GRAŻYNA DEMENKO

ANALIZA CECH SUPRASEGMENTALNYCH JĘZYKA POLSKIEGO NA POTRZEBY TECHNOLOGII MOWY

GRAŻYNA DEMENKO

ANALIZA CECH SUPRASEGMENTALNYCH JĘZYKA POLSKIEGO NA POTRZEBY TECHNOLOGII MOWY

Spis treści

UNIWERSYTET IM. ADAMA MICKIEWICZA W POZNANIU
SERIA JĘZYKOZNAWSTWO STOSOWANE NR 17

WYDAWNICTWO NAUKOWE

POZNAŃ 1999

ABSTRACT. Demenko Grażyna, *Analiza cech suprasegmentalnych języka polskiego na potrzeby technologii mowy* (Analysis of Polish Suprasegmentals for Speech Technology) Poznań 1999. Adam Mickiewicz University Press. Seria Językoznawstwo Stosowane nr 17, pp. 232. ISBN 83-232-1002-0. ISSN 0137-1444. Polish text with a Summary in English.

The present dissertation presents problems arising in the analysis of suprasegmentals in speech, their modelling, classification, synthesis and automatic recognition. On the basis of linguistic premises related to the general theory of suprasegmentals and on empirical verification of given hypotheses at the acoustic, perceptual and structural level, a model of the Polish intonational phrase is proposed. The results of a comprehensive analysis of the tunes in Polish speech may be directly used, above all, in systems of Automatic Speech Recognition and Text-to-Speech Synthesis, which are currently carried out in Poland with increasing intensity.

Grażyna Demenko, Uniwersytet im. Adama Mickiewicza, Instytut Lingwistyki, ul. Międzychodzka 5, 60-371 Poznań, Polska - Poland.

Recenzent: *prof. zw. dr hab. inż. Ryszard Tadeusiewicz*

Projekt okładki: *Ewa Wąsowska*
Redaktor: *Renata Filipowicz*
Redaktor techniczny: *Elżbieta Rygielska*
Łamanie: perfekt, ul. Grodziska 11, 60-363 Poznań

ISBN 83-232-1002-0
ISSN 0137-1444

WYDAWNICTWO NAUKOWE UNIWERSYTETU IM. ADAMA MICKIEWICZA W POZNANIU
Wydanie I, Nakład 320 + 80 egz. Ark. wyd. 20,50. Ark. druk. 14,5
Papier offset, kl. III. Podpisano do druku w styczniu 2000 r.
POZNAŃSKA DRUKARNIA NAUKOWA Poznań, ul. Heweliusza 40

Pliki przygotowane na podstawie wydania oryginalnego. Numeracja stron i przypisów może być zmieniona w stosunku do oryginału. W przypadku odwołań i cytowań, prosimy o korzystanie z numeracji wydania oryginalnego.

Język: polski

1 CECHY SUPRASEGMENTALNE W TECHNOLOGII MOWY

W okresie obecnego, szybkiego rozwoju techniki cyfrowej oraz postępu prac w zakresie analizy i przetwarzania języka naturalnego istnieje realna szansa, że rozpoznawanie i synteza mowy będą podstawowymi środkami komunikacji w komputerowych systemach dialogowych. Problematyka łączności słownej w układach technicznych obejmuje szeroki zakres zagadnień z różnych dziedzin wiedzy związany z analizą fonetyczno-akustycznych cech mowy, jej rozpoznawaniem, syntezą oraz transmisją. Automatyczne przetwarzanie dźwiękowej postaci języka stanowi przedmiot badań wielu dyscyplin naukowych, takich jak: technologia mowy, fonetyka, lingwistyka komputerowa, psycholingwistyką, informatyka, telekomunikacja, foniatria i audiologia.

Złożoność problematyki, zarówno na etapie wytwarzania, percepcji, jak i analizy akustycznej sygnału, wynika — niezależnie od tego, czy układem rozpoznającym jest mózg człowieka czy komputer — ze specyficznych własności mowy. Mimo wielu badań prowadzonych intensywnie w przeciągu ostatnich kilkudziesięciu lat, relacje między sygnałem akustycznym i strukturą języka nie zostały w pełni ustalone. Powstałe w procesie artykulacji zespoły dźwięków są nośnikami różnorodnych informacji językowych, paralingwistycznych oraz pozajęzykowych. Przeprowadzenie kompleksowej analizy uwzględniającej oddziaływanie wielu interaktywnych źródeł zmienności wymaga obszernej bazy danych i pracochłonnych eksperymentów. Określenie tych źródeł i opisanie ich funkcjonowania jest zadaniem tak skomplikowanym, że istnieje pogląd sceptyczny, według którego sformułowanie odpowiednich algorytmów rozpoznawania oraz syntezy mowy wyłącznie na bazie teorii jest wątpliwe. W związku z tym, zauważa się w ostatnich latach rozwiązania typowo techniczne. Powstają układy rozpoznawania oraz syntezy mowy oparte głównie na statystyczno-matematycznych algorytmach (np. sieciach neuronowych, procesach Markowa) umożliwiających uczenie systemów bez konieczności uwzględniania złożonych związków między językowymi a akustycznymi cechami sygnału. Tego rodzaju opracowania nie zapewniają sformułowania uniwersalnych, poprawnie funkcjonujących algorytmów, niezależnych od doboru materiału językowego, głosu mówcy oraz akustycznych uwarunkowań otoczenia. Przygotowanie zaś reprezentatywnej bazy danych na potrzeby automatycznego uczenia systemu rozpoznawania mowy lub syntezy, możliwe w pewnym stopniu dla tekstów czytanych, wydaje się, w przypadku mowy spontanicznej problemem nie do rozwiązania. Jeśli wziąć pod uwagę fakt, że wyniki rozpoznawania mowy przy zastosowaniu różnych algorytmów często dają podobne rezultaty, można przypuszczać, że trudności w przetwarzaniu sygnału wynikają nie tyle z nieadekwatności stosowanych metod czy matematycznych modeli decyzyjnych, co z powodu nieuwzględniania w opisie mowy inwariantów w zakresie poszczególnych typów informacji.

Pomimo szybkiego tempa prac poświęconych automatycznemu przetwarzaniu języka naturalnego, praktyczne implementacje analizy, syntezy a zwłaszcza rozpoznawania mowy są ciągle ograniczone. W komputerowych systemach komunikacji słownej konieczne jest uwzględnienie informacji nie tylko segmentalnej, ale również informacji suprasegmentalnej w bardzo znacznym stopniu wykorzystywanej zarówno przez mówcę, jak i przez słuchacza. Podstawowe problemy związane z parametryzacją i modelowaniem struktur melodycznych poszczególnych języków nie są jednak zadowalająco dobrze rozwiązane dla praktycznych implementacji. Duża liczba stosowanych technik w zakresie ekstrakcji, opisu cech suprasegmentalnych mowy oraz kwantytatywnych modeli intonacji opartych na różnego rodzaju manualnych transkrypcjach struktur melodycznych świadczy o tym, że jak dotychczas nie jest opracowana odpowiednia metodologia badań w tej dziedzinie.

Podobnie jak w przypadku cech segmentalnych mowy, również w przetwarzaniu suprasegmentaliów próbuje się w najnowszych opracowaniach, wykorzystujących automatyczne uczenie, pominąć metodologiczne problemy związane z niedostateczną wiedzą o interakcji różnych informacji zawartych w sygnale. Sposób „ślepego" uczenia układu nie wymaga ani obszernych eksperymentów, ani manualnej transkrypcji złożonych struktur suprasegmentalnych. Wydaje się więc rozwiązaniem optymalnym. Jak dotąd zauważa się tylko nieliczne opinie krytyczne tego rodzaju rozwiązań (por. np. Collier 1992, s. 205).

Zasadniczą rolę odgrywają cechy suprasegmentalne w syntezie mowy. Modelowanie melodycznych struktur zwiększa zrozumiałość i w sposób decydujący wpływa na naturalność wypowiedzi. Trudno obecnie zaakceptować system dialogowy wytwarzający monotonną mowę. Problem sterowania suprasegmentaliami na potrzeby syntezy w zakresie mowy czytanej został już częściowo rozwiązany, zwłaszcza dla języka angielskiego (np. de Pijper 1983, Santen 1997a i b), japońskiego (Fujisaki 1988, 1997), holenderskiego ('t Hart et al. 1990, Terken 1993), niemieckiego (Kohler 1991, 1995, Portele et al. 1997, Portele 1997, Traber 1997), francuskiego (Hirst et al. 1991, Veronis et al. 1997).

W systemach rozpoznawania mowy prozodia nie jest niezbędna, jednak jej uwzględnienie może zwiększyć efektywność pracy systemu, skrócić czas obliczeń oraz ułatwić korekcję błędów. W językach nietonicznych, takich jak polski, angielski, niemiecki, francuski udział intonacji w przekazywaniu informacji polega na tym, że sygnalizuje ona pewne stany emocjonalne mówcy, jego stosunek do treści wypowiedzi lub do słuchacza. W językach tonicznych, jak np. szwedzki, japoński oraz tonalnych, jak np. chiński i wietnamski, intonacja spełnia funkcję podwójną. W językach tych identyczne fonematycznie wypowiedzi o różnej dystynktywnie intonacji mogą stanowić odrębne części mowy. Różnice dystynktywne w intonacji tych języków występują na tle takich samych sekwencji fonemów i mają związek ze znaczeniem leksykalnym wyrazów.

Dla praktycznych implementacji suprasegmentaliów konieczne jest rozwiązanie podstawowych problemów metodologicznych oraz technicznych w zakresie:

a. wiarygodnej ekstrakcji parametrów suprasegmentalnych — głównie częstotliwości podstawowej,

b. kwantytatywnego opisu cech suprasegmentalnych oraz modelowania intonacji,

c. automatycznej transkrypcji struktur melodycznych,

d. integracji cech suprasegmentalnych z cechami segmentalnymi.

Suprasegmentalne cechy sygnału mogą być uwzględniane na różnych poziomach analizy:

a. fonetycznym — badanie efektów koartykulacyjnych, specyficznych wartości częstotliwości podstawowej poszczególnych samogłosek,

b. składniowym — określenie granic frazy, struktury syntaktycznej wypowiedzi,

c. pozajęzykowym — wykrycie emocji lub patologii w głosie mówcy.

Systemy rozpoznawania mowy dzieli się na systemy rozpoznające krótkie wypowiedzi (z większego lub mniejszego słownika) i mowę ciągłą (teksty czytane oraz wypowiedzi spontaniczne). W rozpoznawaniu pojedynczych wypowiedzi wykorzystanie suprasegmentaliów koncentruje się głównie na ustaleniu dla danego wyrazu wzorca akcentowego oraz na pozajęzykowych aspektach — np. wykrywaniu patologii w głosie.

W systemach rozpoznawania mowy ciągłej cechy suprasegmentalne jako źródło informacji stają się bardziej istotne, ale ich ekstrakcja jest trudniejsza i może być obarczona wieloma błędami. W mowie ciągłej istotny jest wzorzec akcentowy

w obrębie frazy lub zdania (por. np. Lea 1979, Waibel 1986, Price et al. 1991, Nakai et al. 1997). Jego weryfikacja i odnalezienie najistotniejszych fragmentów wypowiedzi pozwalają na ograniczenie czasu przeszukiwania leksykonu. Para- lingwistyczne i pozajęzykowe aspekty suprasegmentaliów odgrywają w tym przypadku drugorzędną rolę, jeśli pominąć zadanie szybkiej adaptacji systemu i konieczność wstępnego opracowania sygnału, (np. mowy z chrypką) lub identyfikację głosu.

Modelowanie suprasegmentaliów uwzględnia się obecnie w każdym systemie syntezy, w rozpoznawaniu mowy obserwuje się w tym zakresie tylko sporadyczne, pilotażowe eksperymenty (np. Komatsu et al. 1986, Nöth 1991, Kompe et al. 1993, Nöth et al. 1993, Dumouchel et al. 1993, Taylor et al. 1997). Najszerzej wykorzystano integrację cech segmentalnych i suprasegmentalnych w prototypowym, automatycznym systemie tłumaczenia tekstów — Yerbmobil (Hirose et al. 1994, Mast et al. 1996, Hirose 1997, Lehning 1996a i b, Hess 1992, Hess et al. 1997, Niemann et al. 1997, 1998).

Na podstawie 242 cech opisujących cechy suprasegmentalne sylaby (względem 6 poprzedzających i 6 następujących sylab), uzyskano dla mowy spontanicznej poprawność rozpoznawania akcentu i granic frazy w zakresie 82,5-91,7%. Nie do końca jednak znany jest zakres wykorzystania suprasegmentaliów w implementacjach praktycznych (por. np. Hess et al. 1997).

Analiza cech suprasegmentalnych sygnału mowy jest przedmiotem intensywnych badań wielu zespołów naukowych na całym świecie, można się więc spodziewać, że już wkrótce odpowiednią metodologią uda się z sygnału mowy wyodrębnić większość ukrytych informacji i wykorzystać je nie tylko w technologii mowy, ale również w innych dyscyplinach nauki.

Pierwsze prace poświęcone fonetyce akustycznej, akustyce mowy oraz łączności przy pomocy języka naturalnego powstały w Polsce w latach 50. zapoczątkowane między innymi przez Skorupkę (1955), Dłuską (1957), Jassema (1949, 1952). Od tego czasu obserwuje się w zróżnicowanych kręgach specjalistów z zakresu informatyki, telekomunikacji, fonetyki, lingwistyki i medycyny szerokie zainteresowanie nową w Polsce dziedziną nauki — technologią mowy. Przekrojową problematykę z tej dziedziny przedstawiają np. prace Tadeusiewicza (1988) i Basztury (1989, 1993). W zakresie rozpoznawania mowy w latach 70. oraz 90. powstały liczne algorytmy i metody opracowane między innymi przez Gubrynowicza (1967, 1968), Kacprowskiego et al. (1970), Gubrynowicza et al. (1990), Tadeusiewicza (1994), Majewskiego (1994), Baszturę (1994), Izworskiego (1995), Kubzdelę (1986, 1997) oraz Grocholewskiego (1995b). Jakkolwiek pierwsze prace nad syntezą mowy polskiej powstały już w latach 60. i 70. (np. Kacprowski 1965, Kacprowski et al. 1968, Myślecki 1979), to dopiero w latach 90. zwraca się uwagę na konieczność starannego modelowania suprasegmentaliów (por. między innymi Imiołczyk et al. 1993, 1994, Demenko et al. 1993). Opracowywane ostatnio bazy danych np. przez Grocholewskiego (1995a, 1997), Gubrynowicza (1998) stwarzają możliwości włączenia języka polskiego do europejskich komputerowych systemów dialogowych. Powstają obecnie również w Polsce prace poświęcone automatycznemu tłumaczeniu tekstów (Jassem 1996b, Jassem 1997).

Do tej pory dla języka polskiego brak kompleksowych badań struktur suprasegmentalnych, zarówno na potrzeby syntezy, jak i rozpoznawania mowy. Najobszerniejsze monografie dotyczące przebiegu melodii w obrębie wypowiedzi opracowane przez Steffen-Batogową (1963, 1966, 1996), Dłuską (1976), Jassema (1962) oraz Dukiewicz (1978) skoncentrowane są głównie na aspektach lingwistycznych analizy intonacji. Nieliczne badania poświęcono akustycznej strukturze suprasegmentaliów (np. Renowski 1967a, b i c, Majewski et al. 1969, 1973). Rozpoczęty dopiero pod koniec lat 80. cykl prac poświęcony automatycznej analizie intonacji języka polskiego jest wciąż rozwijany (Demenko et al. 1988, Demenko 1984, 1986, 1987, 1995c, 1998).

Niniejsze opracowanie dotyczy zagadnień związanych z analizą akustyczną cech suprasegmentalnych mowy (głównie intonacji), z ich modelowaniem, klasyfikacją oraz automatycznym rozpoznawaniem. Lingwistyczne podstawy analiz

cech suprasegmentalnych mowy zawarte w rozdziale 2 mają na celu wyjaśnienie najważniejszych pojęć z dziedziny przedmiotu stosowanych w dalszej części opracowania.

W rozdziale 3 przedstawiono główne aspekty fizjologicznych uwarunkowań oraz słuchowej percepcji wybranych, fizycznych parametrów sygnału mowy. W dziedzinie tej istnieje stosunkowo dużo szczegółowych, kompetentnych opracowań, dlatego też ograniczono się tylko do zarysowania problematyki.

Złożoność, wieloaspektowość badań nad modelowaniem cech suprasegmentalnych, szczególnie trudności w zobiektywizowaniu ich opisu wymagają analizy porównawczej z badaniami dla innych języków. Z uwagi na brak w języku polskim publikacji ujmującej syntetycznie obecny stan wiedzy w zakresie suprasegmentaliów, część niniejszego opracowania (rozdziały 4 i 5) poświęcono przedstawieniu uniwersalnej problematyki suprasegmentaliów mowy dla różnych języków oraz dokonaniu oceny stanu badań związanych z tematyką pracy prowadzonej na świecie. Rozdział 4 zawiera akustyczno-percepcyjne podstawy opisu cech suprasegmentalnych mowy oraz wyznaczników akcentu i granicy frazy. W rozdziale 5 przedstawiono oparte na kryteriach fonetyczno-akustycznych najczęściej wykorzystywane opisy oraz modele intonacji opracowane na świecie.

Rozdział 6 dotyczy definicji akcentu oraz frazy suprasegmentalnej. Sformułowano hipotezy w zakresie modelowania struktury intonacji dla języka polskiego na poziomie lingwistycznym.

W dalszej części pracy (rozdziały 7, 8 oraz 9) przedstawiono analizę akustyczną cech suprasegmentalnych języka polskiego w wypowiedziach dialogowych. Rozdział 10 poświęcono akustycznej analizie suprasegmentaliów w mowie ciągłej. Zweryfikowano hipotezy postawione w rozdziale 6 na poziomie fizycznym, percepcyjnym i strukturalnym.

Problematykę związaną z ekstrakcją cech suprasegmentalnych, parametryzacją konturu intonacyjnego przedstawiono schematycznie w rozdziale 11. Zagadnienia analizy instrumentalnej częstotliwości podstawowej omówione są w sposób wyczerpujący w obszernym opracowaniu Hessa (1983).

