Janusz Imiołczyk
Ryszard Ciarkowski

SYNTEZA OGRANICZONEGO ZBIORU WYRAZÓW POLSKICH ZE ZMIENNYM PARAMETREM F_0

12/1986

WARSZAWA 1986

ISSN 0208-5658

Praca wpłynęła do Redakcji dnia 22 listopada 1985 r.

N a p r a w a c h r ę k o p i s u

Instytut Podstawowych Problemów Techniki PAN

Nakład 160 egz. Ark.wyd. 1,4 Ark.druk. 2,5
Oddano do drukarni w marcu 1986 r.
Nr zamówienia 202/86.

Warszawska Drukarnia Naukowa, Warszawa,
ul. Śniadeckich 8

Janusz Imiołczyk
Ryszard Ciarkowski
Pracownia Fonetyki Akustycznej
IPPT PAN

SYNTEZA OGRANICZONEGO ZBIORU WYRAZÓW POLSKICH ZE ZMIENNYM PARAMETREM F_0 [1)]

Streszczenie

Dokonano syntezy czterech wyrazów polskich o zróżnicowanej liczbie sylab: "dał", "dobra", "normalny" i "naturalnie". Poprzez odpowiednie wysterowanie parametru F_0 w obrębie każdego z nich uzyskano 6 podstawowych dla języka polskiego typów intonacji: intonację niską, pełną i wysoką rosnącą, niską i pełną opadającą oraz równą (dla wyrazu "dobra" - dodatkowo niską i pełną rosnąco-opadającą). Każdy z przebiegów zrealizowano w wersji quasi-naturalnej oraz w trzech wzorowanych na niej wersjach opartych na aproksymacjach (aproksymacja łamaną, prostoliniowa i liniowo-schodkowa). Produkt syntezy, obejmujący 96 przebiegów intonacyjnych, poddano szczegółowej i uproszczonej ocenie percepcyjnej w badaniach odsłuchowych przeprowadzonych z udziałem dwu grup respondentów (łącznie 18 osób).

Niezależnie od opisu doświadczenia, praca zawiera również wnioski dotyczące syntezy intonacji jako jednego z istotnych aspektów globalnej syntezy mowy.

1. Techniczne podstawy syntezy intonacji.

Eksperyment z syntezą intonacji dzielił się na dwa etapy, stanowiące pod względem technicznym dwie odrębne części:

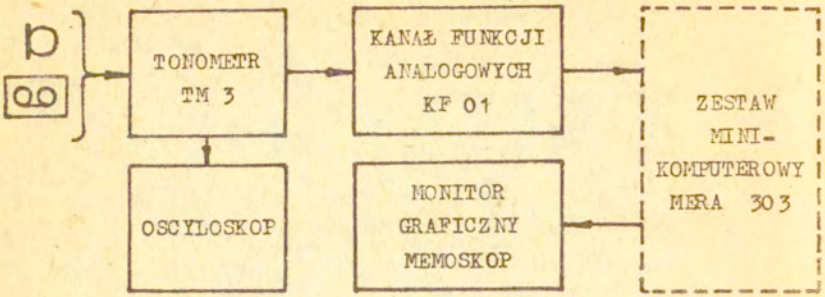
- analizę przebiegów parametru F_0 w wypowiedziach naturalnych,
- właściwą syntezę intonacji w procesie elektronicznej syntezy mowy.

1.1. Analiza przebiegów intonacyjnych.

Schemat układu pomiarowego częstotliwości podstawowej przedstawia rys.1.

---

1) Praca wykonana w ramach planu C-1



Rys.1. Schemat układu pomiaru i analizy parametru F_0.

Z nagranego na taśmie magnetofonowej materiału językowego dokonywano ekstrakcji parametru F_0 przy pomocy tonometru TM 3. Wyjściowy sygnał ekstraktora obserwowano na ekranie oscyloskopu z długą poświatą. Wynik ekstrakcji przesyłano w postaci ciągu impulsów do czasomierza (jeden z bloków funkcjonalnych kanału funkcji analogowych KF-01), gdzie dokonywał się cyfrowy pomiar okresu, a następnie do pamięci minikomputera MERA 303. Dane o wartości F_0 wprowadzano do pamięci co 10 ms. Po ich programowym przetworzeniu uzyskiwano wynik końcowy w trojakiej postaci:

- wydruku wartości częstotliwości podstawowej w odstępach 10-milisekundowych,
- wydruku dyskretnego wykresu przebiegu parametru F_0

lub

- przedstawienia w formie ciągłej przebiegu F_0 na ekranie monitora graficznego MEMOSKOP.

Dokładne omówienie sposobu analizy przebiegu parametru F_0 zawiera praca [8] (str.4-11).

1.2. Elektroniczna synteza mowy.

Elektroniczną syntezę czterech wybranych wyrazów przeprowadzono w oparciu o zestaw minikomputerowy MERA 303, obejmujący

jednostkę centralną MOMIK 8B o pamięci operacyjnej 8Kbajtów wraz z typowymi urządzeniami peryferyjnymi:
- zestawem: drukarka znakowo-mozaikowa DZM 180 i klawiatura operatorska (alfanumeryczna i numeryczna),
- modułem pamięci na dyskach elastycznych MDE-300,
- perforatorem (DT-105) i czytnikiem taśmy (CT 1001 A)

oraz wyspecjalizowanymi urządzeniami peryferyjnymi:
- 63-kanałowym analogowym analizatorem widmowym,
- kanałem funkcji analogowych KF-01 (zawierającym m.in. 8 bitowy przetwornik A/C),
- formantowym syntezatorem mowy COMPUTALKER CT-1

oraz
- monitorem graficznym MEMOSKOP.

Sterujący zestawem system programowy SPOS1 umożliwiał syntezę fragmentu mowy o czasie trwania do 910 ms oraz bieżącą analizę widmową mowy syntetycznej i naturalnej, która prowadziła do uzyskania spektrogramu wyświetlanego na ekranie monitora ([3], [5]). W przypadku syntezy wyrazów o długości większej niż 910 ms korzystano z innej wersji systemu, umożliwiającego syntezę mowy o czasie trwania do 3 s przy rezygnacji z analizy syntezowej mowy.

2. Synteza intonacji.

2.1. Podstawowe składniki akcentu.

Istnieje dość powszechne przekonanie, że zasadniczym fizjologicznym korelatem zjawiska akcentu w mowie naturalnej jest wzrost ciśnienia podkrtaniowego (por.np. [15], [16], [20]). Wzrost ten pociąga za sobą podwyższenie chwilowych wartości częstotliwości podstawowej oraz intensywności w obrębie samogłoski stanowiącej centrum sylaby akcentowanej [1]. Trzecim czynnikiem, którego zmiany wpływają w pewnym stopniu na percepcję akcentu jest czas trwania samogłoski ([10], [11], [15], [19]).

W dosyć obszernej literaturze przedmiotu uznaje się

[1] brak jest ścisłej korelacji między zmianami ciśnień podkrtaniowych a czasem trwania samogłoski akcentowanej

na ogół, że na zjawisko akcentu składają się zmiany wszystkich trzech z wymienionych wyżej parametrów akustycznych (m.in. [2], [9], [15]). Nie wszyscy jednak autorzy przypisują tym parametrom (zwłaszcza F_0 i iloczasowi) jednakowo ważną rolę. W oparciu o wyniki prac eksperymentalnych większość z nich stwierdza, że dla percepcji akcentu najistotniejsze są zmiany częstotliwości podstawowej ([10], [14], [15], [21]). Spotyka się jednak także pogląd według którego głównym sygnałem akcentu jest długość samogłoski ([2]). Większa zgodność opinii zaznacza się natomiast w odniesieniu do trzeciego z rozważanych czynników: intensywność jest dość powszechnie uznawana za najmniej istotny składnik akcentu ([2], [14] , [15]).

Ze względu na fakt, iż główny przedmiot zainteresowania w niniejszej pracy stanowił wpływ przebiegu częstotliwości podstawowej (F_0) na percepcję akcentu (wyrazowego), przedstawione poniżej wnioski prac eksperymentalnych koncentrują się wokół tego właśnie parametru[1)].

Wśród prawidłowości rządzących percepcją intonacji na plan pierwszy wysuwa się fakt, iż zmiany melodii wypowiedzi postrzegane są wyłącznie w obrębie segmentów samogłoskowych; rola przebiegu F_0 w segmentach spółgłoskowych jest nieznacząca ([9] , [19]). Sam kształt przebiegu F_0 nie zależy przy tym istotnie od dźwięczności lub bezdźwięczności spółgłosek tworzących wypowiedzi (por. [13], 196) , co świadczyłoby o tym, że zmiany wysokości tonu nie są zjawiskiem segmentalnym.

Z punktu widzenia syntezy intonacji istotna jest konkluzja, iż przebiegi F_0 można rozpatrywać zarówno w kategoriach konturów, jak i sekwencji tonów statycznych ([13] , [20]). Oba te podejścia są jednakowo zasadne i dają bardzo zbliżone rezultaty.

Przy analizie funkcji językowych intonacji, zwłaszcza w badaniach percepcyjnych, przyjmuje się najczęściej podział

---

1) Zagadnienia związane z intonacją jednostek językowych o rozciągłości większej niż wyraz (fraza, zdanie) wykraczają poza zakres niniejszej pracy.

jej przebiegów na dwa zasadnicze typy, wyróżniając wypowiedzi: 1) niezakończone (pytajne i kontynuatywne) oraz 2) zakończone (oznajmienia, żądania, pytania o uzupełnienie, por.[10], 30, 31 i 89)[1]. Wypowiedzi pierwszego typu cechuje na ogół intonacja rosnąca, drugiego zaś- opadająca. Przy percepcyjnej identyfikacji każdego z tych typów odgrywają rolę trzy podstawowe czynniki:

1) poziom częstotliwości podstawowej F_0 w punkcie bezpośrednio poprzedzającym jej wzrost bądź spadek na końcu wypowiedzi,

2) kierunek tej zmiany wysokości oraz

3) jej zakres.

Przykładowo, subiektywne wrażenie pytania o rozstrzygnięcie jest tym wyraźniejsze, im większy jest wzrost F_0 na końcu wypowiedzi ([12], 177-78, [17] , 452) oraz im wyższa jest wartość F_0 w punkcie bezpośrednio poprzedzającym zmianę kierunku tonu ([12] , 180) . Wpływ wartości F_0 w punkcie przegięcia na powstanie takiego wrażenia może być istotniejszy niż wpływ końcowego wzrostu F_0, zwłaszcza gdy ten drugi cechuje stosunkowo niewielki zakres (por.[12], 180, [17], 455).

Subiektywne wrażenie akcentu jest ponadto tym silniejsze, im szybszy jest wzrost F_0 (lub im mniej stromy jest jej spadek) w obrębie samogłoski sylaby akcentowanej ([19], 34). Niewielkie znaczenie przy klasyfikacji danej intonacji w terminach "pytanie" - "odpowiedź" ma natomiast przebieg F_0 w początkowym fragmencie wypowiedzi ([17], 456).

