16 na 60 wypowiedzi testowych przez 2 głosy,
głos 3-ci dostarczył wzorców,

lecz w dalszym ciągu jest niekorzystny.

Wiadomo było, iż spektrogramy binarne reprezentują zarówno cechy dystynktywne wypowiedzi jak i cechy osobnicze głosu. Dlatego autor czynił próby znalezienia takiej reprezentacji wypowiedzi w formie obrazu binarnego, która w możliwym stopniu wolna byłaby od szczegółów wyrażających cechy osobnicze głosu. Przedstawione wyniki pozwalają wyrazić wniosek, że zaproponowany w tej pracy nowy rodzaj parametryzacji sygnału mowy w dużym stopniu ignoruje widmowe cechy osobnicze głosu i przez to poszerza krąg głosów mogących posługiwać się w automatycznym rozpoznawaniu wypowiedzi wyrazów wspólnym zbiorem wzorców. Średnio dla trzech spośród czterech badanych głosów uzyskano wyniki rozpoznawania przy użyciu obcych wzorców równe: 90%, 100%, 100%. Przy uwzględnieniu czwartego głosu dla dwóch głosów tego rodzaju wyniki wynosiły średnio po 96,5%, a dla pozostałych dwóch głosów uzyskano średnie poprawności rozpoznawania: 73% i 90%.

Analogiczne wyniki dla przypadku użycia spektrogramu binarnego jako formy reprezentacji wypowiedzi wyrazu są znacznie gorsze. Mianowicie dla trzech spośród czterech badanych głosów uzyskano następujące średnie poprawności rozpoznawania: 60% , 75%, 85%. Po uwzględnieniu czwartego głosu, bardzo zindywidualizowanego, średnia poprawność rozpoznawania wyniosła: 56,6%, 63%, 70%, 85%. Te wyniki stanowią jedynie odniesienie dla oceny nowego sposobu parametryzacji opartego o pojęcie funkcji wagi widma. Nie stanowią one dowodu krytycznego wobec parametryzacji sygnału mowy w oparciu o spektrogramy binarne, która nadaje się do stosowania w systemach automatycznego rozpoznawania wyrazów zależnego od głosu.

Przeprowadzono też ocenę odległości obrazów wypowiedzi testowych od obrazów wypowiedzi wzorcowych dla obu rozpatrywanych rodzajów reprezentacji sygnału mowy. Na rys. 8 zamieszczono wy-



Rys. 8. Wykresy wartości ilorazu odległości obrazu wypowiedzi testowej od właściwego wzorca i najbliższego wzorca obcego, dla rozpatrywanych wyrazów, dla dwóch rodzajów reprezentacji i czterech par głosów.

kresy wartości ilorazu odległości obrazu wypowiedzi testowej od obrazu właściwego wzorca i od obrazu drugiego najbliższego wzorca dla poszczególnych wyrazów, dla dwóch rodzajów reprezentacji wypowiedzi wyrazów i dla czterech par głosów. W każdej parze jeden głos dostarczył wzorców (na wykresie przed oznaczeniem tego głosu umieszczono litery wz) a drugi wypowiedzi testowych (na wykresie przed oznaczeniem tego głosu umieszczono literę t). Wykresy te ilustrują istnienie większej względnej dystynktywności pomiędzy obrazami wypowiedzi różnych wyrazów przez różne głosy dla nowej reprezentacji sygnału mowy w porównaniu z analogicznymi odległościami uzyskanymi dla przypadku użycia reprezentacji w formie spektrogramów binarnych.

Przytoczone wyniki świadczą, że przedstawiony w pracy nowy sposób parametryzacji sygnału mowy zapewnia większą niezależność obrazu wypowiedzi od cech indywidualnych głosu. Ponieważ elementami mowy są między innymi wypowiedzi izolowanych wyrazów przez różne głosy uznać należy, iż proponowany w pracy sposób parametryzacji pozwoli znacznie poszerzyć zakres rozpoznawanych elementów.

Streszczenie.

Praca przedstawia propozycję nowych parametrów widmowych sygnału mowy. Wprowadzono pojęcie funkcji wagi widma i zdefiniowano je w sposób matematyczny. Podano przykładowe przebiegi funkcji wagi dla widm wybranych fonemów. Wskazano na opłacalność użycia jako parametru widmowego szerokości zakresu widma, w którym funkcja wagi widma posiada wartość większą od pewnego założonego progu. Jako dodatkowy parametr zaproponowano szerokość najdłuższego fragmentu tegoż zakresu określonego przez niezmienną wartość funkcji wagi.

Przedstawiony w pracy nowy opis parametryczny sygnału mowy powinien skuteczniej neutralizować widmowe cechy indywidualne głosu. Potwierdzają to przytoczone wyniki porównania odległości obrazów reprezentujących wypowiedzi tego samego wyrazu przez różne głosy dla przypadku, gdy obrazy te mają formę spektrogramów binarnych oraz formę opartą o pojęcie funkcji wagi widma.

BIBLIOGRAFIA

[1] DAVIS, S.B., MERMELSTEIN, P.: Comparison of Parametric Representations for Monosyllabic Word Recognition in Continously Spoken Sentences, IEEE Trans. on ASSP, Vol.ASSP-28, 357-366, 1980.

[2] DUKIEWICZ, L., PIELA, R.: Szczegółowe badania wyrazistości i rozróżnialności głosek polskich w różnych warunkach przenoszenia. Cz. II: Wyrazistość głosek języka polskiego w zależności od górnej granicy częstotliwości, Biuletyn WAT, 4, 33-39, 1962.

[3] ITAKURA, F.: Minimum Prediction Residual Principle Applied to Speech Recognition, IEEE Trans. on ASSP, Vol.ASSP-23, No. 1, 67-72, 1975.

[4] JASSEM, W., KRZYŚKO, A., DYCZKOWSKI, A.: Klasyfikacja i identyfikacja samogłosek polskich na podstawie częstotliwości formantów, Prace IPPT 64/72, Warszawa 1972.

[5] JASSEM, W., KUBZDELA, H., DOMAGAŁA, P.: Automatic Acoustic-phonetic Segmentation of the Speech Signal, Acta Universitatis Umensis, From Sound to Words, Umea 1983.

[6] KUBZDELA, H.: Badania nad udoskonaleniem spektrogramów binarnych, Prace IPPT 24/83, Warszawa 1983.

[7] KUBZDELA, H.: Metoda globalnego rozpoznawania wyrazów na podstawie spektrogramów binarnych, Prace IPPT 28/86, Warszawa 1986.