Automatyczne przetwarzanie cech suprasegmentalnych mowy jest zagadnieniem nowym, szczególnie dla języka polskiego. Rozdział 12 poświęcono omówieniu założeń do automatycznej analizy akcentu według ustalonego wcześniej, na podstawie przesłanek lingwistyczno-akustycznych, modelu frazy intonacyjnej. Opracowano i zweryfikowano eksperymentalnie strukturalną parametryzację zmian wysokości tonu. Schemat klasyfikacji akcentów rdzennych i pobocznych ustalono z wykorzystaniem klasycznej sieci neuronowej typu MLP (Multilayer Perceptron).

W rozdziale 13 przedstawiono projekt sieci neuronowej do klasyfikacji 12 struktur intonacyjnych w wypowiedziach dialogowych i 6 struktur intonacyjnych w tekstach czytanych. Przetestowano kilka różnych typów sieci: probabilistyczne, z funkcjami radialnymi oraz klasyczne typu MLP. Zweryfikowanie sieci na danych nie pochodzących ze zbioru uczącego (wypowiedziach z tekstów czytanych) wykazało poprawne uogólnianie nowych przypadków. Dla mowy ciągłej przeanalizowano kilka możliwości parametryzacji struktur suprasegmentalnych. Względnie wysoki procent prawidłowej, zgodnej z oczekiwaną, klasyfikacji (od 70-90% zależnie od typu akcentu) stanowi potwierdzenie dla przyjętego i analizowanego w poprzednich rozdziałach modelu intonacji dla języka polskiego.

Podstawowe zasady sterowania intonacją, przetestowane praktycznie w układzie syntezy, przedstawiono w rozdziale 14. Wskazano możliwości modyfikacji i implementacji reguł modelowania akcentu rdzennego oraz akcentów pobocznych.

Kierunki analizy suprasegmentaliów w różnych dziedzinach nauki omówiono w podstawowym zarysie w rozdziale 15. Wszechstronną ocenę aktualnego stanu wiedzy i dalszych kierunków rozwojowych analiz suprasegmentaliów zawiera publikacja (Sagisaka et al. 1997) przygotowana na podstawie materiałów z konfe-

rencji *Computional Approaches to Processing the Prosody of Spontaneous Speech* w roku 1995 w Kyoto.

W niniejszym opracowaniu z bardzo obszernej literatury przedmiotu szczególny nacisk położono na te prace, które przyczyniły się w sposób bezpośredni lub pośredni do powstania nowego kierunku badawczego w obrębie technologii mowy związanego z modelowaniem cech suprasegmentalnych.

Do analizy instrumentalnej wykorzystano najnowszą wyspecjalizowaną aparaturę, spektrograf cyfrowy Kay 5500 oraz komputer PC z procesorem Pentium II. Do modelowania intonacji wykorzystano w początkowej fazie eksperymentów pakiet programowy w oryginalnej wersji opracowany przez J. L. McClellanda oraz D. E. Rumelharta, opisany w pracy *Explorations in parallel distributed processing* (1987)[1].

Algorytmy te wraz z teoretycznymi i praktycznymi wskazówkami w opisie umożliwiły przetestowanie i wdrożenie techniki sieci neuronowych do analizy intonacji. W dalszej części pracy posłużono się pakietem Statistica zawierającym oprogramowanie przygotowane w roku 1998 przez firmę Statsoft. Program ten, chociaż mało elastyczny w porównaniu z profesjonalnym opracowaniem sieci neuronowych, łącznie z modułem do analizy danych (statystyka podstawowa, analiza wariancji, analiza dyskryminacyjna) stanowi dogodne narzędzie do podstawowych badań. Dla praktycznych implementacji zaprojektowanych modeli sieci (zwłaszcza dla rozpoznawania mowy) konieczne będzie wykorzystanie oprogramowania profesjonalnego, dającego możliwość ingerencji projektanta w strukturę sieci.

Praca powstała w ramach grantu Cooperative Research in Information Technology CRIT2 EP-20288 *Computer Analysis and Synthesis of Suprasegmental Structures in Dialogue Systems* oraz projektu badawczego finansowanego przez KBN (8T11E 04215) i jest ukierunkowana na praktyczne zastosowania w systemach syntezy i rozpoznawania mowy ostatnio intensywnie rozwijanych również w Polsce.

[1] Program udostępniony został przez prof. zw. dr hab. inż. Ryszarda Tadeusiewicza z Katedry Automatyki Wydziału Elektrotechniki, Automatyki, Informatyki i Elektroniki Akademii Górniczo-Hutniczej w Krakowie.

2 LINGWISTYCZNE PODSTAWY ANALIZY SUPRASEGMENTALIÓW

Niniejsza praca poświęcona jest analizie akustycznej określonych cech sygnału mowy. Ze stanowiska ogólnoakustycznego każdy dźwięk posiada następujące cechy: wysokość, natężenie, barwę oraz czas trwania. Te cechy mają charakter ogólny i dotyczą zarówno sygnału mowy, jak i np. sygnału akustycznego wytwarzanego w muzyce i śpiewie, a także innych zjawisk dźwiękowych, tak o charakterze informacyjnym, jak i o charakterze zakłócenia. W mowie każda z tych czterech podstawowych cech wykorzystywana jest wielostronnie, tak w płaszczyźnie językowej, jak i paralingwistycznej oraz pozajęzykowej. Na przykład dana różnica wysokości może sygnalizować określony element systemu: w języku angielskim spadek wysokości tonu ze średniego do niskiego ma inną funkcję niż spadek od tonu wysokiego do średniego. Mamy tu do czynienia z językowo dystynktywnymi zjawiskami. Dokładnie (z punktu widzenia fizycznego, tj. przebiegu parametru F0) taka sama różnica może jednak mieć charakter międzyosobniczy (np. głos męski i kobiecy), i w tym przypadku różnica jest pozajęzykowa. Określone dwa różne dźwięki samogłoskowe mogą sygnalizować różnicę fonematyczną, ale w szczególnych warunkach ta sama różnica może sygnalizować różne głosy (jak w słynnym eksperymencie Ladefogeda i Broadbenta 1957).[2] Dana samogłoska o dwóch różnych iloczasach może w jednym języku sygnalizować różnicę ściśle językową, np. w języku czeskim i fińskim, a także częściowo np. w niemieckim, gdzie różnica iloczasowa jest fonematyczna, natomiast analogiczna różnica określonej samogłoski w języku polskim ma funkcję ekspresywną, a zatem paralingwistyczną.

Każdy z wymienionych czterech aspektów dźwiękowych (akustycznych) sygnału mowy można analizować i opisywać na trzech poziomach:

- a. fizycznym (sygnał),
  b. percepcyjnym (audytywnym) oraz
  c. strukturalnym.

Na przykład w zakresie badań segmentalnych, na poziomie (a) można badać formanty samogłoskowe, mierząc ich częstotliwość i ewentualnie szerokość wstęg. Na poziomie (b) można badać, które sygnały samogłoskowe są rozróżnialne słuchowo, które są realizacjami tych samych fonemów, ale przez różne głosy. Na poziomie (c) można badać bądź metodami indukcyjnymi, bądź dedukcyjnymi, ile dany język ma fonemów samogłoskowych i jakimi cechami się one między sobą różnią.

Powyższy przykład służył do ukazania analogii. Tutaj bowiem przedmiotem badań są cechy suprasegmentalne, głównie jeden szczególny parametr, mianowicie częstotliwość podstawowa. Na poziomie (a) dokonuje się ekstrakcji parametru F0 jednym z wielu środków technicznych i według różnych założeń teoretycznych. W wyniku analizy otrzymuje się określoną funkcję czasową, ciągłą na określonych odcinkach czasowych. Wciąż na poziomie (a) można przeprowadzać różnego rodzaju analizy, porównania i zestawienia. Nie czyni się tego bezładnie. Zawsze jak w każdym przedsięwzięciu naukowym badaniom przyświeca jakaś hipoteza. Można taką hipotezę testować na przykład metodami statystycznymi, wykorzystując dane z analizy przebiegu F0. Można poszukiwać różnych regularności, niekoniecznie związanych z jakąś hipotezą o charakterze lingwistycznym. Określone przebiegi, różniące się wyłącznie albo przede wszystkim, parametrem F0 mogą jednak być też przedmiotem analizy słuchowej.

[2] Eksperyme nt badał wpływ kontekstu poprzedzającego na barwę samogłoski w wyrazie testowym w zależności od głosu mówcy.

Jeśli ocenie podlegać będą, z jakiegokolwiek punktu widzenia, wrażenia słuchowe osób poddanych eksperymentowi, to nasze badania znajdują się na poziomie (b).

Badania takie mogą (choć nie muszą) prowadzić do stwierdzeń, jaka jest w danym języku struktura zjawisk suprasegmentalnych, w szczególności wysokości tonu i jej zmian. Na przykład, czy z lingwistycznego punktu widzenia inną funkcję spełnia intonacja rosnąca od niskiego tonu do wysokiego w porównaniu z intonacją rosnącą od tonu średniego do wysokiego. Jedna z nich może sygnalizować pytanie, a druga nie. W tym przypadku rozpatruje się problemy strukturalno-lingwistyczne na poziomie (c).

Analizę suprasegmentalną można przeprowadzać metodami racjonalistyczno-dedukcyjnymi, stawiając określone tezy na podstawie przesłanek wyższego rzędu (na przykład metajęzykowych) i próbować w sygnale mowy odnaleźć potwierdzenie albo zaprzeczenie postawionej tezy. Taka metoda racjonalistyczno-dedukcyjna stosowana jest w analizie suprasegmentaliów w ramach fonologii suprasegmentalno-metrycznej. Poniżej zastosowano metodę indukcyjno-empiryczną. Niniejsza praca dotyczy zjawisk akustycznych w mowie polskiej w zakresie wszystkich 4 wymienionych cech, chociaż zjawiska barwy mają tutaj znaczenie peryferyjne.

Z lingwistycznego punktu widzenia elementy językowe występujące w sygnale mowy rozpatruje się w dwóch płaszczyznach analizy: segmentalnej i suprasegmentalnej, przy czym zjawiska podlegające analizie suprasegmentalnej są również określane synonimicznie jako prozodyczne. Różnica pomiędzy analizą segmentalną a suprasegmentalną dotyczy dziedziny (w sensie logicznym i eksperymentalnym). Dziedziną analizy segmentalnej są elementy sygnału mowy, które stanowią w płaszczyźnie percepcyjnej elementy dalej nierozkładalne na osi czasu pod względem barwy, a także określone, krótkie ciągi (sekwencje czasowe) takich elementów. Elementy te nazywa się segmentami fonetycznymi. Rozciągłość czasowa pojedynczego segmentu często pokrywa się z rozciągłością czasową głoski. Są jednak w każdym języku głoski polisegmentalne, do których należą np. dyftongi, afrykaty, a także wibranty. Dziedziną analizy suprasegmentalnej jest co najmniej pojedyncza sylaba. Najczęściej jednak jest nią określony ciąg sylab. W tym miejscu napotykamy na pierwszą niejednoznaczność. Przede wszystkim należy rozróżnić sylabę fonetyczną od fonologicznej. Definicja sylaby odnosi się do jednego z najbardziej kontrowersyjnych pojęć w lingwistyce (por. np. Awedyk 1990, Dukiewicz 1995a i b, Dukiewicz et al. 1995), istnieje nawet teoria (mniej rozpowszechniona), która odmawia sylabie statusu lingwistycznego (por. np. Dziubalska-Kołaczyk 1995). Zróżnicowania definicyjne sylaby z punktu widzenia lingwistycznego komplikują nieco analizę akustyczną, ale na ogół w tej analizie nie mają znaczenia decydującego. Ciągami sylab stanowiącymi czasowe elementy suprasegmentalne są zestroje akcentowe, jednostki rytmiczne, wzorce intonacyjne oraz frazy intonacyjne. Pojedynczy zestrój akcentowy tworzy na przykład (z materiałów użytych w niniejszej pracy) część wypowiedzi sygnalizowaną w piśmie jako *poradnia,* stanowiącą jeden zestrój akcentowy, w przeciwieństwie do części wypowiedzi pisanej *pora dnia,* która stanowi dwa zestroje akcentowe. Spacjowanie w językach europejskich niekoniecznie zgodne jest z podziałem wypowiedzi na zestroje akcentowe, jako że pojedynczym zestrojem jest także ciąg wyrazów, w którym początkowy jest proklityką, a końcowy enklityką.

Tak na przykład wypowiedź: *To nie jest najlepsza poradnia,* składa się z trzech zestrojów akcentowych: /to’ɲejest//naj’lepʃa//po’radɲa/. Każdy z wymienionych kolejnych zestrojów akcentowych reprezentuje tego samego typu jednostkę rytmiczną, którą można oznaczyć np. [ˊ _]. W polskim języku (podobnie jak w angielskim) zestrój akcentowy ma tę samą rozciągłość czasową co jednostka rytmiczna, której podstawą jest izochronizm (zob. Jassem et al. 1981, 1984). Rozróżnienie jest jednak potrzebne, gdyż w innych językach może nie zachodzić taka tożsamość. Aczkolwiek sylaba ma z definicji pewną rozciągłość czasową, określenie jej granic na osi czasu może (prawdopodobnie w zależności od języka) być nieistotne.

Pomiary, których szczegóły pojawią się w dalszym ciągu niniejszej pracy, wskazują, że w przypadku języka polskiego istotne są dla rozstrzygnięcia określonych problemów suprasegmentalnych cechy szczytu sylabowego, którym jest samogłoska, a w określonych warunkach samogłoska wraz z przynależną do danej sylaby, poprzedzającą spółgłoską sonorną (np. [n] lub [l]). Na potrzeby analizy suprasegmentalnej, tak w zakresie wysokości, jak i iloczasu, pomiary na odcinku czasowym odpowiadającym samogłosce (wraz z ewentualną poprzedzającą głoską sonorną) okazują się wystarczające (i konieczne), a granica takiej relewantnej części sylaby jest akustycznie jednoznaczna.

Akcent był w przeszłości w literaturze lingwistycznej często traktowany tak, jakby był niezależnym parametrem opisu fonetycznego, chociaż nie przypisywano mu jakichś niezależnych cech fonetyczno-akustycznych. Bywał określany bądź w niejasnych terminach subiektywnych (np. Jones 1956), bądź w terminach artykulacyjnych (np. Ladefoged, Draper, Whitteridge 1958). Jeśli wiązano akcent z jakąś cechą akustyczną, to z reguły przypisywano mu korelacje z wymiarem amplitudy, co oznaczało, że przyjmowano, iż sylaba akcentowana jest głośniejsza niż nieakcentowana (np. Heffner 1949). Dopiero w latach 50. i 60. zaczęły się pojawiać prace postulujące dominujący związek akcentu z wysokością tonu (Fry 1955, 1958, Bolinger 1958) oraz z intonacją i iloczasem (Jassem 1962). Zarazem stało się jasne, że wbrew wcześniejszym przypuszczeniom, w żadnej z cech akustycznych akcent nie oznacza lokalnego maksimum (tj. sylaba akcentowana nie musi być najwyższa ani najgłośniejsza, ani też dłuższa od sylab nieakcentowanych). Wykazywać natomiast zaczęto pewne konfiguracje cech, które wyróżniały sylabę akcentowaną. Na przykład Jassem już w 1949 roku postulował, że w języku angielskim sylaba akcentowana jest początkiem zestroju quasi-izochronicznego (Jassem 1949), a Bolinger (1958) ujawnił 3 konfiguracje wysokości (pitch accent) dla języka angielskiego jako sygnalizujące akcent. Jassem (1962) wykazał na podstawie pomiarów przebiegu parametru F0 oraz obwiedni amplitudowej, że akcent polski też jest przede wszystkim zależny od przebiegu wysokości tonu.

Do tradycji fonetycznej należą również kategorie akcentu zdaniowego i wyrazowego (ang. word stress, sentence stress, np. u Jonesa 1956). Obecnie te kategorie zaczyna się zarzucać. Natomiast pojawiają się koncepcje akcentu realnego i akcentu potencjalnego (Ladd 1996). Te pojęcia pozwalają wyraźnie określić związki pomiędzy płaszczyzną syntaktyczną oraz leksykalną z jednej a fonetyczną lub fonologiczną z drugiej strony. Dziedziną akcentu realnego jest fraza intonacyjna. W obecnej chwili trwają jeszcze dyskusje nad definicją frazy intonacyjnej i na jej ostateczną definicję być może trzeba będzie jeszcze poczekać (Ladd 1996, s. 155, s. 222), ale na razie można przyjąć, że fraza intonacyjna jest pojęciem wiążącym składniowy poziom analizy z poziomem fonologicznym. Tymczasowo określić można frazę intonacyjną jako połączenie określonego wzorca intonacyjnego z określoną spójną strukturą składniową, przy czym wzorzec ten wykazuje jeden (i tylko jeden) ośrodek (ang. focus, nucleus, ictus). Ośrodek z kolei jest wzorcem intonacyjnym w każdym języku zdefiniowanym określonymi regułami odnoszącymi go do kierunku zmiany wysokości oraz górnej i dolnej granicy zasięgu głosu, ewentualnie również do tonu średniego. Z semantyczno-pragmatycznego punktu widzenia „ośrodek", na który przypada intonacja rdzenna, a tym samym realny akcent główny, jest związany z tą minimalną frazą syntaktyczną (np. frazą nominalną, werbalną, przymiotnikową itd.), która niesie lokalne maksimum informacji, tzn. jest przez odbiorcę najbardziej nieoczekiwana. Neutralną pozycją takiej minimalnej frazy syntaktycznej jest w wielu językach pozycja końcowa w obrębie frazy intonacyjnej. Fakt ten ma szczególne znaczenie w programowaniu modułu „grapheme-to-phoneme" w syntezie typu text- to-speech. Jeśli taki program jest krańcowo uproszczony pod względem składniowym, to zakłada się, że ośrodek intonacyjny, a zatem główny akcent realny, jest zarazem ostatnim akcentem realnym we frazie intonacyjnej.

Istotność głównego akcentu realnego w słuchowej ocenie granic frazowych potwierdziło obszerne doświadczenie percepcyjne, którego szczegóły podano w pra-

cach Demenko et al. (1996b), Demenko (1997). Jako materiał eksperymentalny przyjęto odczytane przez profesjonalnego spikera radiowego oraz dwóch fonetyków fragmenty felietonów prasowych. Doświadczenie przebiegało dwuetapowo i polegało na:

- 1. zaznaczeniu sylab akcentowanych,
  2. wyznaczeniu granic frazowych.