2.2. Synteza wyrazów. Zasady doboru czasów trwania i przebiegów amplitudy.

Dla celów niniejszego doświadczenia dokonano syntezy czterech następujących wyrazów o zróżnicowanej liczbie sylab: dał, dobra, normalny i naturalnie[2].

Sposób syntezy był analogiczny jak w dwóch poprzednich doświadczeniach ([4], [6]) i nie będzie tu przedmiotem odrębnego opisu. Na przypomnienie zasługuje natomiast fakt, iż percepcyjny efekt akcentu wyrazowego został w obu wymienionych pracach uzyskany

---

[1] por. także [19], 32

[2] Łącznie zsyntezowano pięć wyrazów, lecz jeden z nich - dalej- nie został wykorzystany w przeprowadzonych później badaniach odsłuchowych.

bez różnicowania przebiegów częstotliwości podstawowej, a jedynie poprzez dobór odpowiednich czasów trwania oraz przebiegów amplitudy w obrębie samogłosek akcentowanych. W identyczny sposób postąpiono również w odniesieniu do wyrazów daj, dobra, normalny i naturalnie : punktem wyjścia do syntezy różnych przebiegów intonacyjnych było nadanie każdemu z tych wyrazów akcentu "iloczasowo-amplitudowego".

Ze względu na istotność informacji dotyczących sposobu kształtowania przebiegów amplitudy oraz zasad doboru czasów trwania samogłosek akcentowanych w kontekście rozważanego tu problemu akcentu wyrazowego, poniżej omówione zostaną odpowiednie zasady przyjęte zarówno w obu wcześniejszych pracach, jak i w niniejszej.

Zmiany amplitudy w torze formantowym miały niemal we wszystkich przypadkach charakter wyłącznie segmentalny: w obrębie stanu ustalonego danej głoski (w niniejszym materiale z wyjątkiem /b/, /d/, /t/ i /r/) amplituda miała równą, niezróżnicowaną wartość. (Pominięty jest tu przypadek narastania i spadku amplitudy na końcu i początku wypowiedzi.) Ze względu na zaniedbywaną w literaturze (por. str. 5), stosunkowo niewielką rolę intensywności w sygnalizacji akcentu, również w obrębie samogłosek akcentowanych amplituda była "płaska" (identyczną zasadę przyjęli w syntezie intonacji m.in. Abramson [1], str.321 i Mattingly [18], 2) [1]. Poziom amplitudy w obrębie samogłoski akcentowanej wynosił:

w wyrazie daj - 16 dB [2]
w wyrazie dobra - 20 dB
w wyrazie normalny - 18 dB
w wyrazie naturalnie - 16 dB.

Jako podstawę do ustalenia optymalnych długości poszczególnych głosek w ramach czterech wybranych wyrazów przyjęto wyniki pracy [22](str.22). Główny nacisk położono na dobranie odpowiednich czasów trwania dla samogłosek akcentowanych, dostosowując do nich następnie iloczasy pozostałych segmentów

---

1) Optymalny (tzn. dający efekt akcentu) poziom amplitudy w obrębie samogłoski akcentowanej ustalany był w oparciu o bieżącą ocenę percepcyjną.

2) względem przyjętego poziomu odniesienia 0 dB.

głoskowych. W oparciu o bieżącą ocenę słuchową ustalono, że optymalna długość stanu ustalonego samogłosek akcentowanych w czterech zsyntezowanych wyrazach wynosiła:

w wyrazie	<u>dał</u>	- 190 ms
	<u>dobra</u>	- 130 ms
	<u>normalny</u>	- 120 ms
	<u>naturalnie</u>	- 90 ms

(przy całkowitej długości każdego z wyrazów równej odpowiednio: 480 ms, 670 ms, 910 ms i 1060 ms).

Przedstawione dalej wykresy (rys.2) ilustrują przebiegi amplitudy oraz iloczasy głoskowe dla poszczególnych wyrazów syntetycznych.

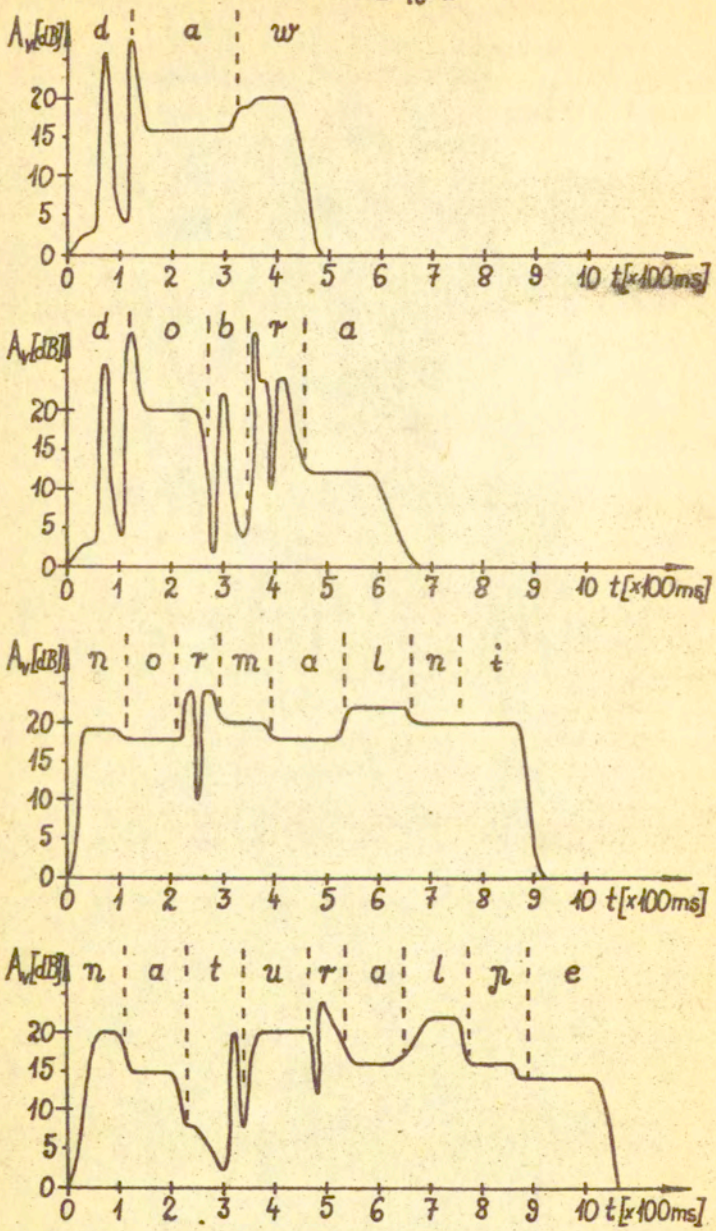
2.3. <u>Opis doświadczenia z syntezą intonacji</u>.

Kolejny, zasadniczy etap pracy obejmował syntezę przebiegów intonacyjnych. Bez względu na typ przebiegu F_0 żaden z pozostałych parametrów sterujących syntezą (w tym także oba parametry odpowiedzialne za sygnalizowanie akcentu, t.j. amplituda i iloczas) nie był modyfikowany. U podstaw tej decyzji tkwiło m.in. założenie (którego słuszność potwierdziły badania odsłuchowe), że występujące w mowie naturalnej zróżnicowanie czasu trwania samogłoski akcentowanej w zależności od typu przebiegu F_0 (por.np. [9], 5)nie powinno mieć istotnego wpływu na percepcję akcentu i poprawność typu intonacyjnego.

Z inwentarza intonemów mowy polskiej zaproponowanego przez Steffen-Batogową (zob. [10], str.29) wybrano dla celów syntezy 8 przebiegów odpowiadających najbardziej typowym intonacjom wyrazowym [1]:

(1) niski rosnący (NR)
(2) pełny rosnący (PR)
(3) wysoki rosnący (WR)
(4) równy (R)
(5) pełny opadający (PO)
(6) niski opadający (NO)
(7) pełny rosnącą-opadający (PRO)
(8) niski rosnącą-opadający (NRO)

[1] Nie uwzględniono podziału przebiegów rosnąco-opadających na mocno rosnąco-opadające i słabo rosnąco-opadające, traktując je jako jedną klasę.



Rys.2. Wykresy przebiegów amplitudy A_v wyrazów syntetycznych : dał, dobra, normalny i naturalnie.

Kolejny krok stanowiło nagranie naturalnych wersji wyrazów dał, dobra, normalny i naturalnie, wymówionych we wszystkich uwzględnianych intonacjach przez jedną osobę (niski głos męski) o dużym doświadczeniu fonetycznym. W związku z tym, że przebiegi (7) i (8) brzmiały w wymówieniach wyrazów dał, normalny i naturalnie nieco sztucznie oraz że ich występowanie w mowie polskiej jest niezbyt częste, postanowiono wykorzystać je wyłącznie w syntezie intonacji w wyrazie dobra.

Przebiegi parametru F_0 we wszystkich wypowiedziach naturalnych zanalizowano przy użyciu tonometru TM 3, uzyskując ciągi bezwzględnych wartości tego parametru (w Hz) w przedziałach 10-milisekundowych. Niezależnie od tego, dokonano wąskopasmowej analizy spektrograficznej poszczególnych wypowiedzi. W oparciu o otrzymane spektrogramy dokonano segmentacji wyrazów i ustalono kształt przebiegów F_0 w obrębie poszczególnych głosek, jak również zweryfikowano dane będące efektem analizy tonometrycznej[1].

We wstępnym etapie syntezy intonacji oparto się na wartościach F_0 wyekstrahowanych z wypowiedzi naturalnych[2]. Przy pierwszych próbach okazało się jednak, że w niskich zakresach częstotliwości (poniżej 90 Hz), które osiągnął głos "wzorcowy" w intonacjach opadających, efekt akustyczny uzyskiwany na wyjściu syntetyzatora nie był zadawalający. Teoretyczny zakres częstotliwości podstawowej możliwej do generowania w syntetyzatorze COMPUTALKER CT-1 wynosi 74 - 460 Hz, jednakże przy wartościach F_0 mniejszych niż 90 Hz brzmienie głosu syntetycznego cechuje niska jakość (efekt "chrypki"). Za średnią, typową dla głosu męskiego można przyjąć wartość ok. 120 Hz. W tej sytuacji uznano za konieczne podwyższenie wyjściowych wartości F_0 o interwał czterech półtonów. Synteza w oparciu o te przetworzone wartości dała zdecydowanie lepsze rezultaty.

---

1) Na spektrogramie wąskopasmowym (szerokość pasma 45 Hz) uwidocznione są kolejne składowe harmoniczne, których kształt ilustruje przebieg intonacji w obrębie wypowiedzi.