Wykorzystano trzy stopnie pewności ocen akcentu i trzy stopnie pewności ocen wyznaczania granic frazowych.

Akcenty uznane przez słuchaczy za silne (75% - 100% łącznych ocen, zależnie od grupy odsłuchującej) przypadają na ostatni zestrój akcentowy we frazie. Ten najsilniejszy akcent związany jest z wyrazem treściowo najważniejszym w obrębie frazy. Nawet powierzchowna analiza semantyczna załączonych tekstów świadczy, że rzeczywiście wyrazy z akcentem głównym są dla treści tekstu najważniejsze (załącznik 1). W kilku przypadkach, w których na końcu frazy występował jeszcze akcent słabszy zachodził ewidentny przykład emfazy.

Jedną z podstawowych różnic pomiędzy analizą intonacji w tradycji tzw. szkoły brytyjskiej a analizą generatywną autosegmentalno-metryczną, której uwieńczenie stanowi system ToBI, jest to, że pierwsza utrzymuje pojęcie intonacji rdzennej (nuclear tone lub nuclear tune), podczas gdy w drugiej pojęcie to zostało wyeliminowane. Analiza autosegmentalno-metryczna wprowadza za to szereg komplikujących pojęć, które znacznie utrudniają zastosowanie jej w badaniach empirycznych i dlatego tutaj jako lingwistyczną podstawę przyjęto tzw. analizę brytyjską. Istnieją zresztą próby konwersji tych dwóch systemów: ToBI i brytyjskiego (Roach 1994, Ladd 1996).

Dla języka polskiego Steffen-Batóg (1963) zdefiniowała, posługując się metodami logiki formalnej (w formie zwerbalizowanej), pojęcie frazy intonacyjnoakcentowej. Opisała szeroki materiał eksperymentalny zebrany na płaszczyźnie analizy audytywnej w terminach postulowanych kilkudziesięciu intonemów. Praca ta w okresie jej pisania (lata sześćdziesiąte) stanowiła ogromny skok w zakresie wiedzy o intonacji polskiej. Nie dokonano tam próby zintegrowania uzyskanego systemu z jakimiś uniwersaliami fonologicznymi. Nie pokazano także, czy wszystkie wyróżnione intonemy są rzeczywiście dystynktywne w sensie lingwistycznym na poziomie wyższym niż percepcyjny (audytywny).

Mimo że niniejsza praca pisana jest z punktu widzenia zastosowań w technologii mowy (por. rozdz. 1), to pewne elementy analizy lingwistycznej na poziomie fonologicznym wydają się niezbędne.

3 FIZJOLOGICZNE I SŁUCHOWE UWARUNKOWANIA INTONACJI

3.1. ASPEKTY FIZJOLOGICZNE

W procesie fonacji mówca steruje częstotliwością podstawową poprzez systemy: krtaniowy i oddechowy. Znajdujące się w krtani fałdy głosowe w czasie oddychania są rozsunięte, podczas mowy natomiast naprężone i zbliżone do siebie na odległość kilku dziesiętnych milimetra, tworząc niewielką szczelinę — głośnię. W przypadku wytwarzania głosek bezdźwięcznych, dźwięk jest generowany poprzez turbulentny przepływ powietrza w przewężeniu utworzonym przez narządy mowy (wargi, dziąsła, podniebienie, krtań). Przy wytwarzaniu samogłosek przez głośnię przepływa w ciągu sekundy około 50 - 250 cm3/sek powietrza (około 1 cm3 powietrza na 1 cykl). Cykl wibracji w czasie wytwarzania głosek dźwięcznych składa się z otwarcia oraz zamknięcia głośni. Fałdy głosowe drgają na skutek aerodynamicznych i elastycznych oddziaływań na siebie (według Hessa 1983). Nie wszystkie możliwości fonacji wykorzystywane są w mowie. Sposób fonacji określany jako „vocal fry", „creaky voice" lub laryngalizacja, występujący najczęściej na końcowych fragmentach wypowiedzi, charakteryzuje się bardzo niewielką częstotliwością w zakresie 28 - 73 Hz oraz niewielką prędkością przepływu powietrza. Zakresy zmian parametru F0 dla normalnej fonacji i fonacji falsetem mogą się częściowo pokrywać. W czasie mówienia lub śpiewania falsetem fałdy wytwarzają głośnię tylko w pewnej części, w pozostałej są zwarte i nie poddają się drganiom. W ten sposób długość drgających krawędzi skraca się i ton się podwyższa. Zakres wytwarzanych zmian częstotliwości podstawowej uzależniony jest od fizjologicznych uwarunkowań.

Częstotliwość drgań zmienia się w mowie w głosach kobiecych w zakresie 180 Hz do 400 Hz, w śpiewie przekracza nawet 1000 Hz. W głosach męskich w mowie parametr ten zawiera się przeciętnie w zakresie 60 - 200 Hz. Średnia wartość jittera (absolutnej wartości różnicy między kolejnymi okresami Ti) zależy od chwilowej wartości parametru F0 (jak podaje np. Horii 1979 według Hessa 1983). Dla F0 = 978 Hz wynosi ona 51 µs, dla F0 = 298 Hz — 24 µs. Maksimum prędkości zmian częstotliwości podstawowej wynoszące około 3 oktawy na sekundę ustalił Sundberg (1979).

Przeciętne zmiany wahają się w zakresie 2-6,5% dla średniej wartości częstotliwości = 100 Hz (na podstawie badań Takefuta 1975).

Bardziej gwałtowne zmiany mogą wystąpić:

a. po spółgłoskach zwartych,

b. na początku wypowiedzi,

c. na końcu wypowiedzi — laryngalizacja.

Przeprowadzone szczegółowe badania wykazały następujące charakterystyczne cechy tonu krtaniowego:

a. impulsy tonu krtaniowego są niesymetryczne, w przybliżeniu mają kształt trójkątny,

b. podczas normalnej fonacji głośnia w pewnym przedziale czasu jest zamknięta,

c. stosunek czasu otwarcia głośni do czasu trwania cyklu zmienia się od 0,3 do 0,7,

d. obwiednia widma tonu krtaniowego opada ze stromością — 12 dB na oktawę.

Częstotliwość drgań fałdów głosowych, posiadająca korelat akustyczny w częstotliwości podstawowej mowy, zależy od wielu czynników, z których najważniejsze są: długość oraz masa fałdów, ich napięcie oraz ciśnienie podgłośniowe Ps. Chwilowa wartość częstotliwości podstawowej jest odwrotnie proporcjonalna do długości fałdów.

Korelacja między zmianami ciśnienia podgłośniowego Ps i częstotliwością podstawową nie jest bezpośrednia. Wyższe ciśnienie powoduje zwiększenie amplitudy drgań fałdów, a nie częstotliwości ich wibracji. Efekt pośredniej korelacji może być wyjaśniony przez fakt, że wyższe ciśnienie powoduje zwiększenie powierzchni głośni — a więc większe rozsunięcie fałdów głosowych, co wywołuje większe usztywnienie tych fałdów i powoduje w konsekwencji wyższą częstotliwość podstawową. Pomiary elektromyograficzne aktywności mięśni krtaniowych podczas wytwarzania różnych wzorców intonacyjnych (Collier 1975) wykazały, że zarówno na kierunek, jak i wielkość zmian częstotliwości podstawowej przede wszystkim wpływa aktywność mięśnia CT (cricothyroideus). Przy wypowiadaniu końcowej sylaby frazy, zdania, zmniejsza się znacznie ciśnienie powietrza Ps i obniża się częstotliwość podstawowa.

Przeprowadzone na podstawie wypowiedzi jednego mówcy eksperymenty wykazały następujące prawidłowości.

1. Aktywność mięśnia CT (cricothyroideus) jest w największym stopniu w ścisłej relacji ze zmianami parametru F0. Napięcie/rozluźnienie tego mięśnia powoduje odpowiednio wzrost/spadek częstotliwości podstawowej.

2. Ciśnienie podgłośniowe jest skorelowane ze spadkiem częstotliwości podstawowej na końcowej sylabie wypowiedzi.

3. Mięśnie sternohyoideus, sternothyroideus i thyrohyoideus nie mają bezpośredniego wpływu na zmiany parametru F0 (por. Titze 1993).

Atkinson (1977) na podstawie 12 wypowiedzi *Bev loves Bob,* ilustrujących różne typy intonacji, badał wpływ na częstotliwość podstawową, następujących czynników: ciśnienia podgłośniowego, średniej aktywności EMG (electromyographic activity) mięśnia sternohyoideus — ST, średniej aktywności EMG cricoarytenoideus — CA oraz średniej aktywności EMG cricothyroideus — CT. Największy wpływ na zmiany parametru F0 wykazało funkcjonowanie mięśnia CT (dodatni współczynnik korelacji wyniósł r = 0,7) oraz mięśnia ST (ujemny współczynnik korelacji r = -0,6). Relacje typu semantyczno-składniowego między ciśnieniem podgłośniowym i częstotliwością podstawową okazały się uzależnione od typu wypowiedzi (w oznajmujących otrzymano wysoki współczynnik korelacji, w pytających niski).

Podstawowe zależności zmian częstotliwości podstawowej od długości i napięcia fałdów głosowych sformułował Jafari et al. (1988). Jeżeli rozważymy fałdy jako elastyczną membranę[3], to relacja między częstotliwością wibracji (parametrem F0) i siłą ich napięcia P może być opisana w przybliżeniu równaniem (3.1).

[3] Według podstawowej zależności obowiązującej dla membrany, zakłada się, że jest ona napięta na swoim obwodzie z siłą P (w N/m długości obwodu).

$$F_0 = C_0 \sqrt{P} \tag{3.1}$$

gdzie: C_0 — stała.

Między długością wiązadeł λ i siłą napięcia mięśni P istnieje eksperymentalna zależność (3.2)

$$P = a(e^{b\lambda} - 1) = a(e^{b\lambda}) \tag{3.2}$$

przy założeniu że $e^{b\lambda} >> 1$, gdzie a, b stałe.

Z zależności 3.1 i 3.2 wynika wzór (3.3)

$$\ln(F_0) = (b/2)\,\lambda + \ln(\sqrt{a}C_0) \tag{3.3}$$

Wyrażenie (3.4) otrzymane ze zróżniczkowania zależności (3.3) określa prędkość zmiany długości fałdów, która jest proporcjonalna do prędkości zmiany parametru F0 i odwrotnie proporcjonalna do tego samego parametru.

$$\frac{d\lambda}{dt} = \frac{2}{b}\frac{dF_0}{dt}\frac{1}{F_0} \tag{3.4}$$

Zależność (3.5) opisuje relację między prędkością zmiany napięcia fałdów głosowych i częstotliwości podstawowej.

$$\frac{dP}{dt} = 2C_0^{-2}\frac{dF_0}{dt}F_0 \tag{3.5}$$

Matematyczny model oparty na biomechanice krtani opracowany przez Titze et al. (1993) wykazał użyteczność uproszczonego opisu sterowania parametrem F0 za pomocą kilku mięśni. Wykorzystując wyniki tej pracy Farley (1994, 1996) opracował sieć neuronową (typu MLP z 7 neuronami ukrytymi) symulującą funkcjonowanie 3 mięśni: TA — thyroarytenoideus, CTO — cricothyroideus pars oblique i CTR — cricothyroideus pars rectus, generującą wartość częstotliwości podstawowej. Dla praktycznych zastosowań modelowania wysokości tonu konieczne są modyfikacje układu polegające na uwzględnieniu bardziej złożonej struktury sieci i większej ilości czynników fizjologicznych. Proces fonacji oraz wyjaśnienie fizjologicznych i fizycznych uwarunkowań mających wpływ na przebieg tonu krtaniowego stanowią ważne zagadnienia wymagające szczegółowych opracowań.

3.2. SŁUCHOWA OCENA WYSOKOŚCI TONU

Nowoczesne teorie percepcji częstotliwości podstawowej Wightmana (1973), Goldsteina (1979), Terhardta (1973, cyt. za Hess 1983) postulują, że percepcja wysokości tonu jest przeprowadzana na podstawie procesu rozpoznawania struktury harmonicznej dźwięku. Modele percepcji zawierają wstępny etap, w którym dokonuje się analizy częstotliwościowej i etap centralny, w którym otrzymuje się

wrażenie ogólne, jako percepcyjny odpowiednik częstotliwości podstawowej. Różne są natomiast założenia dotyczące organizacji procesu percepcyjnego rozpoznawania.

W 1979 r. Terhardt ustalił funkcjonalny model percepcji wysokości tonu. Obejmuje on następujące etapy:

- 1. Sinusoidalny ton wywołuje wrażenie wysokości, które jest bezpośrednio odniesione do miejsca największego pobudzenia w organie Cortiego i zwane jest spektralną wysokością tonu.
  2. Spektralne wysokości tonu mogą być percypowane indywidualnie przez słuchaczy, wówczas gdy ich uwaga jest właściwie ukierunkowana.
  3. Dźwięk złożony, niezależnie od efektu wymienionego w p. 2 wywołuje globalne wrażenie wysokości tonu, które w przypadku harmonicznego dźwięku odpowiada częstotliwości podstawowej i określane jest jako ton wirtualny *(virtual pitch, periodicity pitch* lub *residue pitch).*

Podczas gdy spektralna wysokość tonu (spectral pitch) może być uważana jako wynik peryferyjnej, słuchowej analizy, percepcja globalnej wysokości tonu zależy od wyższych — centralnych etapów procesu rozpoznawania.

Teorie Goldsteina (1973) i Terhardta (1979) oparte są na koncepcji centralnego procesora, w którym określony jest największy wspólny podzielnik harmonicznych sygnału. Szczególnie istotne w percepcji wysokości tonu są harmoniczne od trzeciej do szóstej. Słuchacz może zauważyć bardzo niewielkie zmiany częstotliwości podstawowej, nawet poniżej 1 Hz (np. Rakowski 1971, 1991, Hess 1983) i na krótkich sygnałach rzędu 30 ms (w tym przypadku przy założeniu, że F0 wynosi powyżej 100 Hz, tak aby możliwy był odbiór 2-3 okresów). Wysokość tonu jest percypowana w zakresie częstotliwości 40 Hz - 4000 Hz. Dla wyższych częstotliwości (np. 10 KHz) i krótkich bodźców (krótszych od 60 ms) percepcyjna precyzja maleje (’t Hart 1991). DL — próg rozróżnienia (difference limen) maleje wraz z częstotliwością: dla 125 Hz DL wynosi 0,74 Hz, dla 4000 Hz — DL wynosi aż 21 Hz.

Zgodnie z wynikami prac Saslowa i Flanagana (według Hessa 1983), w mowie syntetycznej, w krótkich logatomach zauważalna jest zmiana parametru F0 rzędu 0,35 - 0,5%. W mowie naturalnej próg ten jest nieco wyższy. Isačenko i Schädlich (1966) jako istotne percepcyjnie określili zmiany rzędu 5% (dla częstotliwości 150 Hz), Rossi i Chafcouloff (1972, cyt. za Hess 1983) — 4% (dla częstotliwości 195 Hz). Mikrozmiany parametru F0 (rzędu 5%) nie są zauważalne jednostkowo, ale ich ciągły efekt daje się zauważyć w percepcji. Powoduje on wrażenie naturalności głosu i jest dobrze odróżnialny słuchowo od sztucznego szumu wprowadzanego niekiedy do syntetycznych przebiegów częstotliwości podstawowej (w celu zwiększenia naturalności wypowiedzi).

Biorąc pod uwagę przebieg komunikacji za pomocą języka naturalnego, jako percepcyjnie istotne należy uznać zmiany równe lub większe od 3 półtonów. Jeżeli mają być percepcyjnie odróżnione np. dwa wzrosty F0, z których pierwszy występuje w zakresie 125- 150 Hz (tj. wynosi 3,16 półtonu), to drugi aby został odebrany słuchowo jako większy powinien być np. w zakresie 100- 143 Hz (tj. obejmować 6,19 półtonu). Dodatkowo na percepcję wysokości tonu w mowie mają wpływ skoki amplitudy. Zmiana amplitudy z 10 dB na 20 dB w przedziale czasowym rzędu kilku milisekund, (np. przy przejściach ze spółgłoski na samogłoskę) może przesłonić słuchowo zmianę wysokości tonu w tym fragmencie. Jako językowo istotne przyjmuje się znacznie większe zmiany parametru F0 — rzędu kilkunastu procent, ’t Hart i Collier (1975), ’t Hart (1981), ’t Hart et al. (1990), dla rozróżnienia słuchowego dwóch konturów melodycznych przyjmują zmianę 3-4 półtonów tj. 18%--25%. DL zależy od stopnia złożoności wypowiedzi. Dla zdań krótkich, trwających 2-3 sekundy, słuchacz zauważa zmiany częstotliwości rzędu 5 Hz w obrębie całego konturu. Dla zdań długich zauważalne są dopiero zmiany wynoszące 15 Hz.

DL jest więc dla wypowiedzi złożonych około 50 razy większy niż dla samogłosek izolowanych, ’t Hart stawia hipotezę, że odbierane wrażenie zależy od kategorii słuchania. Przy „szerokim” sposobie odbioru informacji słuchacz otrzymuje całkowite wrażenie, które wykorzystuje do klasyfikacji intonacyjnych całości. Natomiast przy „wąskim” — analitycznym słuchaniu, jego uwaga jest skoncentrowana na szczegółach, a więc istnieje możliwość wykrycia drobnych, szybkich zmian. W ostatnich pracach, np. ’t Hart et al. (1991) i Kohler (1997) zwracają uwagę na zróżnicowanie percepcyjnej ważności poszczególnych fragmentów przebiegu. Słuchacz nie jest jednakowo wrażliwy na wszystkie zmiany w konturze melodycznym. Podkreśla się rolę samogłosek (ewentualnie samogłosek z sąsiednimi głoskami sonornymi) w percepcji struktur melodycznych wypowiedzi.

Interesującym zagadnieniem jest percepcja wysokości tonu w wypowiedziach zawierających segmenty bezdźwięczne. Słuchacz odbiera ogólne wrażenie ciągłości przebiegu melodycznego. Z akustycznego zaś punktu widzenia, zależnie od struktury sylabicznej wypowiedzi, otrzymuje się szereg wyraźnie wyodrębniających się przebiegów parametru F0. Jak dotychczas nie istnieje algorytm, który umożliwiałby symulację zmian wysokości tonu w bezdźwięcznych fragmentach wypowiedzi.

Sformułowanie zasad percepcyjnego przetwarzania zmian wysokości tonu wymaga jeszcze wielu podstawowych badań. Nie wyjaśniono np. w jaki sposób słuchacze przeprowadzają normalizację częstotliwościową oraz czasową sygnału mowy i wyodrębniają charakterystyczne wzorce intonacyjne, cechy głosu mówcy, niezależnie od tempa mowy i skali zmian parametru F0.

4 AKUSTYCZNO-PERCEPCYJNE PODSTAWY OPISU STRUKTUR SUPRASEGMENTALNYCH MOWY

4.1. RELACJE MIĘDZY CZĘSTOTLIWOŚCIĄ PODSTAWOWĄ, ILOCZASEM ORAZ INTENSYWNOŚCIĄ

W większości opracowań dotyczących struktur melodycznych różnych języków (np. angielskiego, niemieckiego i polskiego) przyjmuje się, że częstotliwość podstawowa posiada zasadnicze znaczenie w percepcji akcentu wypowiedzi, natomiast iloczas i intensywność sygnału spełniają funkcje podrzędne.

Percepcja intensywności zależy w dużej mierze od uwarunkowań czasowych. Jeżeli czas trwania obserwacji jest zbyt krótki, to zmiany intensywności są niezauważalne (np. Sorin 1981). Czas trwania samogłoski powinien wynosić około 200 ms, aby intensywność zaistniała jako cecha suprasegmentalna. Analogicznie Rossi (1978) stwierdził, że zmiana intensywności mniejsza niż 11 dB w 200 ms stacjonarnych samogłoskach nie jest zauważalna percepcyjnie. Podobne wyniki, dotyczące wpływu iloczasu trwania samogłoski (w sekwencji logatomów *tatatata…)* na głośność sylaby, otrzymali Nishinuma et al. (1984). Autorzy stwierdzili, że istotny dla percepcji zmian intensywności czas trwania samogłoski powinien wynosić od 200 - 250 ms.

Pośredni związek między intensywnością i częstotliwością podstawową ma podstawę fizjologiczną. Obydwa parametry podczas wytwarzania mowy sterowane są przez ten sam mechanizm polegający na zwiększeniu lub zmniejszeniu ciśnienia podgłośniowego. Brak bezpośredniego związku między intensywnością i częstotliwością podstawową wynika z modyfikacji podczas artykulacji zmian intensywności w torze głosowym oraz z faktu, że ciśnienie podgłośniowe nie jest wyłącznym czynnikiem wpływającym na wibrację wiązadeł głosowych. Zwykle obydwa parametry są dodatnio skorelowane (np. Vaissiere 1983).

Zjawisko dodatniej korelacji pomiędzy częstotliwością podstawową i poziomem intensywności zaobserwowano również dla wypowiedzi izolowanych języka polskiego. Według badań Nowakowskiej (1977) przeprowadzonych na wypowiedziach izolowanych języka polskiego współczynnik korelacji między parametrem F0 i intensywnością wynosi 0,82, natomiast współczynnik korelacji pomiędzy intensywnością i czasem trwania sylaby akcentowanej równy jest 0,4.

Zee (1978) dla 5 tonów języka tajwańskiego wykazał, że wysokie tony mają wyższą intensywność niż średnie, średnie wyższą niż niskie.

Iloczas segmentów akustycznych zmienia się między innymi zależnie od pozycji w wyrazie i zdaniu (zjawisko to badała szczegółowo dla języka hebrajskiego Berkovits 1993, 1994).

Dla niektórych języków (np. szwedzkiego) podjęto próbę ustalenia zależności między iloczasem samogłoski i konturem intonacyjnym. Lyberg (1981) analizował związek między wydłużaniem końcowym samogłosek we frazie i typem zmiany częstotliwości podstawowej w wypowiedziach języka szwedzkiego. W wyniku testowania dwóch hipotez: pierwszej, zakładającej wydłużenie samogłosek końcowych jako konsekwencji występowania akcentu i konieczności realizacji spadku lub wzrostu częstotliwości podstawowej oraz drugiej, przyjmującej niezależność wydłużenia samogłoski końcowej od akcentowania, została ustalona korelacja między zmianami częstotliwości podstawowej i iloczasem samogłoski znajdującej się na końcu frazy. Wpływ na iloczas wzrostu parametru F0 opisuje relacja 4.1 (według Lyberga 1981, s. 102):

$$D_{vw} = a + k_r \Delta F_r \quad (4.1)$$

spadku parametru F0 określa zależność 4.2

$$D_{vs} = a + k_f \Delta F_f \quad (4.2)$$

brak zmian częstotliwości podstawowej związanych z akcentowaniem uwzględnia relacja (4.3)

$$D_{vr} = a + b \quad (4.3)$$

gdzie: D_{vw}, D_{vs}, D_{vr} — iloczasy końcowej samogłoski,
$\Delta F_r/\Delta F_f$ — interwały wzrostu/spadku parametru F_0,
a, b, k_r , k_f — stałe.

Dla języka szwedzkiego Fant et al. (1989, 1994, 1995, 1997) zwrócili uwagę na podrzędną rolę intensywności w akcentowaniu wypowiedzi. Zmiany intensywności na sylabach akcentowanych, mieszczące się zwykle w zakresie od 0 dB do 6 dB, nie są istotne percepcyjnie. W analizie intensywności mowy w wypowiedziach tekstów czytanych, zauważono od początku do końca danej wypowiedzi stopniowe obniżanie się poziomu sygnału (rzędu 9 dB), skorelowane ze spadkową tendencją przebiegu parametru F0. Autorzy zwrócili uwagę na konieczność analizy relacji między cechami suprasegmentalnymi, z uwzględnieniem pozycji poszczególnych fragmentów wypowiedzi: początkowych, środkowych lub końcowych.

Relacje między cechami fizycznymi sygnału mowy: częstotliwością podstawową, intensywnością oraz iloczasem segmentu fonetyczno-akustycznego są złożone i stanowią nadal aktualny do rozwiązania problem, zwłaszcza w syntezie oraz rozpoznawaniu mowy spontanicznej. Z powyższych rozważań wynika, że bardziej szczegółowego omówienia wymaga organizacja czasowa wypowiedzi.

4.2. CZASOWA STRUKTURA WYPOWIEDZI

Badania przeprowadzone w latach 70. przez Hugginsa (1972) oraz Raphaela (1971) wykazały, że percepcja iloczasu odbywa się na poziomie sylaby, a nie pojedynczego segmentu fonetycznego. Ponadto okazało się, że słuchacze są bardziej wrażliwi na zmiany czasu trwania samogłoski niż spółgłoski. Czas trwania samogłoski może zawierać informację o następującej spółgłosce. Słuchacze percypują

końcowy segment konsonantalny jako bezdźwięczny, kiedy go poprzedza krótka samogłoska, i jako dźwięczny, kiedy go poprzedza długa samogłoska. Wykazano również (między innymi Huggins 1972, Lehiste 1970, 1977, Kato et al. 1997), że w percepcji akcentu ważny jest interwał czasowy między początkami akcentowanych sylab.

Dla jednofrazowych neutralnych wypowiedzi języka szwedzkiego, Lindblom (1975) zaobserwował, że im fraza jest dłuższa, tym większe następuje skracanie jej sylab. Czas trwania sylaby został opisany zależnością (4.4).

$$D_s = \frac{\bar{D}_S}{(m_p + 1)^{C_1} + (m_n + 1)^{C_2}} \qquad (4.4)$$

gdzie: D_S — czas trwania sylaby,
$\bar{D}_S$ — uśredniony czas trwania sylaby,
m_p, m_n — liczba sylab poprzedzających daną sylabę oraz następujących po niej,
c_1, c_2 — stałe.

Czas trwania poszczególnych segmentów fonetycznych jest wielokrotnie modyfikowany na różnych poziomach tworzenia wypowiedzi. Artykulacja związana jest z fizjologicznymi uwarunkowaniami wpływającymi na sposób wytwarzania sygnału oraz funkcjonowanie aparatu głosowego. Po akustycznej transformacji sygnał mowy kodowany jest w postaci dystynktywnych, fizycznych wzorców w wymiarze czasu, obserwowalnych w analizie spektrograficznej.

Na czasową strukturę zdania wpływają więc następujące czynniki (według Klatta 1976):

a. niejęzykowe — fizjologiczne i fizyczne warunkujące prędkość mowy,

b. semantyczne — uwydatniające znaczenie określonych informacji,

c. syntaktyczne — warunkujące wydłużanie struktur fonetyczno-akustycznych w pobliżu granic frazowych,

d. czynniki wynikające z pozycji zdania, frazy, wyrazu w tekście,

e. czynniki fonologiczne — określające specyficzny czas trwania segmentu (tzw. iloczas właściwy),

f. akcentuacja i rytm,

g. efekty segmentalne.

Według badań Klatta, typowa szybkość wypowiedzi waha się w granicach 150 - -250 wyrazów na minutę, 4-7 sylab na sekundę (przy nieuwzględnianiu pauz dłuższych od 200 ms). Zmienia się ona między innymi zależnie od głosu parlatora, rodzaju wypowiedzi (np. komentarz sportowy, wykład). Pauzy zajmują 20% czasu podczas płynnego czytania i około 50% w dialogach. Wpływ czynników semantycznych zauważa się przy podkreśleniu nowej informacji — poprzez tzw. akcent emfatyczny. Udział syntaktycznych czynników, obserwowany w badaniach eksperymentalnych ujawnia się (nawet w przypadku braku fizycznej pauzy w sygnale) znacznym, często dochodzącym do kilkudziesięciu procent, wydłużeniem końcowych sylab (głównie ostatniej samogłoski lub też samogłoski i spółgłosek sonornych lub spółgłosek trących występujących po samogłosce). Nie jest wiadome, czy mówca wydłuża sylaby znajdujące się w pobliżu granic frazowych, aby zwiększyć uwagę słuchacza, czy jest to tendencja wynikająca z zakończenia pracy artykulatorów. Wydłużanie sylab w pobliżu granic wyrazowych nie zawsze występuje.

Fonetyczne aspekty iloczasu ujawniają się w 3 głównych postaciach:

- 1. Specyficznego, fonologicznego czasu trwania fonetycznego segmentu; np. samogłoska [z] jest krótsza niż inne samogłoski bardziej otwarte,

spółgłoski trące bezdźwięczne są około 40 ms dłuższe niż ich dźwięczne odpowiedniki.

2. Redukcji sylab nieakcentowanych. Największą różnicę obserwuje się na końcowej sylabie frazowej, gdzie nieakcentowana samogłoska może być około 65% krótsza, niż gdyby była akcentowana.
3. Wpływu następującej spółgłoski. Istnieje tendencja w wielu językach skracania samogłoski, jeżeli następuje po niej spółgłoska bezdźwięczna.

Próbę implementacji reguł Klatta dla języka angielskiego w syntezie MIT podjęli Allen et al. (1987, s. 95). Formuła obliczania czasu trwania fonemu uwzględniała jego charakterystyczny czas trwania Dpw — (typowy czas trwania fonemu w pozycji początkowej wyrazu przed sylabą akcentowaną i nie występującego w zbitce spółgłoskowej) oraz minimalny czas trwania fonemu Dpmin w zależności od kontekstu uwzględniającego iloczasowe uwarunkowania oraz czynnik korekcyjny Cp (procentowy stopień skracania fonemu określony 10 regułami wynikającymi między innymi z syntaktycznej oraz akcentowej struktury wypowiedzi).

Czas fonemu Dp ujęto w postaci wzoru (4.5).

$$D_p = (D_{pw} - D_{pmin})\, C_p/100 + D_{pmin} \tag{4.5}$$

Próbę statystycznej normalizacji iloczasu fonemu (zależność 4.6) przeprowadzili Campbell i Isard (1991, s. 39). Na podstawie 200 zdań określili parametry statystycznych rozkładów wartości iloczasu: średniej μi, oraz odchylenia standardowego δi iloczasu poszczególnych fonemów (zależność 4.6).

$$D_p' = (D_p - \mu_i)/\delta_i \tag{4.6}$$

gdzie: D_p' — iloczas znormalizowany,
D_p — iloczas obserwowany.

W nowszej pracy Campbell (1993 s. 346) przedstawił zmodyfikowaną regułę obliczania iloczasu w sylabie (zależność 4.7).

Iloczas fonemu w sylabie określony jest średnią i odchyleniem standardowym rozkładu logarytmów iloczasów fonemów z bazy danych odpowiadającej danej sylabie (zależność 4.7).

$$D_{psyl} = \sum_{i=1}^{n} \exp\,(\mu_i + k\delta_i) \tag{4.7}$$

gdzie: D_{psyl} — sumaryczny czas wszystkich fonemów w sylabie,
n — liczba fonemów w sylabie,
μ_i oraz δ_i — średnia i odchylenie standardowe rozkładu,
k — stała zależna od średniej długości segmentu.

Wightman et al. (1992) dla mowy czytanej (na podstawie analizy 280 zdań języka angielskiego) podał statystyczną normalizację czasu trwania fonemu analogiczną do zastosowanej przez Campbella (1992).

Stosując podaną powyżej normalizację, autorzy określili 4 percypowane poziomy wydłużenia jednostek fonetycznych (jeden na poziomie wyrazu, dwa na poziomie frazy, jeden na poziomie zdania). Stwierdzono, że w największym stopniu wydłużeniu ulega ostatnia samogłoska frazy lub zdania.

Na podstawie analiz 11 różnych struktur sylab (np. typu: cv, v, cvc, cvccc), pochodzących z mowy czytanej (tekst w języku angielskim), Crystal et al. (1990)

ustalili, że głównymi wyznacznikami statystycznego iloczasu sylaby jest średnia liczba fonemów na sylabę i akcent (jeżeli występuje).

Multiplikatywny model, określający całkowity czas trwania samogłoski (zależność 4.8) z uwzględnieniem pozycji samogłoski oraz struktury jej fonetycznego otoczenia, przedstawił Santen (1997a, s. 235).

$$D_V\,(V, VOI, POS) = S_1\,(V)\,S_2\,(VOI)\,S_3\,(POS) \tag{4.8}$$

gdzie: D_V — iloczas samogłoski,
V — typ samogłoski,
VOI — cecha dźwięczności postwokalicznej spółgłoski,
POS — pozycja samogłoski we frazie,
S_1 oraz S_2, S_3 — wagi skalujące.

Grover, Terken (1994) na podstawie eksperymentów przeprowadzonych dla języka szwedzkiego oraz niemieckiego stwierdzili, że mówcy nie kontrolują precyzyjnie czasu trwania poszczególnych fonemów w sylabie. Następstwem tego stwierdzenia jest wniosek, że czas trwania sylaby nie może być wyjaśniony jako suma poszczególnych składowych — fonemów sylaby. Kontrola iloczasu powinna uwzględniać również wpływy rytmiczne.

Jednym z czynników decydujących o iloczasie głosek jest rytm. W wielu językach rytm mowy jest związany ze zjawiskiem izochronizmu polegającym na tendencji do zachowania względnie stałej długości jednostek rytmicznych, niezależnie od liczby sylab. W językach, w których obowiązuje zasada izochronizmu, segmenty fonetyczne ulegają w mniejszym lub większym stopniu skracaniu wraz z wydłużeniem jednostki rytmicznej (np. Lindblom 1975), Nakatani et al. (1981). Proponowane teorie rytmu i izochronizmu są przedmiotem dyskusji. Między innymi Lehiste (1977, s. 256) stwierdziła, że niektóre aspekty danych przemawiają za obecnością izochronizmu, a inne przeciwko niemu. Publikacje z zakresu iloczasu głoskowego, rytmu i tempa mowy są w języku polskim nieliczne; np. Łobacz (1976), Richter (1983, 1987), Steffen-Batogowa (1987). Dla języka angielskiego teoria rytmu i izochronizmu, zaproponowana przez Jassema (1984), zakłada dwie jednostki: anakruzę (nie wykazującą izochronizmu) i ścisłą jednostkę rytmiczną.

Dla mowy spontanicznej, Batliner et al. (1996) i Batliner (1997) do formuły normalizacji iloczasu fonemu (zależność 4.9) wprowadzili czynnik τ uwzględniający tempo mowy (zależność 4.10), obliczany dla dłuższych fragmentów tekstu np. frazy, zdania.

$$D_p' = \frac{1}{l}\sum_{i=1}^{l}\frac{D_{pi} - \tau\mu_{(i)}}{\tau\delta_i} \tag{4.9}$$

$$\tau = \frac{1}{k}\sum_{i=1}^{k}\frac{D_{pi}}{\mu_{(i)}} \tag{4.10}$$

gdzie: D_p' — średni znormalizowany czas fonemu,
D_{pi} — iloczas i-tego fonemu,
μ_i — średni czas trwania fonemu i,
δ_i — odchylenie standardowe dla fonemu i,
l — liczba fonemów w sylabie,
k — liczba fonemów w ciągu mowy nie zawierającym pauz.

Na konieczność rewizji reguł w zakresie modelowania iloczasu fonemów na podstawie szczegółowych badań 5000 izolowanych wyrazów, 200 zdań izolowanych, 200 zdań mowy ciągłej oraz 20 minut mowy spontanicznej zwrócił uwagę w ostatnich pracach Campbell (1997). Wyniki jego pracy wykazały, że dla izo-

lowanych wypowiedzi parametry statystyczne rozkładów iloczasu: średnie i wariancje mają określony, niewielki zakres, w tekście czytanym rozrzut parametrów wzrasta, natomiast w mowie spontanicznej rozrzut jeszcze bardziej się zwiększa (wariancja wzrasta prawie 2 razy).

4.3. AKUSTYCZNE WYZNACZNIKI AKCENTU

Od kiedy zaistniała możliwość syntezy oraz resyntezy mowy, zaczęto badać w sposób systematyczny wpływ poszczególnych cech zmienności parametru F0 na percepcję akcentu. Obszerne badania w tym zakresie przeprowadzili między innymi de Pijper (1983), Collier (1990, 1991), 't Hart et al. (1990). Opracowali metodę pozwalającą określić, które zmiany częstotliwości podstawowej są istotne w percepcji melodii mowy. Przebieg parametru F0 aproksymowano minimalną liczbą prostych odcinków w taki sposób, aby różnice percepcyjne były niezauważalne. Wynikowy przebieg zwany „dokładną kopią konturu" — (close copy) opisano w kilku wymiarach:

a. kierunek (wzrost, spadek),
b. rodzaj zmiany (gwałtowna, stopniowa),
c. wielkość (duża, mała),
d. synchronizacja (timing) zmiany odniesiona do początku samogłoski (wczesna, późna).