2) Ponieważ różnice w długości głosek i w długości całkowitej były między wyrazami naturalnymi i syntetycznymi stosunkowo niewielkie, zachodziła jedynie konieczność dokonania prostej normalizacji czasowej.

Quasi-naturalne intonacje syntetyczne stały się punktem wyjścia do opracowenia trzech następujących typów aproksymacji przebiegów F_0 :

1) aproksymacji łamaną (P) składającą się z trzech bądź czterech odcinków

Jest to najpowszechniej stosowany rodzaj aproksymacji (por. [7] str.463, [17]str.451). Aproksymacji czteroodcinkowej użyto w intonacjach rosnąco-opadających; aproksymację intonacji równej stanowiła oczywiście linia prosta.

2) aproksymacji prostoliniowej (P1)

Zmiana wysokości F_0 (wzrost lub spadek) w przedziale wartości określanym na podstawie przebiegu quasi-naturalnego jest w tej aproksymacji realizowana równomiernie na całej długości wyrazu.

oraz 3) aproksymacji liniowo-schodkowej (P2)

Ten typ aproksymacji polegał na prostoliniowej realizacji przebiegu intonacji na obranych dwóch wysokościach F_0 i raptownej zmianie (10 Hz/10 ms) pomiędzy nimi. Zmiany dokonywano w obrębie sylaby akcentowanej.

Wybór wymienionych wyżej typów aproksymacji podyktowany został przez konieczność stworzenia możliwości automatycznej realizacji przebiegów intonacyjnych w syntezie. Wykorzystanie aproksymacji wyrażonych funkcjami matematycznymi uznano za zbyt złożone dla praktycznej realizacji. Skoncentrowano się na prostych aproksymacjach, przyjmując kompromis pomiędzy naturalnością brzmienia a wymaganiami automatycznej realizacji (złożoność procedur aproksymacyjnych).

Ogółem dokonano syntezy 96 przebiegów (pominięto aproksymacje P i P2 dla intonacji równej)[1]. Całkowity zakres częstotliwości podstawowej wykorzystany w syntezie wynosił od 77 Hz do 250 Hz. Te skrajne wartości wystąpiły odpowiednio: w aproksymacjach P1 przebiegów opadających i pełnych rosnących. W zdecydowanej większości przypadków F_0 nie osiągała jednakże

[1] Przeprowadzona dodatkowo próba "podłożenia" wartości F_0 z poszczególnych przebiegów wyrazu dobrze w wyrazie dalej wykazała, że ze względu na zbliżoną strukturę fonetyczną obu wyrazów możliwa była niemal bezpośrednia ekstrapolacja.

wartości poniżej 88 Hz i powyżej 240 Hz.

Kształt poszczególnych przebiegów w czterech wyrazach syntetycznych ilustrują wykresy zamieszczone w Dodatku do niniejszej pracy.

3. Percepcyjna ocena intonacji syntetycznych.

Podstawę do zobiektywizowanej oceny jakości syntetycznych przebiegów intonacji stanowiły wyniki badań odsłuchowych przeprowadzonych z udziałem 18 respondentów podzielonych na dwie grupy (A i B). Zadanie grupy A, w skład której wchodziło 8 osób mających doświadczenie zawodowe w zakresie melodii mowy (6 fonetyków oraz 2 muzyków), polegało na dokładnym określeniu typu prezentowanego przebiegu intonacji, tj. na podaniu, czy usłyszany wyraz syntetyczny "wymówiony" został z intonacją niską rosnącą, pełną opadającą, równą itd. Odpowiedzi udzielono poprzez wpisanie odpowiedniego oznaczenia (NR, PO, R itd.) na liście, zawierającej zapis poszczególnych wyrazów w (losowej) kolejności ich występowania na taśmie odsłuchowej.

Grupa B, złożona z 10 "naiwnych" respondentów (osób bez przygotowania w zakresie intonacji), miała za zadanie zakwalifikować poszczególne wypowiedzi syntetyczne jako pytania bądź jako odpowiedzi, wpisując na listach symbole P (pytanie) lub O (odpowiedź).

3.1. Metoda prezentacji bodźców.

Nagranie dokonane dla celów badań odsłuchowych zawierało część adaptacyjną, na którą składały się występujące w 2-sekundowych odstępach wyrazy z intonacjami quasi-naturalnymi, oraz właściwy materiał testowy, obejmujący wszystkie zsyntetyzowane przebiegi, zarejestrowane w porządku losowym i oddzielone 6-sekundowymi pauzami.

Część adaptacyjną nagrania odtworzono bezpośrednio przed właściwą prezentacją. Jej użycie miało na celu oswojenie respondentów z brzmieniem mowy syntetycznej, niecc odbiegającym od brzmienia mowy naturalnej, oraz zapoznanie ich ze skalą zmian intonacyjnych. Badania odsłuchowe poprzedzano każdorazowo instrukcją, mającą wyjaśnić respondentom istotę eksperymentu i ich w nim zadanie.

Odsłuchy przeprowadzono w pomieszczeniu charakteryzującym się pewnym pogłosem i dość znacznym tłem dźwiękowym otoczenia (przejeżdżające w niewielkiej odległości samochody i pociągi, dobiegające z zewnątrz odgłosy rozmów itp.) przy użyciu toru akustycznego, w skład którego wchodziły:

- stereofoniczny magnetofon kasetowy M601 SD
- stereofoniczny wzmacniacz PA-107

oraz

- zestawy głośnikowe ZG-10-C.

3.2. Wyniki.

3.2.1. Grupa A.

Wyniki rozpoznawania przebiegów intonacyjnych przez grupę A przedstawione są w Tabelach 1,2,3 i 4. Na ogólną liczbę 768 odpowiedzi (96 przebiegów x 8 osób) zanotowano 532 odpowiedzi poprawne (ok. 69 %). Poprawność identyfikacji przebiegów intonacyjnych w obrębie czterech wyrazów syntetycznych była najwyższa w przypadku aproksymacji liniowo-schodkowej (76 %), najniższa zaś - w przypadku aproksymacji prostoliniowej (63 %). Przebiegi aproksymowane łamaną i quasi-naturalne zostały rozpoznane odpowiednio w 71 i 69 procentach przypadków.

Spośród ośmiu typów zsyntetyzowanych przebiegów najlepiej została rozpoznana intonacja niska rosnąca (84 % poprawnych odpowiedzi), najgorzej zaś - intonacja pełna rosnąco-opadająca (19 %). Rezultaty uzyskane dla pozostałych przebiegów przedstawiają się następująco:

- niski opadający: 77 %
- równy: 75 %
- pełny opadający: 73 %
- wysoki rosnący : 69 %
- pełny rosnący: 63 %
- niski rosnąco-opadający: 34 %

Poprawność identyfikacji przebiegów intonacyjnych w obrębie poszczególnych wyrazów była także zróżnicowana: dla wyrazów normalny i naturalnie uzyskano nieco lepsze wyniki (odpowiednio: 81 % i 77 %) niż dla wyrazów dał i dobra (odpowiednio: 68 % i 56 %).

DOBRA

N \ O	NR	PR	WR	NO	PO	R	NRO	PRO
NR	7					1		
PR		7	1					
WR	1	1	6					
NO				8				
PO				8				
R				1		7		
NRO				2	1		5	
PRO				1	2		1	4

DAŁ

N \ O	NR	PR	WR	NO	PO	R
NR	6	1				1
PR	3	5				
WR	4		4			
NO				6	2	
PO				3	4	1
R	1					7

NORMALNY

N \ O	NR	PR	WR	NO	PO	R
NR	7	1				
PR	1	5	2			
WR	3	3	2			
NO				8		
PO				1	7	
R	1		1	1		5

NATURALNIE

N \ O	NR	PR	WR	NO	PO	R
NR	7					1
PR	2	5	1			
WR	1	1	6			
NO				5	2	1
PO				1	7	
R	5					3

N \ O nadano \ odebrano

typy intonacji:

NR niska rosnąca
PR pełna rosnąca
WR wysoka rosnąca
NO niska opadająca
PO pełna opadająca
R równa
NRO niska rosnąco-opadająca
PRO pełna rosnąco-opadająca

Tabela 1. Wyniki rozpoznawania przebiegów intonacyjnych dla realizacji quasi-naturalnej.

DOBRA

N \ O	NR	PR	WR	NO	PO	R	NRO	PRO
NR	7						1	
PR	1	6	1					
WR	1	1	6					
NO	1			4		1	1	1
PO				4	3			1
NRO				4		1	3	
PRO				1	5		1	1

DAŁ

N \ O	NR	PR	WR	NO	PO
NR	7			1	
PR	2	5	1		
WR	2	1	5		
NO				6	2
PO				3	5

NORMALNY

N \ O	NR	PR	WR	NO	PO
NR	8				
PR		8			
WR	1	2	5		
NO				8	
PO					8

NATURALNIE

N \ O	NR	PR	WR	NO	PO
NR	8				
PR	1	5	2		
WR	2	3	3		
NO				6	2
PO					8

N \ O nadano \ odebrano

typy intonacji:

NR niska rosnąca
PR pełna rosnąca
WR wysoka rosnąca
NO niska opadająca
PO pełna opadająca
R równa
NRO niska rosnąco-opadająca
PRO pełna rosnąco-opadająca

Tabela 2. Wyniki rozpoznawania przebiegów intonacyjnych dla realizacji aprokaymacji Łamana (P).

DOBRA

N \ O	NR	PR	WR	NO	PO	R	NRO	PRO
NR	5					3		
PR	2	2	4					
WR		2	6					
NO	1			4		1	1	1
PO				2	4		1	1
R		1		1		5	1	
NRO				3	1	1	3	
PRO			1		6			1

DAŁ

N \ O	NR	PR	WR	NO	PO	R
NR	3					5
PR	2	1	5			
WR	2		5			1
NO				6	1	1
PO				2	6	
R				2		6

NORMALNY

N \ O	NR	PR	WR	NO	PO	R
NR	6					2
PR	1	3	4			
WR	2		6			
NO				7	1	
PO				1	7	
R				1		7

NATURALNIE

N \ O	NR	PR	WR	NO	PO	R
NR	5		1			2
PR		5	3			
WR	1		7			
NO				6		2
PO				1	7	
R						8

N \ O nadano \ odebrano

typy intonacji:

NR niska rosnąca
PR pełna rosnąca
WR wysoka rosnąca
NO niska opadająca
PO pełna opadająca
R równa
NRO niska rosnąco-opadająca
PRO pełna rosnąco-opadająca

Tabela 3. Wyniki rozpoznawania przebiegów intonacyjnych dla realizacji aproksymacji prostoliniowej (P1).