Maksymalna liczba kategorii jest uzależniona od danego języka i od uwarunkowań percepcyjnych. Stylizacja tego rodzaju (Collier 1991) wykazała użyteczność dla opisu intonacji języka flamandzkiego, angielskiego, niemieckiego i rosyjskiego.

Vaissière (1988, 1995) zwróciła uwagę na uniwersalne, niezależne dla danego języka cechy zmienności przebiegów częstotliwości podstawowej w zakresie:

a) wytwarzania archetypowego rosnąco-opadającego wzorca przebiegu parametru F0 oraz podobnego przebiegu intensywności, b) podobieństwa w zakresie realizacji wzorców intonacyjnych występujących na końcach frazy, c) powtarzania wzorców intonacyjnych.

Przeprowadzone dla poszczególnych języków eksperymenty w zakresie akcentacji, między innymi przez Berkovits (1994), Verhoeven (1994), Hermesa (1995), Hermesa i Rumpa (1994), Terkena (1993, 1997), Streefkerk et al. (1996), Streefkerk (1997), Skorka (1997), Swerts et al. (1994) i Swerts (1997), często znacznie różniące się metodologiami, można poklasyfikować według badanych cech częstotliwości podstawowej, wpływających na akcentuację wypowiedzi, takich jak np.: umiejscowienie zmiany parametru F0, szybkość, interwał zmiany oraz typ intonacji. Dodatkowo także analizuje się strukturę sylaby oraz kontekst (np. Volskaya 1998). Czynniki te są ze sobą ściśle skorelowane i wszystkie związane są ze zmiennością częstotliwości podstawowej w czasie. Ponieważ w literaturze spotyka się pięć poniższych czynników, wszystkie one zostaną pokrótce omówione.

4.3.1. UMIEJSCOWIENIE ZMIANY PARAMETRU F0

Hill i Reid (1977) analizowali dla języka angielskiego percepcyjną wrażliwość słuchaczy na zmiany pozycji początków wzrostu częstotliwości podstawowej w parach wyrazowych. Słuchacze oceniali, czy wzrost w drugim wyrazie pojawił się

później niż w pierwszym. Początki wzrostów parametru F0 zmieniano w zakresie 10-70 ms. W wyniku eksperymentu otrzymano 3 kategorie zmian:

a. wzrosty, które zazwyczaj zaczynają się na prewokalicznej spółgłosce i trwają aż do stanu ustalonego samogłoski,

b. wzrosty rozpoczynające się w ustalonej części samogłoski i trwające aż do ustalonego stanu następującej spółgłoski,

c. wzrosty całkowicie usytuowane poza stanem ustalonym samogłoski.

Dla języka angielskiego istotność wczesnej i późnej zmiany parametru F0 w percepcji akcentu potwierdzili między innymi Pierrehumbert i Steele (1989). Dla języka duńskiego Thorsen (1978, 1982, 1988), określiła jako istotną zmianę umiejscowienia wzrostu parametru F0 (rzędu 2 półtonów) między pierwszą i drugą sylabą (przed lub po interwokalicznej spółgłosce sonornej). W języku szwedzkim o znaczeniu wyrazu (segmentalnie identyczne wyrazy mogą być realizowane albo z akcentem 1 — grave, albo 2 — acute) i o wyborze akcentu decyduje późniejszy spadek parametru F0 na drugiej sylabie akcentowanej (Gardning at al. 1989).

Dla języka holenderskiego ’t Hart et al. (1990) określili 5 typów zmian parametru F0 niosących akcent: dwa z nich związane ze wzrostem częstotliwości podstawowej, a różniące się umiejscowieniem w obrębie sylaby (jeden rozpoczynający się wcześnie i jeden późno), dwa następne związane ze spadkami częstotliwości podstawowej (różniące się interwałem zmian w obrębie sylaby) oraz ostatni typ związany ze zmianami stanowiącymi kombinację wczesnego wzrostu i spadku. Synchronizacja czasowa wzrostów względem początku samogłoski okazała się decydująca w ocenie akcentowania, synchronizacja spadków nie była istotna. Wzrost rozpoczynający się przed początkiem samogłoski oraz wzrost połączony ze spadkiem powodowały podobne percepcyjne wrażenie obecności akcentu. Dwa typy spadków parametru F0 zróżnicowane zostały przez ich wielkość, a nie synchronizację czasową względem samogłoski.

House et al. (1997) na podstawie analiz percepcji akcentu w językach holenderskim, szwedzkim i francuskim wyróżnił 4 kategorie wzrostu oraz 6 kategorii spadku częstotliwości podstawowej.

W 77% akcentowanych wyrazów lokalne maksimum przebiegu parametru F0 leży w obrębie samogłoski akcentowanej. Stwierdzenie to na podstawie materiału składającego się z 7 wypowiedzi języka holenderskiego czytanych przez 8 osób dokonane zostało przez Streefkerk et al. (1996). Słuchacze oceniali, który z wyrazów był akcentowany. Ważną różnicę między sylabami akcentowanymi i nieakcentowanymi stanowiła synchronizacja zmiany częstotliwości podstawowej z obecnością — początkiem, środkiem lub końcem trwania samogłoski.

4.3.2. SZYBKOŚĆ ZMIANY PARAMETRU F0

Maksymalna prędkość zmiany częstotliwości podstawowej, jak stwierdził Sundberg (1979), może wynosić 120 półtonów na sekundę. W większości języków występują jednak znacznie wolniejsze zmiany — rzędu 50 półtonów na sekundę. Obszerne doświadczenia poświęcone percepcji szybkości zmian parametru F0 przeprowadzili Hasegawa et al. (1992). Wykazali, że dla języka japońskiego zarówno lokalizacja, jak i prędkość zmiany częstotliwości wpływa na akcent. Im później pojawia się maksimum przebiegu parametru F0 na danej sylabie, tym bardziej konieczna jest większa prędkość spadku po to, aby słuchacze odebrali poprzedzającą sylabę jako akcentowaną. Spadek częstotliwości podstawowej zmieniał się w zakresie od 0,44 Hz/ms do 2,4 Hz/ms. Jeżeli występowała mała prędkość spadku np. 0,44 Hz/ms, to akcent percypowany był na tej sylabie, na której pojawiło się maksimum przebiegu parametru F0. Jeżeli po maksimum przebiegu występował

stromy spadek, np. 2,4 Hz/ms, to słuchacze jako akcentowaną słyszeli poprzedzającą sylabę.

Możliwość percepcji akcentu na sylabie, na której nie występuje maksimum przebiegu parametru F0 potwierdzono również dla innych języków (np. ’t Hart et al. 1990; Kohler 1991, 1995; Bruce 1995; Bruce et al. 1991, 1995).

4.3.3. WIELKOŚĆ ZMIANY CZĘSTOTLIWOŚCI PODSTAWOWEJ

W obrębie jednej sylaby dwa wzrosty przebiegu częstotliwości podstawowej nie są tak długo percepcyjnie rozróżnialnc, aż różnica w ich interwałach nie osiągnie 3,5 półtonu (por. ’t Hart 1976). W mowie ciągłej w zakresie zmian parametru F0 rzędu oktawy mogą być percepcyjnie wyróżnione co najmniej 3-4 zakresy (np. ’t Hart 1981).

4.3.4. KONTUR

Próbę oceny melodii krótkich, syntetycznych wypowiedzi języka angielskiego, w których sterowano typem konturu intonacyjnego (modelowanego na podstawie 6 punktów) oraz zakresem zmian parametru F0 podjęli Ladd et al. (1985, 1994). W konturze typu pierwszego na ostatniej sylabie akcentowanej dwóm przedostatnim punktom konturu nadano niskie wartości parametru F0, w konturze typu drugiego przyjęto na tym fragmencie przebiegu maksymalne wartości parametru. Pozostałe cztery kontury posiadały wartości pośrednie. Okazało się, że niezależnie od indywidualnych różnic w percepcji najistotniejsze są zmiany w konturze i zakresie zmian częstotliwości podstawowej.

Zitter (1992) analizował percepcję zmian kształtu konturów, należących do tego samego wzorca intonacyjnego oraz wielkości interwałów przebiegu parametru F0, wywołujących wrażenie akcentu. Jako materiał eksperymentalny przyjął dwusekundowe zdanie złożone z 7 sylab. Decydujący o akcencie wzrost przebiegu parametru F0 rozpoczynał się 70 ms przed początkiem samogłoski. W konturach intonacyjnych składających się ze wzrostu oraz spadku (pointed hat), decydujący o akcencie spadek przebiegu rozpoczynał się 80 ms po początku samogłoski. W przebiegach płaskich (fiat hat), prowadzący akcent spadek rozpoczynał się 20 ms przed początkiem samogłoski. Czas trwania zarówno wzrostu, jak i spadku wynosił 120 ms. Dla pierwszego maksimum przebiegu przyjęto interwał zmian w zakresie 5 i 9 półtonów, dla drugiego 5, 7, 9 półtonów. Stwierdzono percepcyjną hierarchię: różnica w kształcie konturu okazała się zawsze istotna, różnica w wysokości pierwszego maksimum przebiegu parametru F0 była ważna tylko wtedy, kiedy kształt konturu nie zmieniał się. Różnice w wysokości drugiego maksimum konturu okazały się percepcyjnie istotne tylko wtedy, jeżeli zarówno kształt konturu, jak i wysokość pierwszego maksimum były stałe i jeżeli zmiany te wynosiły przynajmniej 4 półtony.

4.3.5. TYP INTONACJI

Verhoeven (1994) oceniała, czy słuchacze są jednakowo wrażliwi na zmiany w intonacji rosnącej i opadającej. W dwóch syntetycznych wypowiedziach języka angielskiego przesuwano (co 10 ms od początku sylaby akcentowanej) wzrost lub spadek częstotliwości podstawowej (rzędu 5 półtonów). Wyniki doświadczenia wykazały

większą wrażliwość słuchaczy na intonację opadającą niż rosnącą (jako wartość progową przyjęto 70 ms dla spadków i 95 ms dla wzrostów), przy czym umiejscowienie wzrostu lub spadku częstotliwości podstawowej w obrębie sylaby odgrywało zasadniczą rolę. Większą wrażliwość słuchaczy na intonację opadającą potwierdzili również Hermes i Rump (1994) oraz Hermes (1995).

Na syntetycznej wypowiedzi *mamamamama* (o długości 0,77 s) z założoną deklinacją w zakresie: 3,17 Erba-2,63 Erba (93 Hz-75 Hz lub 0,5 półtonu/s), przyjęto 6 wielkości zmian parametru F0 na drugiej sylabie wypowiedzi w zakresie: 0,56-1,94 Erba[4]. Analizowano, czy słuchacze jednakowo percypują różne typy zmian częstotliwości podstawowej. Okazało się, że wcześnie rozpoczynający się wzrost lub zmiana typu wzrost-spadek, jeżeli miały analogiczne zakresy zmian, były podobnie percypowane, natomiast spadek częstotliwości podstawowej wywoływał większe uwydatnienie percepcyjne niż wzrost lub wzrost-spadek, niezależnie od umiejscowienia spadku w obrębie sylaby.

4.3.6. STRUKTURA SYLABY

Rietveld, Gussenhoven (1995) zwrócili uwagę na tendencję w przesunięciu lokalizacji początku zmiany parametru F0, zależnie od długości sylaby oraz jej segmentalnej struktury. W dłuższych sylabach występowało późniejsze maksimum przebiegu parametru F0 (z opóźnieniem rzędu 15 ms). Podobne opóźnienie w umiejscowieniu maksymalnej wartości częstotliwości podstawowej (rzędu 15 ms) względem początku samogłoski zauważono również w przypadku pojawienia się sono- rantu przed samogłoską.

4.3.7. KONTEKST

Przeprowadzano także eksperymenty poświęcone percepcji intonacji w wypowiedziach zawierających kilka akcentów. Pierrehumbert (1979) modyfikowała parametr F0 w wypowiedzi syntetycznej *ma Ma mama Ma ma*. Jeżeli obydwie akcentowane sylaby miały takie same wartości parametru F0 i taką samą intensywność, to druga sylaba akcentowana wydawała się wyższa. Jeżeli amplituda drugiej sylaby była 4 dB niższa niż pierwsza, to aby wywołać podobne percepcyjne uwydatnienie obu sylab, wartość parametru F0 dla drugiej sylaby musiała być o 11 Hz niższa niż dla pierwszej. Jeżeli obie sylaby miały taką samą intensywność, to aby wywołać wrażenie równej wysokości obu sylab, druga z nich musiała być o 17 Hz niżej niż pierwsza akcentowana sylaba. Na różną czułość percepcyjną na zmiany tonu w obrębie wypowiedzi zwrócili również uwagę 't Hart (1976, 1981), Thorsen (1978), Rietveld, Gussenhoven (1985, 1992/1993). Ladd et al. (1994) zweryfikowali doświadczenie przeprowadzone przez Gussenhovena i Rietveld (1992/1993), w którym słuchacze oceniali uwydatnienie drugiego akcentu w syntetycznej wypowiedzi zawierającej 2 akcenty. W wyniku tych eksperymentów stwierdzono, że obniżenie parametru F0 na pierwszym akcencie wprowadziło mniejsze percepcyjne uwydatnienie drugiego akcentu. Ladd et al. (1994) potwierdzili to spostrzeżenie, ale tylko dla wartości parametru F0 poniżej 145 Hz na drugiej sylabie akcentowanej. Powyżej wartości 145 Hz wystąpił efekt odwrotny.

Rump i Collier (1995) analizując semantyczne umiejscowienie akcentu w wypowiedzi, zauważyli, że aby otrzymać percepcyjne uwydatnienie pożądanych fragmentów wypowiedzi, musi wystąpić określona kombinacja zmian częstotliwości podstawowej na poszczególnych sylabach akcentowanych.

[4] Szcz egółowe informacje dotyczące psychoakustycznej jednostki, jaką jest Erb por. rozdz. 11.1.

Terken (1997) na podstawie oceny aktualnego stanu badań w zakresie percepcji intonacji stwierdził, że chociaż wiadomo, że uwydatnienie sylaby jest proporcjonalne do wielkości zmiany parametru F0, to w dalszym ciągu niejasne są percepcyjne reguły normalizacji konturu intonacyjnego. W szczególności brak jest odpowiedzi na następujące pytania:

a. które punkty konturu są wykorzystywane w percepcyjnej normalizacji,

b. jak słuchacze oceniają wielkość maksimów i minimów konturu,

c. jak ważna jest odległość między poziomami parametru F0 na poszczególnych samogłoskach akcentowanych.

4.4. AKUSTYCZNE WYZNACZNIKI GRANICY FRAZY

W większości prac poświęconych analizie granic frazowych zwraca się uwagę na dominującą rolę wydłużenia sylab końcowych wypowiedzi (np. Delattre et al. 1965, Delattre 1966, Umeda et al. 1981, Kohler 1983 i Gardning et al. 1991). Streeter (1978) przeprowadziła doświadczenie, w którym słuchacze lokalizowali granice frazy w wypowiedziach typu (A + *E)* × *O* [ej plas i times ow] oraz *A* + *(E*×*O)* [ej plas aj times ow]. Wykazała, że najważniejszy dla percepcji granicy frazowej jest czas trwania sylaby końcowej oraz przebieg parametru F0. W doświadczeniu przeprowadzonym przez Harris et al. (1981) 9 słuchaczy analizowało percepcyjnie tekst złożony z 3500 wyrazów czytanych przez 5 mówców. Każdy słuchacz zaznaczał granice akcentu oraz określał kryteria swojej decyzji. W 83% słuchacze byli zgodni co do granic, natomiast kryteria wyboru cech różnicujących akcent bądź jego brak różniły się. Najczęściej słuchacze jako istotne cechy granicy frazy określali pauzę, wydłużenie końcowych segmentów oraz zmiany częstotliwości podstawowej.

Steffen-Batogowa i Katulska (1984) na podstawie obszernego materiału językowego (3500 sylab znajdujących się w różnorodnych testach), zwróciły uwagę na indywidualne różnice w percepcji akcentu. Wyniki pracy wykazały, że rodzimi użytkownicy języka, analizujący słuchem strukturę akcentową wypowiedzi w języku polskim (Steffen-Batóg 1990), percypują z reguły mniej końcowych granic zestrojów akcentowych, aniżeli akcentów głównych. Różnicę tę wykorzystano obliczając, na podstawie wyników odsłuchów dwudziestoosobowego zespołu, wskaźnik struktury wyrażający stosunek łącznej sumy identyfikacji końcowych granic zestrojów akcentowych do łącznej liczby identyfikacji akcentów głównych FBSG/MS (final boundaries of stress groups/main stress). Wykazano, że wskaźnik ten jest istotnie różny dla poszczególnych odmian polszczyzny mówionej.

Helfrich (1985) stwierdziła, że dla rozumienia mowy przebieg parametru F0 odgrywa centralną rolę w podziale wypowiedzi na syntaktyczne jednostki znaczące. Odsłuchy 3 tekstów: jednego naturalnego, drugiego zmodyfikowanego (z nałożonym przebiegiem parametru F0 w sprzeczności z granicami syntaktycznymi) oraz trzeciego (z przebiegami częstotliwości podstawowej, naturalnymi, ale nałożonymi na zmodyfikowany, niegramatyczny tekst) wykazały, że aby lokalne zmiany częstotliwości podstawowej określić jako znaczące dla podziału wypowiedzi na frazy, potrzebna jest ocena wypowiedzi o długości 1 - 2 s i zapamiętanie około dwusekundowego przebiegu parametru F0, pozwalające na oszacowanie rozkładu akcentów. Zmiany częstotliwości podstawowej okazały się bardziej efektywne w podziale na frazy niż informacja syntaktyczną. Słuchacz magazynuje w pamięci około 2-sekundowe fragmenty wypowiedzi konieczne do analizy syntaktycznej zdania.