DOBRA

N \ O	NR	PR	WR	NO	PO	R	NRO	PRO
NR	8							
PR		7	1					
WR	1	1	6					
NO				5		1	1	1
PO				2	5			1
NRO				8				
PRO					6		2	

DAŁ

N \ O	NR	PR	WR	NO	PO
NR	7				1
PR	1	5	2		
WR	1		7		
NO				5	3
PO					8

NORMALNY

N \ O	NR	PR	WR	NO	PO
NR	8				
PR		7	1		
WR		2	6		
NO				8	
PO				1	7

NATURALNIE

N \ O	NR	PR	WR	NO	PO
NR	8				
PR	1	5	2		
WR		1	7		
NO				6	2
PO					8

N \ O nadano \ odebrano

typy intonacji:

NR niska rosnąca
PR pełna rosnąca
WR wysoka rosnąca
NO niska opadająca
PO pełna opadająca
R równa
NRO niska rosnąco-opadająca
PRO pełna rosnąco-opadająca

Tabela 4. Wyniki rozpoznawania przebiegów intonacyjnych dla realizacji aproksymacji liniowo-schodkowej (P2).

Liczba błędów popełnionych przez poszczególnych respondentów wahała się w granicach od 19 (ok. 20 %) do 36 (ok. 38 % odpowiedzi). Nie zaobserwowano przy tym różnic, które mogłyby mieć podłoże zawodowe: poprawność identyfikacji była w przypadku muzyków i fonetyków bardzo zbliżona.

Należy tu zaznaczyć, że pojęciem "błędu" czy "błędnej identyfikacji" można się posługiwać w odniesieniu do uzyskanych wyników jedynie w sposób umowny. Nie jest wykluczone, że część odpowiedzi określonych w powyższym opisie mianem "niepoprawnych" była w istocie efektem błędów popełnionych przez respondentów przy identyfikacji poszczególnych przebiegów. Podejście takie wydaje się jednakże uzasadnione tylko w tych przypadkach, gdy dany przebieg został rozpoznany jako inny przez jedną lub - najwyżej - dwie osoby.

3.2.2. Grupa B.

Dla celów globalnej syntezy mowy nieodzowne jest opracowanie "modelowych" przebiegów intonacyjnych dla różnego typu wypowiedzi, przede wszystkim dla wypowiedzi o charakterze pytania (o rozstrzygnięcie) oraz stwierdzenia. Eksperyment przeprowadzony z grupą B miał na celu ustalenie, które ze zsyntetyzowanych intonacji są najbardziej typowe dla pytań, które zaś - dla oznajmień. Uzyskane wyniki przedstawione są w Tabeli 5.

Dwie z trzech zsyntetyzowanych intonacji rosnących (PR i WR) okazały się być jednakowo efektywnym sygnałem pytania: zostały one uznane przez respondentów za pytające w ponad 98 % przypadków.

Z wyjątkiem przebiegu NR, przy klasyfikacji którego wystąpiła największa rozbieżność ocen, wszystkie pozostałe intonacje zostały powszechnie uznane za typowe dla oznajmienia (od 88 % do 100 % odpowiedzi). Należy przy tym zauważyć, że dotyczyło to nie tylko przebiegów zakończonych spadkiem F_0, ale również przebiegu równego.

3.3. Omówienie wyników.

Rezultaty przeprowadzonych badań odsłuchowych stanowią podstawę zarówno do zobiektywizowanej oceny syntetycznych przebiegów intonacyjnych oraz weryfikacji przyjętych założeń teoretycznych, jak też do wysnucia szeregu wniosków istotnych

Typ intonacji	Liczba ocen jako :			Razem wystąpień
	„pytanie"	„odpowiedź"	brak oceny	
niska rosnąca (NR)	58	91	11	160
pełna rosnąca (PR)	157	3	—	160
wysoka rosnąca (WR)	157	3	—	160
niska opadająca (NO)	2	158	—	160
pełna opadająca (PO)	3	157	—	160
równa (R)	1	79	—	80
niska rosnąco-opadająca (NRO)	—	40	—	40
pełna rosnąco-opadająca (PRO)	5	35	—	40

Tabela 5. Wyniki klasyfikacji przebiegów intonacyjnych w terminach „pytanie - odpowiedź".

z punktu widzenia potrzeb globalnej syntezy mowy.

Na ogólną liczbę 236 błędnych rozpoznań odnotowanych w doświadczeniu z grupą A, w zdecydowanej większości przypadków (81 %) respondenci podali jako odpowiedź taki typ przebiegu intonacji, w którym kierunek zmian częstotliwości podstawowej był identyczny jak w bodźcu nadanym. Oznacza to, że zasadniczy problem przy identyfikacji stanowił na ogół zakres zmian F_0, a nie ich kierunek. Jedynie w 45 przypadkach wystąpiły substytucje drugiego rodzaju, przy czym najczęściej mylonymi przebiegami były NR i R oraz NO i R. Wynikałoby stąd, iż albo wzrost/spadek wartości F_0 w obrębie przebiegów NR i NO był zbyt mały, albo też w ich przypadku - inaczej niż dla pozostałych przebiegów - ustalenie kierunku zmiany intonacji było trudniejsze niż określenie jej zakresu. Biorąc pod uwagę fakt, że NR i NO zostały rozpoznane najlepiej ze wszystkich przebiegów (NR: 84 %, NO: 77 % poprawnych odpowiedzi), należałoby raczej uznać za słuszną drugą z podanych alternatyw. Przyczyna tego zjawiska tkwi zapewne w percepcyjnym podobieństwie intonacji niskich, charakteryzujących się niewielkim zakresem zmian F_0, i intonacji równej.

Wyniki rozpoznawania przebiegów prostych są dosyć zbliżone (od 63 % do 84 % poprawnych odpowiedzi). Zdecydowanie różnią się od nich natomiast rezultaty identyfikacji obu przebiegów złożonych (NRO: 34 %, PRO: 19 %). Przypuszczalnie ze względu na ograniczone występowanie tych intonacji w mowie polskiej (zwłaszcza w wypowiedziach jednowyrazowych) były one najczęściej odbierane jako proste intonacje opadające.

Pewien wpływ na wyniki identyfikacji przebiegów intonacyjnych miała niewątpliwie liczba sylab wyrazów, w obrębie których przebiegi te występowały. W przypadku jednosylabowego wyrazu dał i dwusylabowego wyrazu dobra poprawność rozpoznawania była nieco niższa (odpowiednio: 68 % i 56 %) niż w dwu pozostałych (normalny: 81 %, naturalnie: 77 % poprawnych odpowiedzi). Wiąże się to zapewne z faktem, że w wyrazach zawierających trzy i więcej sylab wyraźniej zaznaczony (łatwiej uchwytny percepcyjnie) jest "poziom wyjściowy" częstotliwości podstawowej, poprzedzający jej wzrost lub spadek, który stanowi

w procesie identyfikacji swoisty poziom odniesienia.

Z punktu widzenia przyszłych prac nad techniczną realizacją syntezy intonacji zdaniowej najbardziej istotny wynik doświadczenia przeprowadzonego z grupą A stanowi ustalenie, w których z przyjętych typów aproksymacji rozpoznawanie przebiegów intonacyjnych było najefektywniejsze. Jak należało oczekiwać, w przypadku najprostszej realizacyjnie wersji prostoliniowej uzyskano najgorsze rezultaty identyfikacji (63 % poprawnych odpowiedzi). Aproksymowane przy jej użyciu przebiegi cechowała ponadto pewna nienaturalność brzmienia, co zaznaczało się szczególnie jaskrawo w wyrazach normalny i naturalnie, w których akcent przesuwał się z sylaby przedostatniej na pierwszą.

Przebiegi zrealizowane w wersji quasi-naturalnej i przebiegi aproksymowane łamaną zostały poprawnie rozpoznane w zbliżonej liczbie przypadków (odpowiednio: 69 % i 71 %). Powodem, dla którego najlepsze wyniki uzyskano dla aproksymacji liniowo-schodkowej była charakterystyczna, bardziej raptowna niż w innych wersjach, zmiana wartości F_0 oraz występowanie na początku i końcu wypowiedzi dość długich (a więc łatwo uchwytnych percepcyjnie) odcinków, w obrębie których F_0 zachowywała równą wartość. Ten sposób realizacji przebiegów powodował jednakże równocześnie powstanie percepcyjnego efektu śpiewności produkowanej mowy.

Rezultaty doświadczenia przeprowadzonego z grupą B stworzyły podstawę do podziału intonacji syntetycznych na pytające i oznajmujące. Jak wynika z Tabeli 5, na typ odpowiedzi miały wpływ trzy zasadnicze czynniki: kierunek (ostatniej) zmiany F_0 w sylabie akcentowanej, zakres tej zmiany oraz poziom F_0 w części przed-/poakcentowej wyrazu. Przypadek intonacji równej wskazuje na występowanie tendencji polegającej na preferowaniu alternatywy typu "oznajmienie". Można na tej podstawie wysnuć wniosek, iż wzrost intonacji w końcowym fragmencie wypowiedzi jest warunkiem koniecznym odpowiedzi typu "pytanie". Wzrost ten musi ponad to charakteryzować dostatecznie duży zakres, bądź też dostatecznie wysoki poziom F_0 w części przedakcentowej wyrazu. Co najmniej jeden z tych warunków był spełniony w przy-

padku intonacji PR i WR, niemal jednomyślnie uznany za pytające. Zasadnicza rozbieżność ocen zaznaczyła się natomiast w odniesieniu do przebiegu NR. Co więcej, ze względu na charakteryzujący ten przebieg niski poziom początkowy oraz niewielki zakres wzrostu F_0, zdecydowana większość respondentów uznała go za typowy dla oznajmienia.

Intonacje opadające, rosnąco-opadające i równą zakwalifikowano jako charakterystyczne dla oznajmienia w ok. 98 % przypadków, przy czym największą rozbieżność odpowiedzi odnotowano dla przebiegu PRO (przy kwalifikacji tego przebiegu część respondentów uznała najwyraźniej za decydujący stosunkowo wysoki - a więc typowy dla pytania - poziom F_0 we fragmencie przedakcentowym).

3.4. Wnioski końcowe.

Uzyskane w niniejszej pracy wyniki dają podstawę do wysnucia kilku wniosków istotnych z punktu widzenia dalszych prac nad syntezą intonacji. Pozwalają one m.in. stwierdzić, iż zarówno ze względu na dość wysoką efektywność rozpoznawania, jak i dużą naturalność brzmienia przebiegów intonacyjnych optymalną wersję realizacyjną stanowi wersja wykorzystująca aproksymację łamaną. Z dwu intonacji rosnących, uznanych za reprezentatywne dla pytania, jako "modelową" należałoby raczej wybrać intonację pełną rosnącą, którą cechuje szerszy zakres występowania w mowie polskiej, i która jest w związku z tym bardziej neutralna (uniwersalna) niż wysoka rosnąca. Z podobnych powodów za "modelową" intonację oznajmującą należałoby uznać przebieg niski opadający oraz, ewentualnie, także pełny opadający, który ze względu na dość dużą częstość występowania w mowie polskiej mógłby również okazać się przydatny, nawet przy syntezie stosunkowo prostych wypowiedzi.