Systematyczne badania cech akustycznych wykorzystywanych w podziale wypowiedzi na frazy podjęli Bruce et al. (1991). Jako istotne uznali nie tylko cechy rozdzielające wypowiedź (demarcative boundary signals), ale też cechy spójności frazy (connective signals). Do cech tych należą:

a. stopniowe obniżanie się przebiegu parametru F0 (connective downstepping),

b. wydłużanie segmentów fonetycznych na granicy frazy (boundary lenghtening),

c. dodatkowy akcent na pierwszym wyrazie po granicy frazowej,

d. duży końcowy spadek częstotliwości podstawowej (final fall).

Odsłuchowe badania wykazały, że obniżanie się konturu intonacyjnego (downtrend), szczególnie jego przerwanie, jest krytyczne dla wrażenia frazowania (grupowania wyrazów), ale nieistotne w percepcji różnic w akcentuacji. Często uwzględnianym czynnikiem, ułatwiającym percepcyjną segmentację wypowiedzi, jest cisza występująca po granicy frazowej.

House (1995) postawił trzy związane z tym zjawiskiem pytania: czy cisza wpływa na percepcję granicy frazowej, czy końcowe sonoranty we frazie niosą istotną percepcyjną informację oraz czy ważność percepcyjną tonalnego końcowego fragmentu przed pauzą wzrasta proporcjonalnie do długości pauzy. W obrębie wypowiedzi syntetycznej typu *amant…ama* w pierwszej wersji eksperymentu — nie umieszczono pauzy, w drugiej wersji eksperymentu umieszczono pauzy długości: 100 ms oraz 1000 ms. Na drugiej sylabie modelowano granicę frazową poprzez spadki częstotliwości podstawowej co 10 Hz w zakresie 140 Hz - 160 Hz. Kiedy między frazami nie było pauzy, tylko połowa słuchaczy zauważała granicę. Obecność pauzy (zarówno krótkiej, jak i długiej) zdecydowanie ułatwiała percepcję granicy frazowej. Istotny również dla słuchowego odbioru ostatniej sylaby przed granicą frazową okazał się udział sonorantu.

Gussenhoven et al. (1992) zauważył, że słuchacze angielscy spodziewali się znaczniejszego wydłużenia segmentów fonetycznych przed granicą, jeżeli ranga granicy frazowej była wyższa.

Stangert (1997) badała wpływ dwóch czynników — struktury i długości wyrazu końcowego na czas trwania sylaby końcowej oraz występującej po niej ciszy. Interwał ciszy był o 68 ms dłuższy w zdaniach z 4-sylabowym wyrazem końcowym niż w zdaniu z 2-sylabowym wyrazem końcowym. Prostsza struktura końcowego wyrazu skróciła czas trwania występującej po nim ciszy.

Systematyczne badania wydłużenia końcowego wyrazu we frazie, zależnie od miejsca wystąpienia akcentu, przeprowadziła Berkovits (1993, 1994). Na podstawie 24 zdań przeczytanych przez 7 mówców analizowała wydłużenie kluczowego wyrazu w pozycji końcowej i niekońcowej we frazie oraz wpływ kontrastywnego akcentu na to wydłużenie. Wyraz na końcu wypowiedzi podlegał w 44% wydłużeniu, a jego zaakcentowanie powodowało dodatkowe wydłużenie sylaby końcowej o 17%. Wyniki wykazały, że zjawisko wydłużania sylaby na końcu frazy występuje niezależnie od akcentu, sylaby nieakcentowane są również wydłużane, przy czym główny efekt obserwuje się na samogłoskach.

De Pijper i Sanderman (1993, 1994) przeprowadzili szczegółowe badania wpływu zakresu zmian parametru F0 oraz konturu na percepcję siły granicy frazowej — PBS (perceived boundary strength). Doświadczenia powtórzono również na materiale zdeleksykalizowanym.

Wyniki wykazały wysoką zgodność w odpowiedziach osób badanych. Słuchacze mogą więc, pomimo braku informacji syntaktycznej, percypować granice frazowe. Siłę granicy frazowej opisano równaniem 4.11.

$$PBS = P + M + R \quad (4.11)$$

gdzie: P — pauza (podzielona na 6 kategorii),
M — typ melodycznego konturu — (7 typów),
R — reset — załamanie linii deklinacyjnej (3 kategorie — brak, reset w górę i reset w dół).

W późniejszych pracach Sanderman i Collier (1996) na podstawie syntezy 8 zdań holenderskich wyznaczyli siłę granic frazowych (PBS) według 5 typów kategoryzacji. Podobne doświadczenia weryfikujące istotność zakresu oraz typu zmian parametru F0, występujących na końcu wypowiedzi przeprowadzili Swerts et al. (1994) i Swerts (1997).

Szereg opracowań dotyczących automatycznego rozpoznawania kategorii suprasegmentalnych wyłącznie na podstawie informacji językowej powstało w ostatnich latach głównie na potrzeby syntezy text-to-speech. Altenberg (1987) stwierdził, że akcent może być przewidziany na podstawie analizy leksykalnej i wyodrębnienia ze słownika wyrazów tzw. pomocniczych oraz wyrazów potencjalnie niosących akcent (wyrazów treściowych i wyrazów funkcjonalnych). Poprzez analizę wielopoziomowej hierarchii rozpoznał w 57% poprawnie miejsce wystąpienia akcentu. Hirschberg (1995) na podstawie analizy drzewa decyzyjnego trenowanego automatycznie opracowała zbiór reguł z informacji zawartej w tekście. Uzyskana dokładność klasyfikacji akcentu wynosiła 77-85%. Ross, Ostendorf (1996) na potrzeby syntezy przedstawili model wykorzystujący procesy Markowa, przewidujący umiejscowienie akcentu wyłącznie na podstawie informacji z tekstu. Modelowanie akcentu przeprowadzono na poziomie sylaby. Doświadczenia odsłuchowe wykazały w 85% zgodność wystąpienia modelowanych oraz percepcyjnie rozpoznanych przez słuchaczy akcentów. Wang i Hirschberg (1992) badali możliwość wykorzystania informacji z tekstu zawierającego 300 wypowiedzi. Zastosowana technika CART (Classification And Regression Tree) wykazała w 90% poprawne modelowanie struktur suprasegmentalnych.

4.5. CECHY SUPRASEGMENTALNE MOWY SPONTANICZNEJ

Nieliczne, przeprowadzane ostatnio na świecie badania dotyczące cech suprasegmentalnych mowy spontanicznej mają fragmentaryczny, pilotażowy charakter, a ich wyniki nie tworzą podstawy do uogólnień (Sagisaka et al. 1997). Jako wstępne zagadnienie rozważa się różnice między cechami melodycznymi mowy spontanicznej oraz czytanej. Bruce (1995, s. 34) na podstawie analizy 13-minutowej konwersacji oraz wersji czytanej tej konwersacji stwierdził, że tekst czytany charakteryzuje się stereotypami w zakresie intonacji. Jako cechę charakterystyczną dla mowy spontanicznej określa wzrost lokalnego zakresu parametru F0 na wybranej jednostce słownej w celu zwiększenia uwagi słuchacza. Mowa spontaniczna wykazuje przeciętnie większe wartości maksymalne częstotliwości podstawowej (wartości średnie i minimalne parametru F0 w mowie spontanicznej i czytanej są podobne). To stwierdzenie autor poparł obszernymi eksperymentami (Bruce 1995).

Hirschberg (1995) zwróciła uwagę na generalną tendencję w mowie spontanicznej do wymawiania wypowiedzi oznajmujących z intonacją rosnącą. W mowie czytanej z opadającą intonacją wymówione jest zwykle 84% wszystkich wypowiedzi, w spontanicznej tylko 70,5%. Druga istotna różnica dla mowy spontanicznej i tekstów czytanych, to większa szybkość mówienia dla tekstów czytanych. Stosunek szybkości (w sylabach na sekundę) dla 17 mówców mieści się w zakresie 0,93- 1,57.

Fujisaki (1997) analizował różnice w zakresie lingwistycznych aspektów mowy spontanicznej i czytanej. W mowie spontanicznej częstym zjawiskiem jest między innymi zmiana porządku wypowiedzi — ważne wyrazy umieszczane są na początku wypowiedzi, istnieje tendencja do powtarzania fragmentów wypowiedzi, popełniania błędów językowych. Nieistotne fragmenty wypowiedzi są wymawiane szybko, ze zredukowaną artykulacją, mogą być całkowicie lub częściowo pozbawione akcentuacji.

Analiza suprasegmentaliów mowy spontanicznej jest niezbędna dla praktycznych zastosowań komputerowych systemów dialogowych. Głównie w tym kierunku należy prowadzić dalsze badania i zmierzać do wykrycia uniwersalnych mechanizmów sterujących prozodią mowy, niezależnie od specyficznych cech określonego języka.

5 MODELE I OPISY INTONACJI W SYSTEMACH DIALOGOWYCH

5.1. OGÓLNE TENDENCJE

Pierwsze prace poświęcone opisom intonacji miały przede wszystkim charakter podręczników (np. Palmer 1922, Klinghardt 1925 i von Essen 1956), w których poprzez impresjonistyczne transkrypcje usiłowano zilustrować w postaci graficznej zmiany wysokości tonu. Obecnie istnieje co najmniej kilka różnych modeli intonacji opartych na kryteriach fonetyczno-akustycznych. Aktualne nadal pozostają jednak zapoczątkowane w latach 30. niektóre kierunki badawcze związane ze sposobem modelowania zmian wysokości tonu. Wyróżnia się zazwyczaj dwa sposoby analizy — globalny i lokalny. **Globalny** opis jest próbą scharakteryzowania konturu jako całościowej struktury, która rozpościera się na całe zdanie (np. Armstrong & Ward 1926 oraz Jones 1956). Angielska intonacja daje się opisać jako dwa podstawowe wzorce (tunes) z licznymi wariantami.

W przeciwieństwie do wyżej wymienionej koncepcji całościowych wzorców Palmer (1933) opisał intonację w terminach **lokalnej** zmiany (nuclear pitch movement) skojarzonej z sylabą akcentowaną i stanowiącej centrum tzw. jednostki tonicznej, która opcjonalnie mogła jeszcze zawierać początek (prehead) i rozciągnięcie (tail).

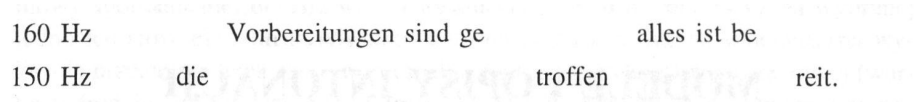
W latach 70. podjęto pierwsze próby modelowania intonacji ukierunkowane na algorytmizację i implementacje praktyczne.

Delattre et al. (1965, s. 135) podzielił przebieg intonacyjny według grup znaczeniowych:

John left Henry (A)	running fast (B)	to find out who had come (C)
minor continuation	major continuation	finality

W wyniku statystycznej oceny otrzymano 5 dystynktywnych wzorców, określających grupy znaczeniowe: *tail, back, neck, head, beak*. Niejasne są jednak kryteria podziału na poszczególne elementy. Eksperymenty prowadzone przez Delattre'a nie uwzględniały percepcyjnego aspektu intonacji.

Na podstawie badań odsłuchowych, mających na celu kategoryzację bodźców na 4 kategorie: wypowiedź, pytanie, kontrast lub kontynuację, Isăcenko i Schädlich (1966, s. 19) zastosowali 2-stopniowy opis intonacji, który ilustruje poniższy przykład:



Halliday (1967) podjął próbę kategoryzacji podstawowych typów intonacyjnych, według której każda wypowiedź może być podzielona na serię grup fonicznych. Każda toniczna grupa składa się z jednej lub więcej stóp rytmicznych i może zawierać aż do 7 sylab. Najważniejsza sylaba oznaczona jest jako toniczna. Toniczna grupa podzielona jest na opcjonalne pretoniczne sylaby poprzedzające te toniczne oraz na obligatoryjne toniczne. Najważniejsza jest grupa toniczna niosąca dystynktywną informację o typie przebiegu.

Wyróżniono następujące znaczące typy przebiegów:

a. spadek — oznaczający stwierdzenie,

b. wzrost lub spadek oraz wzrost — pytanie,

c. przebieg równy — słabe stwierdzenie,

d. wzrost-spadek-wzrost — wypowiedź z rezerwą,

e. spadek-wzrost-spadek — wypowiedź z emfazą.

Tak prosta zależność między typem intonacyjnym i typem zdania nie ma w świetle licznych późniejszych badań uzasadnienia.

Problem segmentacji konturu intonacyjnego na odrębne jednostki opisowe podjęli między innymi Crystal (1969), Altenberg (1987), Isăcenko i Schädlich (1966).

Istotne dla modelowania przebiegu częstotliwości podstawowej 4 fragmenty wypowiedzi ilustruje następujący przykład (według Crystala 1969, s. 207):

you	might get	per MIS	sion
prehead	head	nucleus	tail

Segment oznaczony jako „nucleus” stanowi najbardziej prominentną sylabę, „prehead” oznacza jakiekolwiek sylaby przed akcentowanymi, „head” — sylabę akcentowaną przed „nucleusem”.

Wyróżnia się następujące typy przebiegów (Crystal 1969, s. 210):

a. prosty — (opadający, rosnący lub równy),

b. złożony — (opadająco-rosnący lub rosnąco-opadający),

c. sumaryczny — (opadający + rosnący, rosnący + opadający, opadający + równy etc.),

d. kombinacje powyższych tonów.

W najnowszych pracach poświęconych suprasegmentaliom wyłoniły się 2 kierunki: modelowanie hierarchiczne oraz sekwencyjne.

Różnica między **hierarchicznymi** modelami intonacji, w których przebieg parametru F0 interpretowany jest jako złożony wzorzec składający się z superpozycji wielu składowych i **sekwencyjnymi** modelami, zakładającymi przypisanie przebiegom częstotliwości podstawowej dystynktywnych tonów, polega w pierwszym rzędzie na sposobie definiowania relacji między lokalnymi zmianami i globalnymi tendencjami. Koncepcję liniowego, sekwencyjnego oraz hierarchicznego modelu (w pewnym stopniu wykazującego analogie do akustycznej analizy harmonicznej) przyjęto w pracy Ladd (1983) oraz Thorsen (1988). Trzeci punkt widzenia zakładający prymat intonacji nad akcentuacją reprezentuje np. ’t Hart et al. (1990).

Hipoteza, że mówca chce osiągnąć pewien poziom częstotliwości, a wynikowe przesunięcia przebiegu są tylko przejściami między poszczególnymi poziomami, reprezentowana we wcześniejszych pracach (np. Pike 1947, Pierrehumbert 1980, Ladd 1983), jest obecnie powszechnie krytykowana.

Różne położenie zmian parametru F0 na skali częstotliwości, odmienne wartości średnie, indywidualne wysokości głosu, konieczność uwzględniania w akcentacji synchronizacji czasowej położenia maksimum lub minimum przebiegu parametru F0 względem początku samogłoski, świadczy o istnieniu w mowie konfiguracji tonalnych (por. między innymi ’t Hart et al. 1990, Hermes 1995 oraz Kohler 1997).

5.2. TRANSKRYPCJE STRUKTUR MELODYCZNYCH

Opracowana w 1989 roku kolejna wersja transkrypcji IPA (International Phonetic Alphabet, IPA 1989) zawiera następujące kategorie prozodyczne (c — spółgłoska, v — samogłoska):

Symbol	**Kategoria**
brak symbolu	brak akcentu
,CV	akcentowana
'cv	silnie akcentowana — akcent 1
'cv̀	silnie akcentowana — akcent 2
''cv	ekstra silnie akcentowana — akcent 1
''cv̀	ekstra silnie akcentowana — akcent 2
cv | cv	słaba granica frazowa
cv| |cv	silna granica frazowa
cv | | | cv	ekstra silna granica frazowa

Bruce et al. (1995) w ramach projektu KIPROS (Contrastive Interactive Prosody) prowadzonego w latach 88-90 i obejmującego język francuski, grecki i szwedzki zweryfikowali system transkrypcji suprasegmentaliów IPA oraz system ToBI. Wybrano następujące kategorie prozodyczne:

Symbol	**Kategoria**
brak akcentu	brak symbolu
akcent 1	HL*
akcent 2	H*L
uwydatniony akcent 1	HL*H
uwydatniony akcent 2	H*LH
akcent 2 złożony	H *L...L*H

Akcenty 1 i 2 różnią się synchronizacją czasową w obrębie sylaby.

Popularnym systemem transkrypcyjnym struktur melodycznych stał się w ostatnich latach ToBI (Tones and Break Indices), weryfikowany dla różnych języków (np. Grice et al. 1996, Mixdorf et al. 1997). System ten wykorzystuje jednotonalne zmiany:

H* — akcent szczytowy — *(peak accent)* na akcentowanej sylabie — odpowiadający maksimum,

L* — niski akcent *(low accent)* — odpowiadający często minimum przebiegu na akcentowanej sylabie.

Także zmiany dwutonalne:

L* + H — *(scooped accent)* akcent szczytowy — maksimum przebiegu występuje poza akcentowaną sylabą (późny wzrost),

L + H* — akcent ze wzrostem *(rising peak accent)* — wzrost parametru F0 następuje na sylabie akcentowanej (wczesny wzrost),

H + L* — (*step down to low*) wysoki lub średni ton na przedakcentowej sylabie z następującym spadkiem,

H + !H* — *(step down)* wczesne maksimum, sylaba akcentowana niska (wczesny spadek).

Na końcu frazy mogą wystąpić:

L_ — niski ton (opadający przebieg) usytuowany w pobliżu końca akcentowanego wyrazu,

H_ — bez zmiany tonu — na tym samym poziomie.

Campione et al. (1997) opracowali system kodowania przebiegów częstotliwości podstawowej INTSINT dla języka francuskiego oraz włoskiego oparty na cechach akustycznych sygnału. INTSINT (INTernational Transcription System for INTonation) zakłada możliwość opisu przebiegów częstotliwości podstawowej przy wykorzystaniu globalnych oraz lokalnych cech akustycznych wypowiedzi. Każdy wzorzec intonacyjny może być reprezentowany jako sekwencja tonalnych segmentów. Segmenty te mogą być interpretowane globalnie oraz lokalnie.