Optymalne dla głosu męskiego zakresy częstotliwości podstawowej przy syntezie za pomocą układu CT-1 kształtują się dla trzech wymienionych przebiegów w granicach:

PR: 100 - 240 Hz
NO: 150 - 90 Hz
PO: 240 - 100 Hz

DODATEK

WYKRESY PRZEBIEGÓW INTONACYJNYCH DLA SYNTETYCZNYCH REALIZACJI WYRAZÓW : DOBRA, DAŁ, NORMALNY I NATURALNIE.

Każdy z wykresów dotyczy przebiegu jednej intonacji (dla danego wyrazu) i przedstawia cztery jej realizacje:

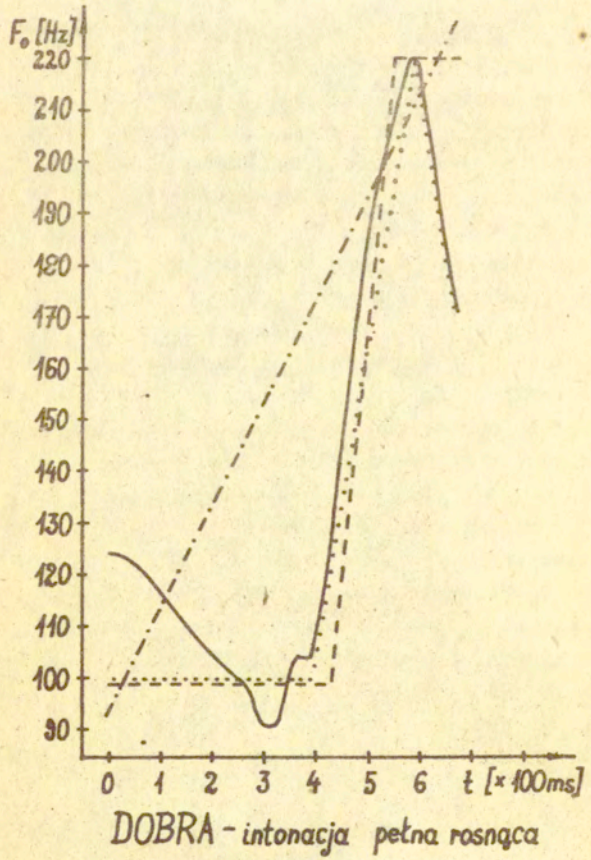
-) quasi-naturalną oznaczoną linią ciągłą ———

oraz trzy oparte na aproksymacjach:

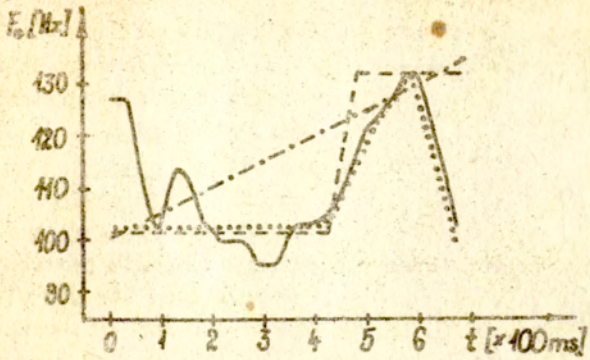
-) łamaną oznaczoną linią przerywaną – – – –

-) prostoliniowej oznaczoną linią ———

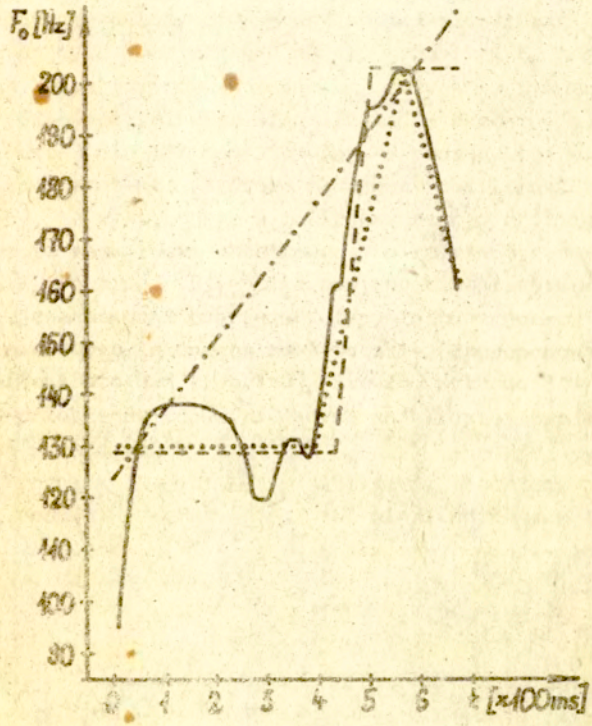
-) liniowo-schodkowej oznaczoną linią kropkowaną .....



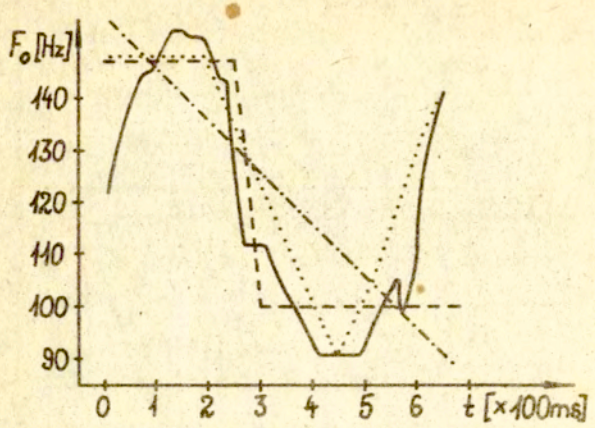
DOBRA - intonacja pełna rosnąca



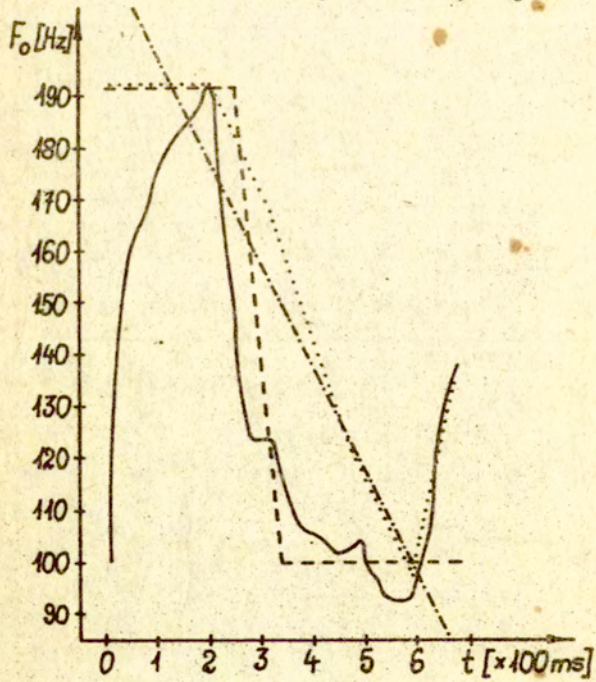
DOBRA - intonacja niska rosnąca



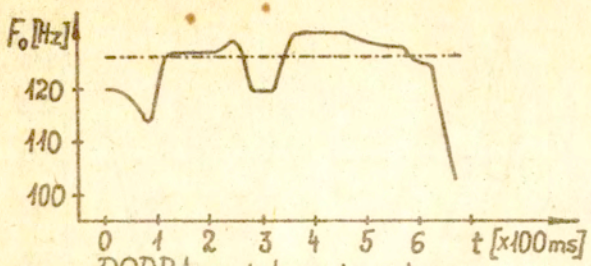
DOBRA - intonacja wysoka rosnąca



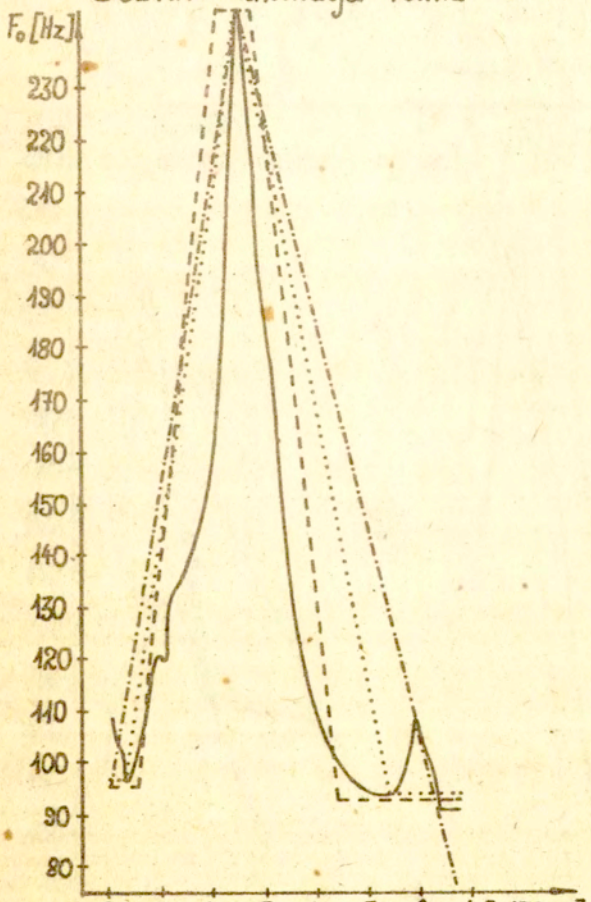
DOBRA – intonacja niska opadająca



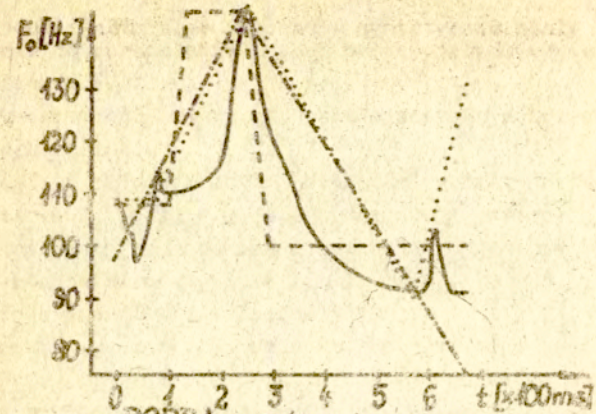
DOBRA – intonacja pełna opadająca



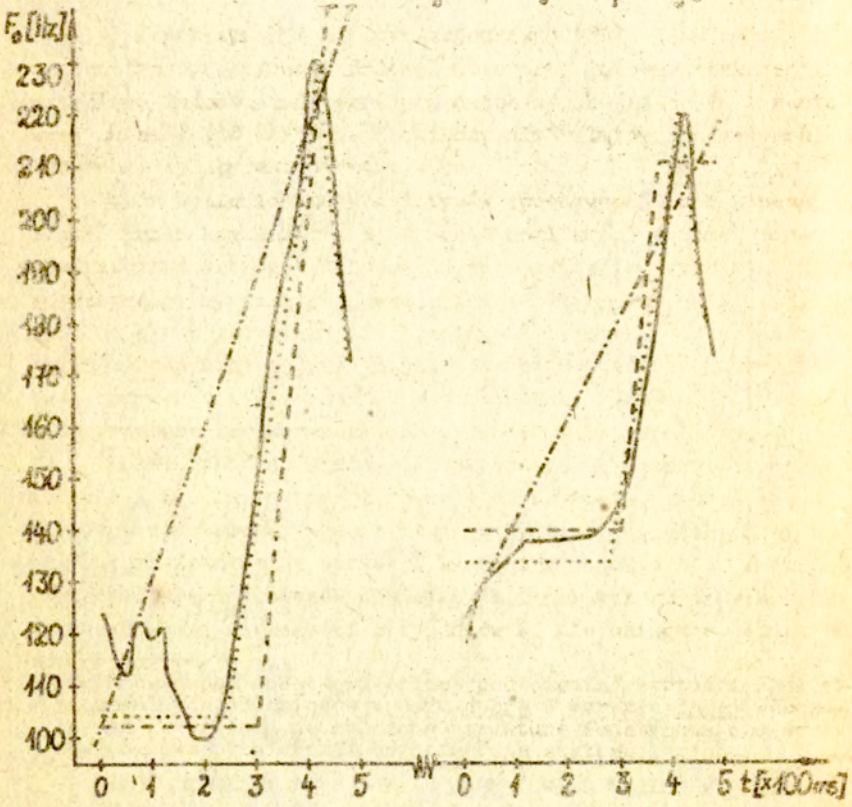
DOBRA - intonacja równa



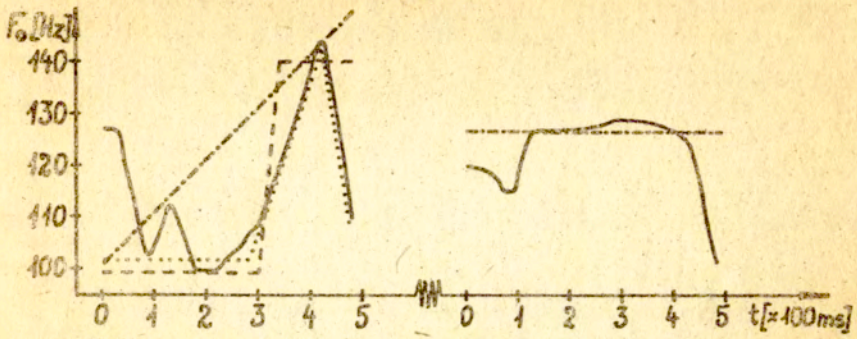
DOBRA - intonacja rosnąco-opadająca pełna



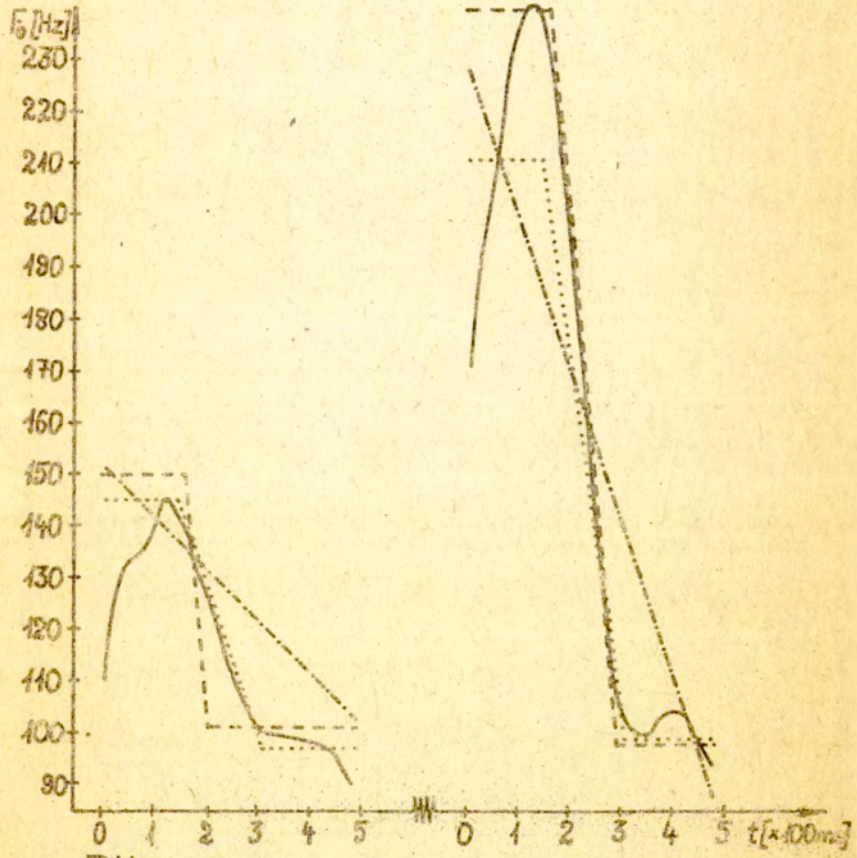
DOBRA - intonacja rosnąco-opadająca niska



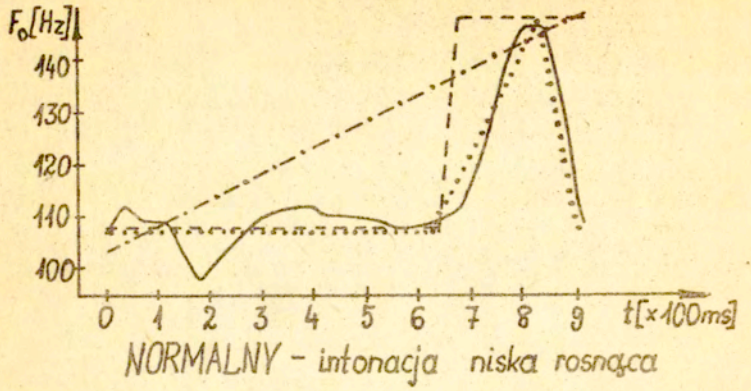
DAŁ - intonacje: -)pełna rosnąca -)wysoka rosnąca



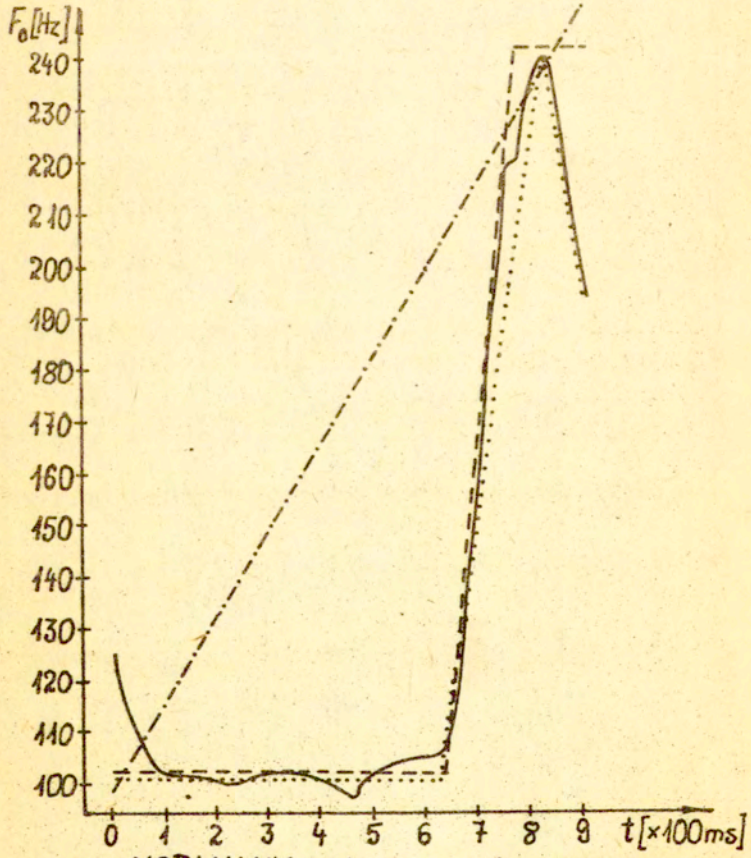
DAŁ - intonacje: →niska rosnąca →równa



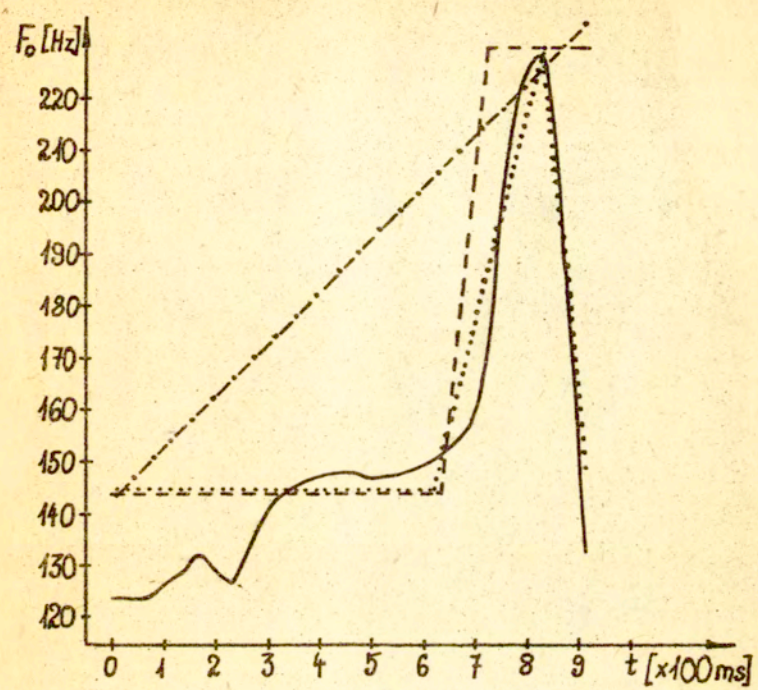
DAŁ - intonacje: →niska opadająca →pełna opadająca