Segmenty definiowane globalnie odniesione są do indywidualnego zakresu zmian parametru F0:

a. linia górna — Top (T),

b. linia środkowa — Mid (M),

c. linia dolna — Bottom (B).

Segmenty definiowane lokalnie uwzględniają poprzedzające zmiany parametru F0 (na poprzednich sylabach):

a. wyższy — Higher (H): „a peak” ,

b. niższy — Lower (L): „a trough” ,

c. taki sam — Same (S): „an F0 plateau”,

d. rosnący w górę — Upstep (U): „a raised plateau or a smaller peak than H”,

e. opadający na dół — Downstepped (D): „a lowered plateau or a smaller through than L”.

Audytywne transkrypcje są często wykorzystywane w badaniach percepcyjnych struktur melodycznych i przydatne w układach automatycznej klasyfikacji i rozpoznawania struktur suprasegmentalnych. Określenie intonacyjnych kategorii nawet dla doświadczonych specjalistów z zakresu fonetyki akustycznej jest zadaniem trudnym (zwłaszcza w przypadku percepcyjnej etykietyzacji mowy spontanicznej).

5.3. OPISY DEKLINACJI

Z teoretycznego i praktycznego punktu widzenia zjawisko deklinacji należy do najbardziej niejasnych pojęć fonetyki. Ogólnie deklinację określa się jako występującą w głosie tendencję do rozpoczynania wypowiedzi średnim tonem i stopniowego obniżania głosu w trakcie wypowiedzi. Już w pierwszych pracach z zakresu fonetyki doświadczalnej Armstrong i Ward (1926) oraz później Lieberman (1967) zwrócili uwagę na fakt, że kontur melodyczny zdania składa się z 2 części: niekońcowej i końcowej, każdej o różnym nachyleniu, przy czym terminalna część przebiegu parametru F0 skojarzona jest z szybkim wzrostem lub spadkiem, nie- terminalna zaś jest płaska i może uwidaczniać wpływy akcentów. Lieberman (1967) w opisie tego zjawiska wykorzystał teorię tzw. grupy oddechowej, według której przy końcu wypowiedzi zmniejsza się oddech oraz ciśnienie podgłośniowe wpływające na obniżenie intensywności sygnału i częstotliwości podstawowej. Kontur zmian parametru F0 w wypowiedziach oznajmujących daje się opisać globalnie jako opadający przebieg z lokalnymi zmianami parametru F0 (Maeda 1976, Pierrehumbert 1979, Cruttenden 1986). Rekonstrukcję tej globalnej tendencji można otrzymać poprzez aproksymację linii łączącej maksima, minima przebiegu lub punkty środkowe (por. np. Cooper i Sorensen 1977, 1981, Huber 1989). Pierrehumbert (1979) zauważyła istnienie zjawiska deklinacji na podstawie badań odsłuchowych. Stwierdziła, że

dwie akcentowane sylaby odbierane są jako równie wysokie, jeżeli częstotliwość podstawowa drugiej sylaby jest około 17 Hz niższa od częstotliwości pierwszej akcentowanej sylaby. Liebermann et al. (1984) zaprezentowali odmienne stanowisko, uznając, że główne czynniki kształtujące kontur są lokalne, natomiast deklinacja traktowana jest jako połączenie stopniowo obniżających się zmian parametru F0 na kolejnych akcentach. Zjawisko deklinacji związane jest w dużej mierze z rodzajem badanego materiału. 35% wypowiedzi spontanicznych nie wykazywało deklinacji, dla 45% wypowiedzi bardziej odpowiednia okazała się teoria grupy oddechowej umożliwiająca opis przez liniową regresję początkowego i środkowego odcinka konturu intonacyjnego i dodatkowo, oddzielnie opadającego, końcowego fragmentu przebiegu parametru F0 (Lieberman et al. 1985). Zależność występowania deklinacji od rodzaju badanego materiału potwierdzają również inni badacze. Między innymi Umeda (1982) stwierdziła, że zdania przeczytane w izolacji wykazywały deklinację, natomiast w tekście ciągłym efekt deklinacji nie był zauważalny. Jedna interpretacja tego faktu zakłada, że zdanie w kontekście nie ma deklinacji, a końcowy spadek parametru jest oddzielnym fenomenem. Drugie wyjaśnienie przyjmuje, że w przypadku wypowiedzi zdania w izolacji mówca bierze pod uwagę koniec zdania i ma możliwość przygotowania się do niego.

Zagadnienie modelowania nie w pełni precyzyjnie określonego zjawiska, jakim jest deklinacja, jest więc kontrowersyjne. Najobszerniej, głównie dla celów syntezy mowy, modelowanie deklinacji dla tekstów czytanych opracowano dla języka holenderskiego (’t Hart et al. 1990) i japońskiego (Fujisaki 1981, 1983, 1997).

W zdaniach krótszych lub dłuższych od 5 s przyjęto następujące eksperymentalne zależności (5.1) określające deklinację *(DK)* w półtonach na sekundę (półton/s):

$$\begin{aligned} t \le 5s \quad & DK = (-11)/(t + 1{,}5) \\ t > 5s \quad & DK = (-8{,}5)/t \end{aligned} \tag{5.1}$$

Fujisaki (1981) do modelowania deklinacji proponuje funkcję ekspotencjalną Gpi(t) i uzasadnia jej wybór uwarunkowaniami fizjologicznymi (zależność 5.2).

$$G_{pi}(t) = \alpha^2 \, t \exp(-\alpha_i t), \qquad \text{dla } t > 0 \tag{5.2}$$

gdzie: α — współczynnik tłumienia,
t — czas.

Istnieje szereg problemów metodologicznych związanych z formalizacją opisu deklinacji, między innymi:

1. Kontrowersyjny jest wybór punktów, które określać będą deklinację: maksima przebiegu odzwierciedlają wpływy akcentów, minima, z uwagi na często występującą na końcu wypowiedzi laryngalizację, są trudne do pomiaru.

2. Dyskusyjny jest wybór funkcji opisującej deklinację: prostoliniowej, nieliniowej (wypukłej czy wklęsłej).

3. Trudne do ustalenia jest kryterium dopasowania funkcji do wybranych punktów (np. średniokwadratowe, maksimum błędu).

Z punktu widzenia analiz akustycznych spotyka się więc w literaturze przedmiotu wiele różnych metod modelowania deklinacji. Na podstawie tekstów złożonych ze 176 zdań, czytanych przez 4 osoby, Huber (1989) przeprowadził liniową regresję między maksimami i minimami konturu intonacyjnego. Analiza współczynnika korelacji (między danymi eksperymentalnymi i danymi otrzymanymi w wyniku regresji) wykazała w 86% zgodność intonacyjnych jednostek z granicami syntaktycznymi. Istnieje przekonanie, że deklinacja spełnia ważną funkcję w syntaktycznym rozczłonkowaniu wypowiedzi (por. także Fujisaki 1981).

Dotychczas przeważająca większość opracowań dotyczyła analiz zdań izolowanych lub czytanych tekstów, które charakteryzują się pewnymi stereotypami nabytymi podczas nauki czytania. Można przypuszczać, że z tego to powodu w większości prac zakłada się istnienie deklinacji — i konieczność jej modelowania. Ostatnio przeprowadzone badania na zróżnicowanym materiale językowym, jakim jest mowa spontaniczna, pozwalają uznać zjawisko deklinacji jako bardziej wątpliwe.

Aby określić zjawisko deklinacji 't Hart et al. (1990), wyznaczyli histogramy z wartości parametru F0 (odczytywanych z przebiegu co 10 ms) z początkowych, środkowych i końcowych fragmentów wypowiedzi spontanicznych. W wyniku ich analizy oszacowano średni spadek parametru F0 ilustrujący deklinację jako 0,3 pół- tonu/s z odchyleniem standardowym 2,05 półtonu/s. W wypowiedziach czytanych (35 zdań o długości 0,6 - 6,3 s), średni spadek wartości parametru F0 wynosił 0,51 półtonu/s z odchyleniem standardowym 0,35 półtonu/s. Znaczną różnicę w parametrach rozkładu spadków częstotliwości podstawowej — w wariancjach uzyskanych dla mowy spontanicznej i czytanej — autorzy tłumaczą łatwiejszym sterowaniem iloczasem w tekstach czytanych o przewidywalnej długości wypowiedzi.

Należy stwierdzić, że wszystkie proponowane sposoby modelowania deklinacji mogą wystarczająco poprawnie odzwierciedlać pewien ogólny trend polegający na obniżaniu zakresu zmian parametru F0 w trakcie neutralnej, czytanej wypowiedzi. Opisy deklinacji opracowane między innymi przez 't Harta et al. (1990) oraz Fujisaki (1981) dotyczą głównie tekstów czytanych.

Problematyczna wydaje się więc uniwersalność modelowania deklinacji np. w mowie spontanicznej. Między innymi Kohler (1997) na podstawie analizy mowy spontanicznej stwierdził, że zjawiska deklinacji nie zaobserwowano w sposób regularny.

5.4. MODELE INTONACJI

5.4.1. MODELE SUPERPOZYCYJNE

Pierwszy model zawierający formalny opis zmian częstotliwości podstawowej przedstawił Öhman (1967). Przebieg parametru F0 jest syntetyzowany w mechanizmie krtaniowym (larynx model), który posiada 3 wejścia scharakteryzowane poprzez: a) zmienne napięcie wiązadeł głosowych, które jest sumą dwóch składowych reprezentujących intonację zdaniową i intonacją wyrazową, b) akustyczną interakcję sygnału wywołaną drugorzędnymi efektami ciśnienia pod i ponadgłośniowego oraz c) niefonacyjnymi czynnikami artykulacyjnymi. Składowe reprezentujące intonację zdaniową i intonację wyrazową są wyjściami dwóch różnych filtrów reprezentujących dynamiczne charakterystyki systemu krtaniowego. Wejścia filtrów sterowane są funkcjami schodkowymi. Model przetestowano na wyrazach języka szwedzkiego z dwoma typami akcentu (akutowym i grawisowym).

Model intonacji, opracowany przez Fujisaki (w latach 1981, 1983), wielokrotnie modyfikowany (1988, 1997), bazujący na koncepcji Öhmana, został zrealizowany w postaci liniowego systemu na zasadzie superpozycji. Wyjściowe sygnały mechanizmu sterującego składową frazową oraz akcentową dodają się do stałej wartości Fmin charakterystycznej dla danego mówcy.

Fujisaki podaje dwie definicje Fmin. Z pierwszej definicji ukierunkowanej artykulacyjno-fizjologicznie wynika, że Fmin jest zależne od mówcy, ale niezależne od wypowiedzi (Fmin jest najniższą wartością możliwą do osiągnięcia przez określonego mówcę). Druga definicja jest zorientowana fizyczno-matematycznie, ponie-

waż opiera się na cechach sygnału — wyekstrahowanych wartościach parametru F0. Druga definicja uwzględnia więc możliwość zależności Fmin od konkretnej wypowiedzi.

Funkcje systemu sterującego (Fujisaki 1988, s. 348) można opisać następującą zależnością (5.3):

$$\ln F_0 = \ln F_{min} + \sum_{i=1}^{I} K_{pi} G_{pi}(t - T_{0i}) + \sum_{j=1}^{J} K_{aj}[G_{aj}(t - T_{1j}) - G_{aj}(t - T_{2j})]$$

$$G_{pi} = \alpha_i t \exp(-\alpha_i t) \quad t > 0$$
$$0 \quad t \le 0$$

$$G_{aj}(t) = 1 - (1 + \beta_j t) \exp(-\beta_j t) \quad t > 0$$
$$0 \quad t \le 0 \qquad (5.3)$$

gdzie: F_{min} — asymptotyczna wartość częstotliwości podstawowej,
I — liczba frazowych rozkazów,
J — liczba akcentowych rozkazów,
$G_{pi}(t)$ — funkcja sterująca frazą,
α_i — współczynnik tłumienia mechanizmu sterującego frazą,
K_{pi} — określa amplitudę i-tej składowej frazowej,
K_{aj} — określa amplitudę j-tej składowej akcentowej
T_{0i} — punkt czasowy rozpoczęcia komendy frazowej,
$G_{aj}(t)$ — funkcja sterująca akcentem,
β_j — współczynnik tłumienia, mechanizmu sterującego akcentem,
T_{1j} — czasowy punkt początku komendy akcentu,
T_{2j} — czasowy punkt końca rozkazu.

Po wygładzeniu każdy przebieg parametru F0 aproksymujący akcent wzrasta od zera asymptotycznie do określonego poziomu i opada asymptotycznie. W praktycznym zastosowaniu modelu Fujisaki (np. Demenko et al. 1990, Möbius et al. 1991a i b, Mobius 1993) pojawiają się między innymi problemy uzyskania dostatecznie dokładnej aproksymacji na końcu wypowiedzi (konieczne jest wprowadzenie ujemnej wartości frazowego rozkazu). Model Fujisaki stosowano dla różnych języków, między innymi chińskiego, hiszpańskiego i niemieckiego (np. Mobius 1993, Fujisaki 1997). W niektórych implementacjach modelu (np. dla języka chińskiego) odpowiednio dokładna aproksymacja konturu melodycznego osiągnięta została przez wykorzystanie dużej liczby funkcji sterujących frazą i akcentem.

Dla języka duńskiego Thorsen (1978, 1988) sformułowała model intonacji koncepcyjnie podobny do rozwiązań superpozycyjnych. Kontur intonacyjny określony jest czterema składowymi. Na pierwszym poziomie priorytet posiada tzw. *tekst-kontur*. Na przebieg częstotliwości określony konturem tekstowym mają wpływ następne składowe: kontur zdaniowy oraz kontur frazowy. Czwartą składową reprezentuje wzorzec opisujący grupę akcentową, której cechy wynikają między innymi z pozycji w zdaniu. Wszystkie składowe są interaktywne i zależne od mówcy. Podobny superpozycyjny model zastosowała Gardning (1983) dla języka szwedzkiego.

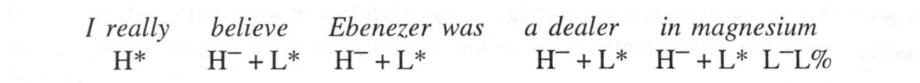
5.4.2. MODELE SEKWENCYJNE

Pierrehumbert (1983) przedstawiła zmiany intonacji jako ciąg niskich tonów L oraz wysokich tonów H w postaci 3 klas:

a) we frazie jest przynajmniej jeden akcent tonalny, który jest zrealizowany albo przez pojedyncze tony H*, L*, albo kombinację 2 tonów: L— + H*, H— + L*, H* + L—, L* + H—, H* + H—,

b) frazowy akcent na końcu wypowiedzi jest typu H— lub L—,
c) graniczny ton na końcowej sylabie frazy jest typu H% lub L%.

Poniższy przykład ilustruje zapis intonacji wypowiedzi (według Pierrehumbert 1983).



Fonetyczne reguły ustalają wartości H— lub L— zależnie od linii bazowej, powyżej której mówca realizuje uwydatnienia poszczególnych sylab. Hipotetyczna linia bazowa umożliwia określenia położenia wysokich oraz niskich tonów. Podstawowa teza wysunięta przez Pierrehumbert (1980) zakłada, że intonacja jest wyłącznie lokalnie zdeterminowana, prezentowany model ma więc typowo sekwencyjny charakter (Pierrehumbert 1983). Model Pierrehumbert w większym lub mniejszym stopniu modyfikowano w różnych opracowaniach. Np. Ladd et al. (1994) uzupełnili ten model o zjawisko „downstepping" — obniżania się konturu intonacyjnego.

Intonacyjny model KIM (Kieler Intonationsmodell) oparty na brytyjskich tradycjach w badaniach intonacji (tonach Hallidaya) opracował dla języka niemieckiego Kohler (1991, 1995). Atrakcyjność modelu polega na uwzględnieniu wpływów segmentalnych i synchronizacji maksimów oraz minimów przebiegu parametru F0 z akcentowanymi samogłoskami. Model obejmuje:

a. akcent leksykalny (3 poziomy: brak akcentu, drugorzędny i główny) oraz akcent zdaniowy (4 poziomy),

b. intonację stanowiącą połączenia między maksimami i minimami, synchronizację ekstremów z samogłoskami akcentowanymi (wczesna, środkowa, późna),

c. granice prozodyczne,

d. tempo mowy,

e. nieciągłości i pauzy.

W modelu uwzględniono 3 położenia maksimów — wczesne, środkowe i późne oraz nadano im funkcje semantyczne. Wczesne oznacza fakt ustalony, środkowe świadczy o pojawieniu się nowej informacji, zaś późne związane jest z emfazą.

5.4.3. MODELE Z CECHAMI SUPERPOZYCJI ORAZ SEKWENCJI

Podstawowym modelem charakteryzującym się cechami sekwencyjności oraz superpozycją jest model 't Harta et al. (1990). Model ten, wielokrotnie modyfikowany, wykorzystywano dla różnych języków. Poniżej przedstawiono najważniejsze założenia tworzenia modelu.

- 1. Słuchacz wrażliwy jest tylko na te zmiany częstotliwości podstawowej, które były zamierzone przez mówcę. Percepcyjnie adekwatna jest stylizacja przebiegu bez gwałtownych przeskoków i aproksymacja zmian wysokości tonu w skali logarytmicznej.
    2. Na pierwszym poziomie opisu najmniejszą jednostką percepcyjnej analizy jest zmiana parametru F0 (pitch movement). Zmiana ta może być opisana następującymi percepcyjnie istotnymi cechami.

	1	2	3	4	5	A	B	C	D	E
kierunek										
wzrost	X	X	X	X	X					
spadek						X	X	X	X	X
synchronizacja										
wczesna	X				X		X			X
późna			X			X				
bardzo późna		X						X		
prędkość zmiany										
szybka		X	X	X		X	X	X		X
wolna				X					X	
wielkość										
pełna	X	X	X	X		X	X	X		
częściowa					X					X

Kategorie umieszczone w tabeli mają interpretację fonetyczną. Np. kategoria 1 (szybki wczesny wzrost) oznacza wzrost 50 półtonów/s z czasem trwania 120 ms i taką synchronizacją czasową, że maksimum jest osiągnięte po 50 ms początku wokalicznej części sylaby. Maksymalna liczba kategorii jest ograniczona. Jeżeli 2 wzrosty przebiegu są krótsze niż 250 ms (najczęstszy przypadek) i pokrywają interwał większy niż 3,5 półtonu to różnice w zmianach ich prędkości pozostają poniżej progu percepcji. Tylko stopniowe wzrosty, które rozciągają się na kilka sylab mogą się percepcyjnie różnić w prędkości zmian.