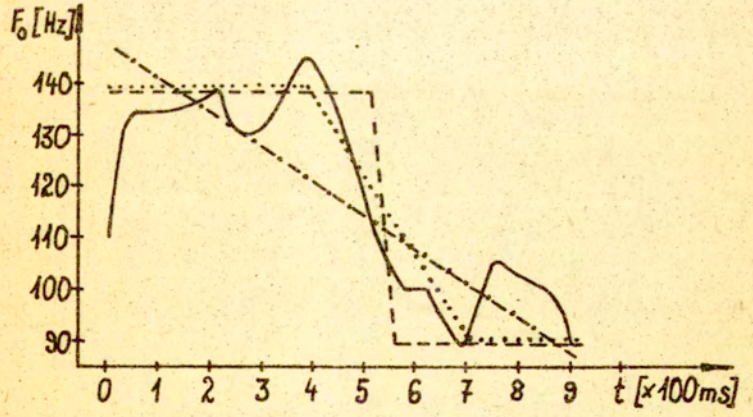
NORMALNY - intonacja niska rosnąca



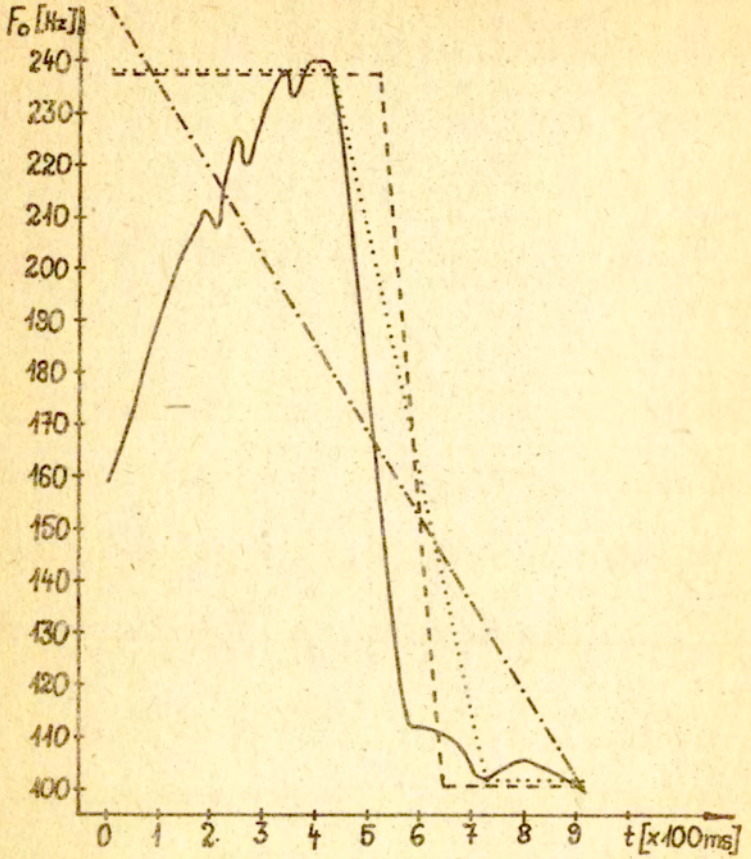
NORMALNY - intonacja pełna rosnąca



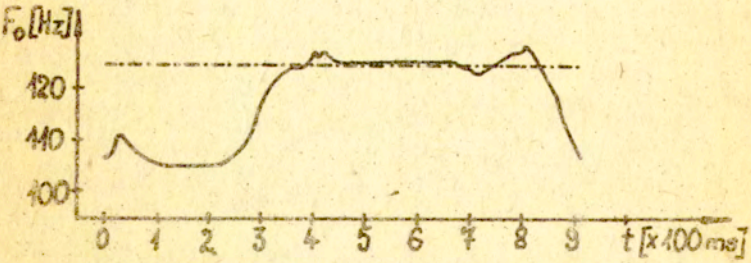
NORMALNY - intonacja wysoka rosnąca



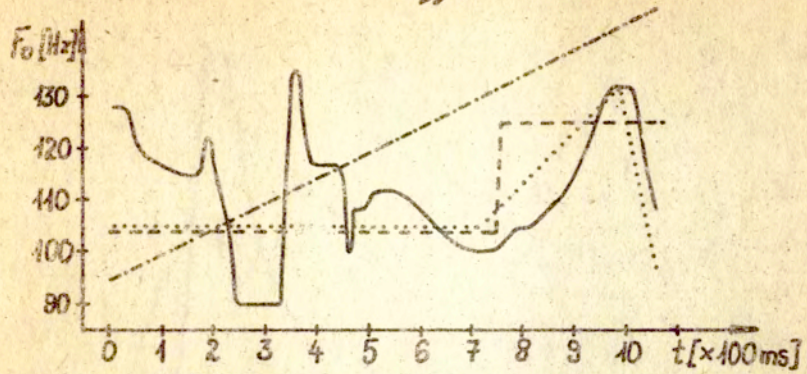
NORMALNY - intonacja niska opadająca



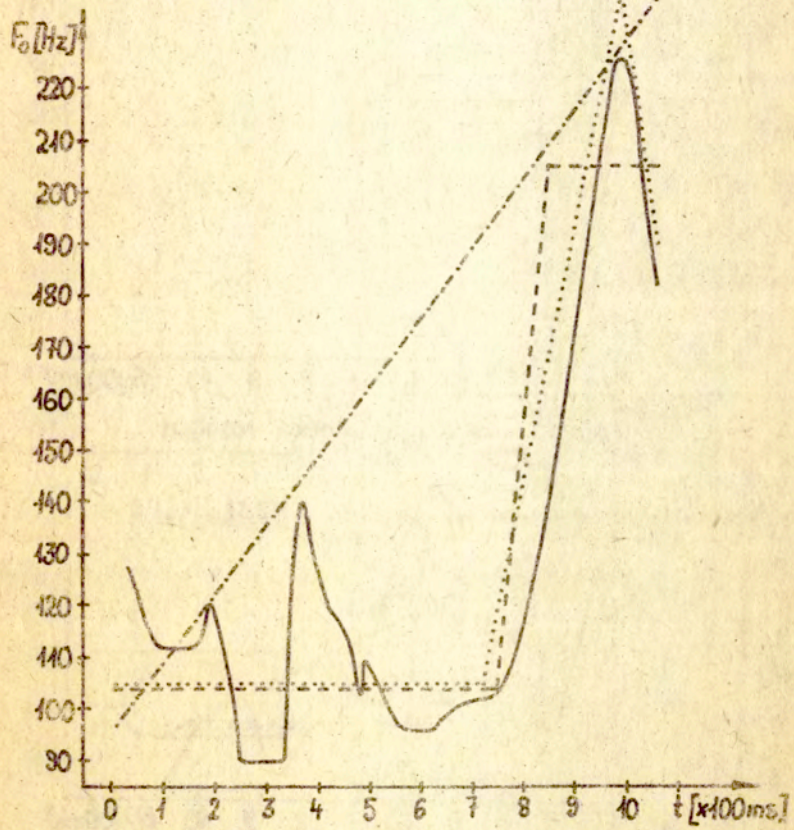
NORMALNY - intonacja pełna opadająca



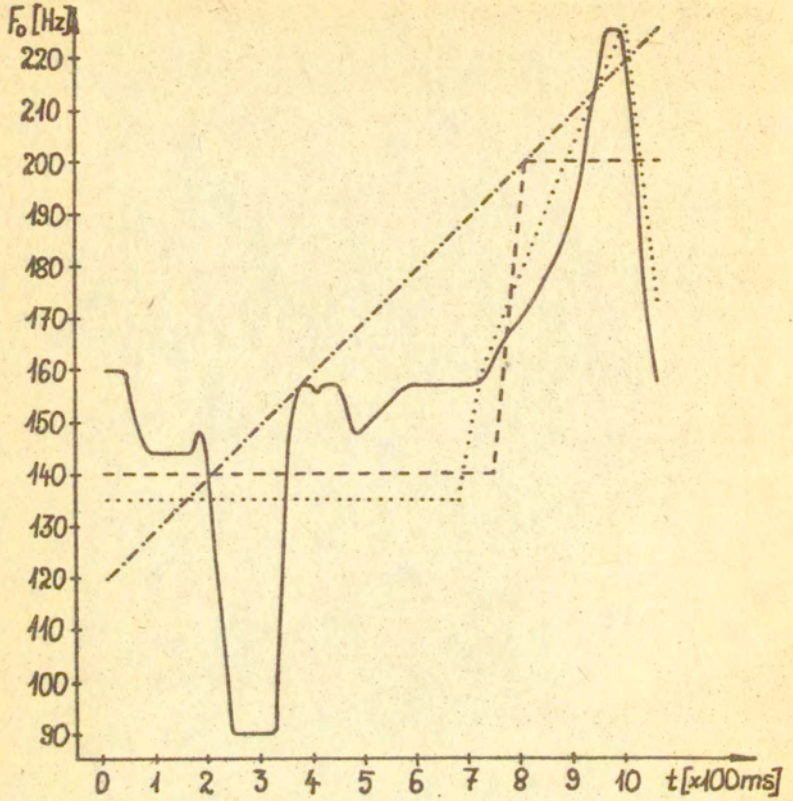
NORMALNY - intonacja równa



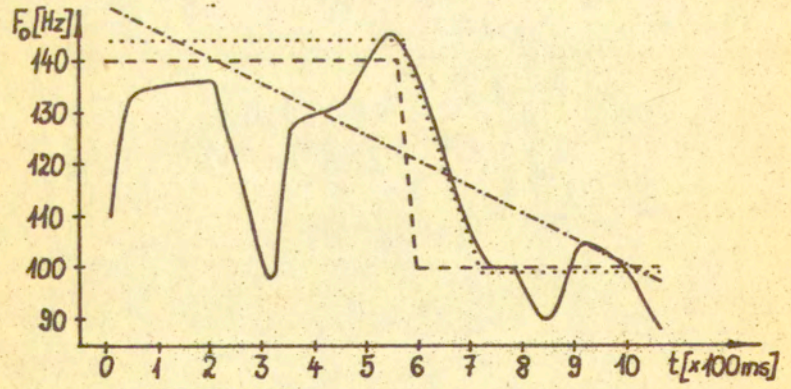
NATURALNIE – intonacja niska rosnąca



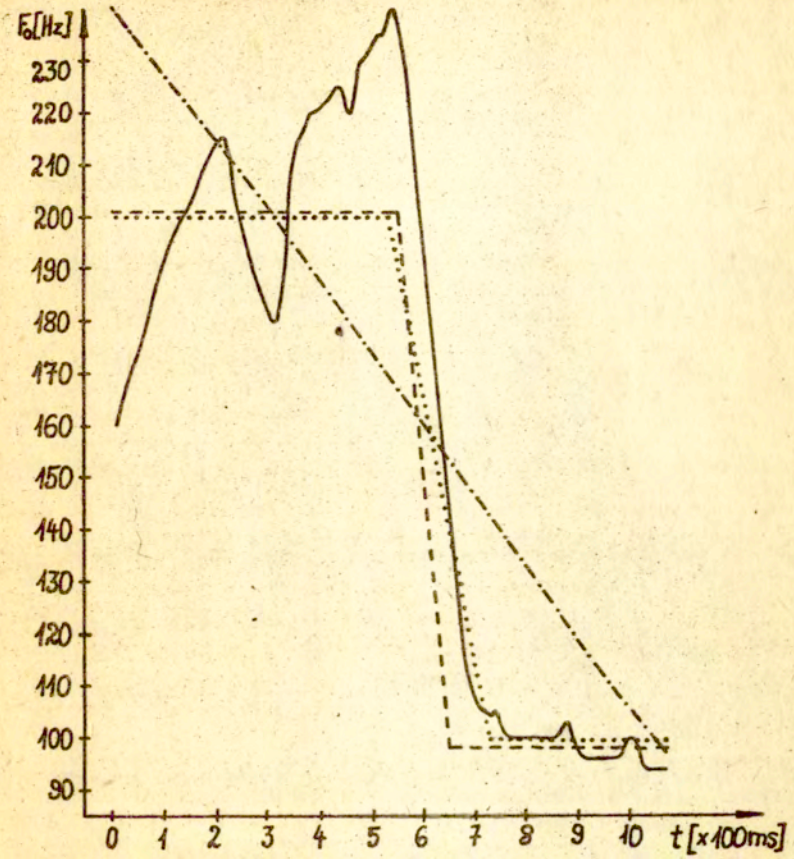
NATURALNIE – intonacja pełna rosnąca



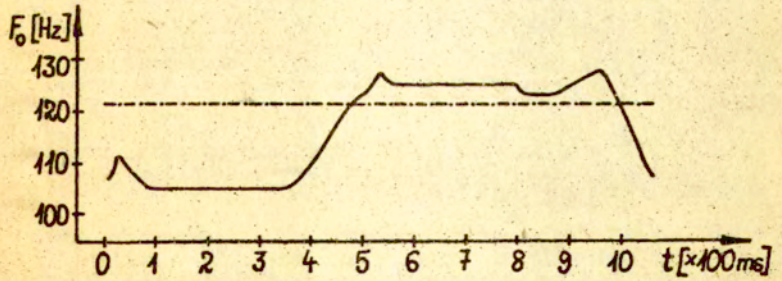
NATURALNIE - intonacja wysoka rosnąca



NATURALNIE - intonacja niska opadająca



NATURALNIE - intonacja pełna opadająca



NATURALNIE - intonacja równa

BIBLIOGRAFIA

[1] ABRAMSON, A.S., Static and dynamic acoustic cues in distinctive tones, Language and Speech 21, 1978, 319-25.

[2] ADAMS, C., MUNRO, R.R., In search of the acoustic correlates of stress: Fundamental frequency, amplitude and duration in the connected utterance of some native and non-native speakers of English, Phonetica 35, 1978, 125-56.

[3] CIARKOWSKI, R., Sterowana z minikomputera MERA 303 synteza wybranych diad polskich i ich percepcja, Prace IPPT 7/1984, Warszawa 1984.

[4] CIARKOWSKI, R., IMIOŁCZYK, J., Synteza wybranych wyrazów polskich i ich ocena percepcyjna, Prace IPPT 1/1985, Warszawa 1985.

[5] CIARKOWSKI, R., IMIOŁCZYK, J., Analysis-aided formant speech synthesis, MELECON '85, vol. II, 171-3, North Holland, 1985.

[6] CIARKOWSKI, R., IMIOŁCZYK, J., Synteza mowy polskiej: diady i wyrazy ze spółgłoskami zwartymi, Prace IPPT 42/1985, Warszawa 1985.

[7] CLARK, J.E., A low-level speech synthesis by rule system, Journal of Phonetics 9, 1981, 451-76.

[8] DEMENKO, G., Statystyczne własności rozkładów chwilowych wartości parametru F_0 w mowie ciągłej, Prace IPPT 31/1980, Warszawa 1980.

[9] DENES, P., MILTON-WILLIAMS, J., Further studies in intonation, Language and Speech 5, 1962, 1-14.

[10] DUKIEWICZ, L., Intonacja wypowiedzi polskich, Ossolineum, Wrocław 1978.

[11] FRY, D.B., Duration and intensity as physical correlates of linguistic stress, JASA 27, 1955, 765-8.

[12] HADDING-KOCH, K., STUDDERT-KENNEDY, M., An experimental study of some intonation contours, Phonetica 11, 1964, 175-85.

[13] HIRST, D., Phonological implications of a production model of intonation, Akten der Vierten Internationalen Phonologie-Tagung, Wien 1980, 195-201.

[14] JASSEM, W., MORTON, J., STEFFEN-BATOG, M., The perception of stress in synthetic speech-like stimuli by Polish listeners, w: Speech Analysis and Synthesis (ed.W.Jassem), vol. 1, 289-308, Warsaw 1968.

[15] LEHISTE, I., Suprasegmentals, The M.I.T. Press, Cambridge, Massachusetts, 1970.

[16] LIEBERMAN, Ph., Direct comparison of subglottal and esophageal pressure during speech, JASA 43, 1968, 1157-64.

[17] MAJEWSKI, W., BLASDELL, R., Influence of fundamental frequency cues on the perception of some synthetic intonation contours, JASA 45, 1969, 450-7.

[18] MATTINGLY, I.G., Synthesis by rule of prosodic features, Language and Speech 9, 1966, 1-13.

[19] ÖHMAN, S., Word and sentence intonation: A quantative model, Speech Transmission Laboratory QPSR 2-3, 1967, 20-54.

[20] ÖHMAN, S., LINDQVIST, J., Analysis-by-synthesis of prosodic pitch contours, Speech Transmission Laboratory QPSR 4, 1965, 1-6.

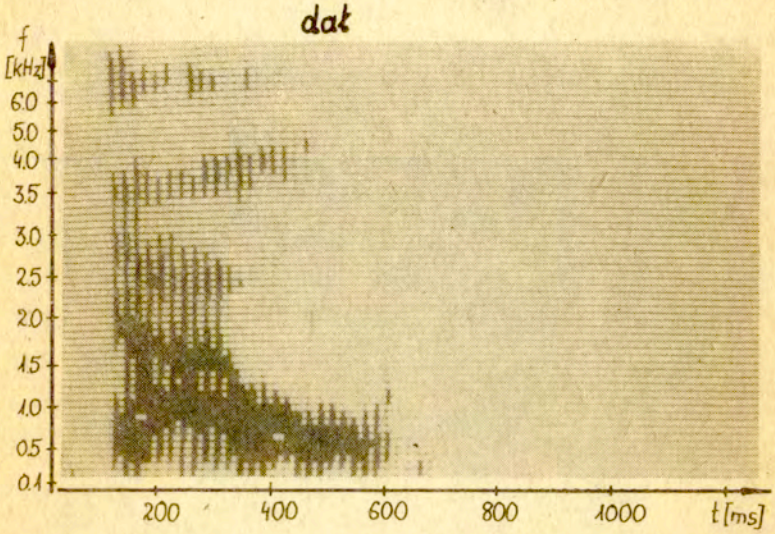
[21] OLIVE, J.P., Fundamental frequency rules for the synthesis of simple declarative English sentences, JASA 57, 1975, 476-82.

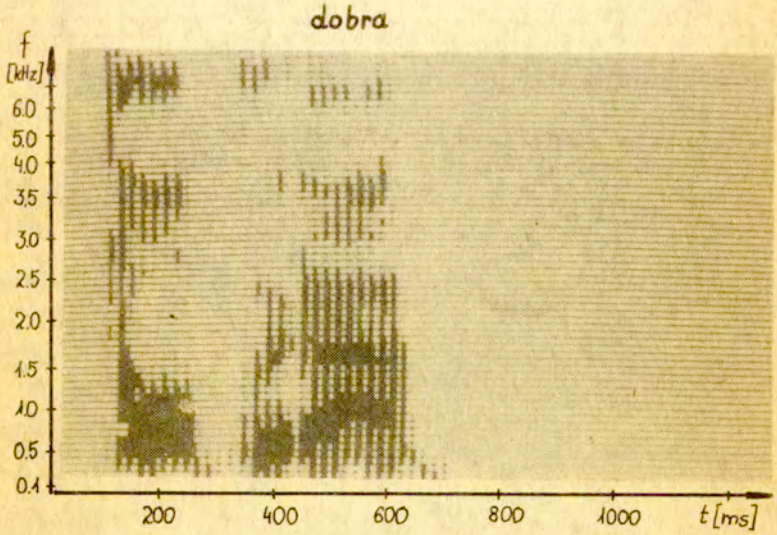
[22] RICHTER, L., Analiza statystyczna rytmicznej struktury wypowiedzi w mowie polskiej, Prace IPPT 8/1984, Warszawa 1984.

SUMMARY

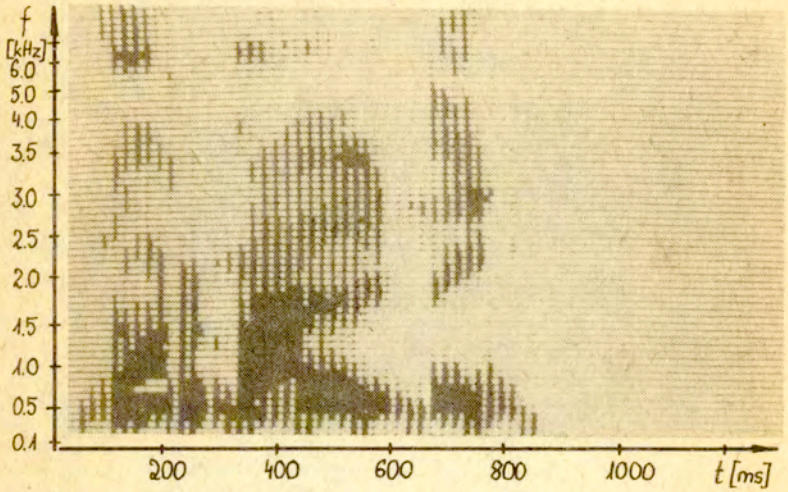
Four Polish words with varying number of syllables (*da!*, *dobra*, *normalny* and *naturalnie*) were synthesized using a COMPUTALKER CT-1 speech synthesizer. For each of the words 6 basic intonation patterns were obtained by appropriately controlling the F_0 parameter. These were: low, full and high rising intonations, low and full falling intonations and level intonation (additionally, low and full rising-falling intonations were synthesized for the word *dobra*). Four versions of each of the intonation patterns were prepared (a quasi-natural version + 3 types of approximation) of which the quasi-natural version, elaborated on the basis of F_0 values extracted from natural utterances of the four words, served as the model for the remaining three. The total of 96 synthetic intonation patterns were tested for recognizability and naturalness in a listening experiment.

SPEKTROGRAMY WYRAZÓW BĘDĄCYCH OBIEKTAMI DOŚWIADCZENIA





normalny



naturalnie