Gramatyka intonacji określa możliwości połączeń zmian parametru F0. Jej najważniejsze reguły są następujące:

- 1. Nie istnieją poziomy intonacyjne tylko konfiguracje.
  2. Zmiany parametru F0 mogą występować tylko w szczególnych melodycznych połączeniach, np. konfiguracje 1A, 1B stanowią unikatowe sekwencje.
  3. Zmiany parametru F0 należą do jednej z 3 klas paradygmatycznych: prefiksu — opcjonalnie rekursywnego, rdzenia — obligatoryjnego i nierekursywnego oraz sufiksu — opcjonalnego i nierekursywnego. Różne kontury należą do tego samego wzorca intonacyjnego, jeżeli mają wspólną konfigurację rdzenną, nawet jeśli posiadają odmienne segmenty wchodzące do klasy prefiksu i sufiksu.
  4. Konfiguracje łączą się w kontury zgodnie z uwarunkowaniami syntagmatycznymi.

Nieograniczona liczba różnych konturów jest przejawem skończonej liczby podstawowych wzorców intonacyjnych. Możliwe kontury tworzą różne melodyczne struktury, zależnie od konfiguracji rdzennych (’t Hart et al. 1990).

Wszystkie relewantne cechy zmiany parametru F0 są kontrolowane w systemie wytwarzania tonu (poprzez mięśnie cricothyroideus CT). Intonacja dominuje nad akcentuacją w procesie kształtowania konturu intonacyjnego. Niektóre zmiany tonu są skojarzone z akcentowanymi sylabami (np. pełny, szybki, bardzo późny wzrost połączony z pełnym, szybkim, późnym spadkiem — 1 A) inne nigdy (np. pełny, szybki, późny wzrost połączony z pełnym, szybkim, wczesnym spadkiem — 2 B). Jak widać omawiany model uwzględnia odwzorowanie czynników fizjologicznych i w związku z tym na potrzeby realizacji automatycznej rozważono dwie hipotezy.

- 1. Zmiany tonu, które się pojawiają na akcentowanych sylabach są całkowicie spowodowane przez instrukcję akcentu, pozostałe zmiany wynikają z cech intonacji. Zmiany częstotliwości kolejno są realizowane

jako polecenie akcentu albo przez instrukcje intonacyjne. Taki pogląd reprezentują między innymi Ladefoged (1975) oraz Thorsen (1978). Stwierdzają oni, że cały kontur melodyczny wypowiedzi powstaje przez liniowe dodawanie akcentuacyjnych i intonacyjnych własności.

2. Melodyczne własności są zdeterminowane przez wybrany wzorzec intonacyjny, wszystkie zmiany częstotliwości tworzą melodię.

Odpowiedniość między intonacją a strukturą syntaktyczną nie jest ani obligatoryjna, ani jednoznaczna. Syntaktyczna granica nie musi być zaznaczona intonacyjnie. W przypadku sprzeczności cech czasowych i intonacyjnych segmentu końcowego frazy, zawsze czynniki czasowe mają priorytet przed intonacyjnymi. Intonacyjne cechy mają ograniczony wpływ na semantyczną interpretację zdania. Można zmienić przy pomocy intonacji wypowiedź twierdzącą w pytanie, ale interpretacja semantyczna jest ograniczona. Programowanie konturu intonacyjnego wymaga fragmentarycznej, przewidywalnej (look ahead) strategii, która integruje akcentowe i syntaktyczne informacje. W najprostszym sekwencyjnym modelu TS (Tone Sequence, według Ladda 1983) wybór zmiany tonu dokonywany jest na podstawie informacji skojarzonej z sylabą. W przebiegu konturu uwzględnia się ograniczenia wynikające z gramatyki intonacji.

6 FONETYCZNO-AKUSTYCZNA DEFINICJA AKCENTU I FRAZY INTONACYJNEJ JĘZYKA POLSKIEGO

Rozdział niniejszy stanowi omówienie podstaw teoretycznych opracowania własnego materiału językowego. Wykorzystanie założeń tzw. szkoły brytyjskiej wydawało się szczególnie dogodne dla powiązania akcentu z jego rolą we frazie intonacyjnej. Operowanie tradycyjną jednostką zestroju akcentowego (por. Dłuska 1957, Steffen-Batogowa 1966, 1996) nie było szczególnie dogodne dla przeprowadzenia w dalszej części pracy własnych badań eksperymentalnych. Główną przeszkodą był brak sformułowania dla zestroju akcentowego definicji akustycznej. Przyjęto, że model tzw. szkoły brytyjskiej można dość prosto skorelować z określonymi parametrami akustycznymi.

W ogólnych ramach teoretycznych tzw. szkoły brytyjskiej Jassem (1996a, 1999) przyjął trzy wysokości tonu (oczywiście zrelatywizowane względem indywidualnego zasięgu wysokości tonu w głosie): niski (L), wysoki (H) oraz średni (M) i wykazał za pomocą statystycznej analizy dyskryminacyjnej, iż angielskie tony rdzenne (ośrodkowe; nuclear tones) można zdefiniować względem tych trzech wysokości. W niniejszej pracy zostanie m.in. podjęta próba określenia polskich intonacji rdzennych (ośrodkowych), a więc zarazem akcentu realnego głównego w terminach względnych wysokości L, M i H. Dalszych badań wymaga ewentualność istnienia dystynktywnej wysokości xL, tj. poziomu szczególnie niskiego. Zachodzi możliwość, że ton xL nie jest poziomem dystynktywnym, lecz wynika z parametru zmiennego rejestru. Wysunięto mianowicie hipotezę, że poza zróżnicowaniem dystynktywnych poziomów, takich jak L, M, H może istnieć parametr zakresu oznaczający przesunięcie na skali wysokości wszystkich (albo może niektórych) z wyróżnionych poziomów (por. ostatnio Cruttenden 1997, s. 123- 125). Z pewnością istnieje ogólne obniżenie rejestru we frazach parentetycznych (por. Jassem 1996a, Cruttenden 1997, Ladd 1996).

Niezależnie od faktu nieuzasadnionych merytorycznie komplikacji w systemie autosegmentalno-metrycznym, takich jak osobny ton „krańcowy” (boundary tone) oraz osobny „ton frazowy” (phrase tone) oraz niezwykle (i kontraintuicyjnie) zawiłych reguł koniecznych do interpretacji tonów H, Low H_, !H itd. w tym systemie, zdecydowano w niniejszej pracy, z dwóch obecnie stosowanych systemów opisu intonacji i akcentu na poziomie lingwistycznym (fonologicznym) wybrać system brytyjski, gdyż pozwala on na bezpośrednie odniesienie postulowanych elementów funkcjonalnych, takich jak poziomy H, M, L, rozróżnienie dystynktywnych akcentów realnych itd. do przebiegu parametru F0, co w systemie autosegmentalno-metrycznym wymaga znacznych zawiłości. Mając na uwadze spostrzeżenia, tak Ladda (1996), jak i Cruttendena (1997), co do istnienia ogólnych zbieżności strukturalnych intonacji (uniwersaliów), postanowiono tutaj dokonać analizy intonacji wraz z akcentem realnym polskiego materiału językowego.

Dziedziną akcentu potencjalnego jest leksem. Mamy tu do czynienia z uściśleniem tradycyjnego pojęcia akcentu wyrazowego. Niezależnie od wieloznaczności pojęcia wyrazu (zob. np. Lyons 1992, Bańczerowski et al. 1982) należy przyjąć, iż jest on pojęciem heterogenicznym, w tym także syntaktycznym. Akcent potencjalny jest zaś pojęciem leksykalno-morfologicznym[5]. Natomiast szczególnego znaczenia w tym kontekście definicyjnym nabiera relacja między akcentem realnym a potencjalnym. Już wczesne prace Bolingera (np. 1958) pozwoliły na dopuszcze-

[5] Leksyka lno-morfologiczna definicja akcentu potencjalnego umożliwia jednoznaczna jego klasyfikację z jednej strony na akcent nieruchomy i ruchomy, a z drugiej na stały i zmienny.

nie prymarności akcentu realnego względem potencjalnego. Akcent potencjalny przypada mianowicie na tę sylabę w obrębie leksemu i jego alternantów fleksyjnych, która w realizacji frazy intonacyjnej niesie akcent realny. Tym samym akcent potencjalny jest abstrakcją wyższego rzędu niż akcent realny. Przykładowo, polski leksem *róża* ma akcent potencjalny na przedostatniej sylabie każdego alternantu fleksyjnego, albowiem w przypadku znalezienia się w pozycji iktycznej we frazie intonacyjnej postać mianownika l.poj. będzie (realnie) akcentowana na sylabie [ruj. We frazie intonacyjnej (stanowiącej w tym przypadku zdanie, co jest dla sprawy obojętne) *To jest róża* iktus (początek intonacji rdzennej) przypada na sylabę [ruj. Podobnie, akcent potencjalny np. w narzędniku l.mn. *różami* przypada na sylabę [£a], albowiem w takiej frazie-zdaniu jak *Ucieszyła się tymi różami* iktus przypada na [£a]. Ponieważ niniejsza praca dotyczy zjawisk suprasegmentalnych, akcent potencjalny nie jest przedmiotem analizy. Konieczne jest jednak ustalenie znaczenia, w jakim pojęcie akcentu będzie tutaj używane.

Praca O'Connora i Arnolda (1973), aczkolwiek pomyślana jako podręcznik, ma istotne znaczenie dla problematyki z pogranicza intonacji i akcentu i jest cytowana przez wszystkich teoretyków i praktyków zajmujących się w ostatnim 20-leciu zagadnieniami z tego zakresu. Autorzy podają w pracy najczęściej używane wzorce intonacyjne standardowej angielszczyzny z obszernymi wyjaśnieniami pragmatyczno-semantycznymi, określającymi użycie poszczególnych wzorców. Pomijając szczegóły, można stanowisko autorów ująć następująco. W angielszczyźnie standardowej każda fraza intonacyjna zawiera jeden ton rdzenny („nucleus", albo „nuclear tone"), który może być rosnący, opadający, opadająco-rosnący, rosnąco-opadający lub równy. Wyróżnione są 2 odmiany rdzennej intonacji opadającej: wysoka, którą można przypisać HL (high-to-low) oraz niska ML, (mid-to-low). Istnieje jeden i tylko jeden przypadek dwóch tonów rdzennych w jednej frazie. Wówczas po tonie opadającym wysokim (który należałoby nazwać opadającym pełnym) następuje bezpośrednio ton rdzenny niski rosnący (np. *My 'mother was born in ' Sheffield* (s.30). Jest to wzorzec intonacyjny szczególnie typowy dla języka angielskiego. Przeglądy w Cruttenden (1997) i Ladd (1996) świadczą, że taki wzorzec nie jest spotykany w dotychczas zbadanych innych językach o bogatej intonacji, które dopuszczają we frazie intonacyjnej tylko jedną intonację rdzenną (a zatem jeden ictus, czyli jeden akcent realny główny). Dogodnie byłoby w tym szczególnym przypadku intonacji angielskiej mówić o tonie rdzennym głównym (opadającym) oraz pobocznym (rosnącym), a to z uwagi na fakt, że we wszystkich frazach tego typu maksimum informacyjne (focus) przypada na część opadającą. Ponadto ważnym szczegółem pracy O'Connora i Arnolda jest postulowanie akcentu postiktycznego, tj. występującego po intonacji rdzennej,

np. ˈCan I ˋhelp you at ₒall (s. 55),

gdzie znaczek ₒ sygnalizuje akcent postiktyczny.

Autorzy nie dają w pełni jasnego wyjaśnienia fonetycznych cech tego akcentu postiktycznego, ale zastrzegają się, że nie jest to akcent intonacyjny. Według Jassema (1984) mamy tu do czynienia z akcentem wyłącznie rytmicznym. System Arnolda i O'Connora oraz wymienione wyżej uwagi Cruttendona (1997) i Ladda (1996) co do możliwości podobieństw w zakresie intonacji pomiędzy językami, nasuwają hipotezę, że w polskim języku wyróżnić można również akcent realny rdzenny, którego cechą relewantną jest określony przebieg zmian wysokości tonu w terminach L, M, H oraz akcent postiktyczny, który nie wykazuje takich zmian jak akcent rdzenny. Mielibyśmy zatem do czynienia z głównym akcentem realnym związanym z intonacją rdzenną oraz pobocznym akcentem postiktycznym, nie wykazującym żadnych cech intonacyjnych (lub ewentualnie takie cechy, które nie mają wpływu na przebieg intonacji rdzennej). Mógłby to być np. akcent nacechowany wyłącznie iloczasowo.

Jednakże wszystkie opisy standardowej angielszczyzny brytyjskiej w zakresie intonacji, począwszy od Palmera (1922) wyróżniają oprócz intonacji rdzennej (nucleus, nuclear tone, ictus itp.) także intonacyjny akcent, który ją poprzedza. Brytyjska terminologia używa tu określeń „head" albo „prenuclear tone". Jak wynika z opisów w pracy O'Connora i Arnolda, akcent preiktyczny („head") może być poprzedzony jedną lub kilkoma sylabami nieakcentowanymi. Tę część frazy intonacyjnej określają ci autorzy jako „prehead". Opis struktury frazy intonacyjnej w języku angielskim przyjęty przez Jassema (1984) jest bardzo zbliżony do struktury przyjętej przez O'Connora i Arnolda. Jassem wyróżnia: (a) anakruzę, (b) intonację przedrdzenną (z akcentem preiktycznym), (c) intonację rdzenną (z akcentem iktycznym) oraz (d) akcent nieintonacyjny postiktyczny.

Struktury O'Connora i Arnolda oraz Jassema odniesione do języka angielskiego można więc przedstawić następująco:

O'Connor i Arnold:

[prehead[head [[nucleus] tail]]]

W. Jassem:

[anakruza] [[intonacja przedrdzenna [intonacja rdzenna]]]

akcenty:

akcent poboczny preiktyczny → akcent główny (iktus) → akcent postiktyczny, gdzie: strzałka oznacza kierunek czasu.

W powyższych schematach użyto nawiasów kwadratowych w sensie zbliżonym do pojęcia składników bezpośrednich, co umożliwia reprezentację dendrytową:

Na każdym poziomie gałąź lewostronna jest fakultatywna.

Akcent realny:

(1) Anakruza jest sylabą lub ciągiem sylab początkowych we frazie intonacyjnej pozbawionych akcentu.

(2) Intonacja preiktyczna zawiera jeden lub więcej akcentów preiktycznych.

(3) Intonacja rdzenna zawiera jeden (i tylko jeden) ictus (główny akcent intonacyjny) oraz fakultatywnie postiktyczny akcent nieintonacyjny (jeden lub więcej).

Celem zaznaczenia w tekście miejsca odpowiednich akcentów, bez wskazywania, jakie intonacje są przezeń realizowane, można użyć następujących oznaczeń:

(a) sylaba nieakcentowana: nie oznaczona [brak symbolu],

(b) sylaba akcentowana preiktyczna [ʼ],

(c) sylaba akcentowana iktyczna [ˈˈ],

(d) sylaba akcentowana postiktyczna [o].

Definicje te i pojęcia zastosowano w konstrukcji polskiego materiału językowego. Poniżej przedstawiono 10 wypowiedzi dialogowych stanowiących jedną z części korpusu danych językowych. Przykłady te będą przedmiotem analizy w dalszym ciągu pracy (por. rozdz. 7 oraz 8):

(1) To jest naj' lepsza 'pora ˈˈdnia.

(2) To jest naj' lepsza poˈˈradnia.

(3) To ˈˈnie jest najolepsza pooradnia.

(4) Móˈˈwiłemˈciˈżeˈtoˈjestˈokiepskiˈonawóz.

(5) Móˈˈwiłemˈciˈżebyśˈonieˈkładłˈobelekˈonaˈwóz.

(6) To jest 'bardzo nie 'dobry ˈˈznak.

(7) ˈˈCo móowiłeś ?

(8) To był 'całkiem 'niezły i ucz'ciwy ˈˈczłowiek.

(9) Dla ''miasta.

(10) 'Dwa ''miasta.

Intonacyjne akcenty: iktyczny i preiktyczny mogą być realizowane różnymi konturami intonacyjnymi.

Intonacja rdzenna (akcent iktyczny)

W języku polskim wyróżnić można następujące intonacje rdzenne (tj. intonacje realizujące akcent iktyczny):

(1) Opadająca pełna **HL**

(2) Opadająca niska **ML**

(3) Opadająca wysoka **HM**

(4) Opadająca ekstra niska **xL**

(5) Rosnąca niska **LM**

(6) Rosnąca wysoka **MH**

(7) Rosnąca pełna **LH**

(8) Rosnąco-opadająca **LHL** lub **MHL**

(9) Równa **MM**

Materiał językowy, który zawiera przykładowe intonacje rdzenne jest następujący:

(1) ‘‘Znowu ten owariat. (Intonacja pełna opadająca, **HL**).

(2) 'Bardzo 'zły znak. (Intonacja preiktyczna równa 2 razy + intonacja iktyczna **ML**).

(3) To był 'całkiem 'niezły i ucz'ciwy człowiek. (Anakruza + intonacja preiktyczna rosnąca 3 razy + intonacja rdzenna **ML**).

(4) 'To jest jakiś ''znak. (Intonacja preiktyczna rosnąca + intonacja rdzenna **HL**).

(5) To jest 'bardzo nie' dobry znak. (Intonacja preiktyczna rosnąca 2 razy + intonacja rdzenna **ML**).

(6) Znowu ten owariat?! (Intonacja rdzenna **LH** + akcent postiktyczny).

(7) Co móowiłeś? (Intonacja rdzenna **LH** + akcent postiktyczny).

(8) 'Ja jem. (Akcent preiktyczny z intonacją równą + intonacja rdzenna **ML**).

(9) Jajem. (Intonacja rdzenna **ML**).

(10) To nie jest najolepsza opora odnia. (Anakruza + intonacja rdzenna **ML** + 3 akcenty postiktyczne).

(11) To nie jest najolepsza pooradnia. (Anakruza + intonacja rdzenna **ML** + 2 akcenty postiktyczne).

W załączniku 5 podano przyjęte do analizy frazy wzorcowe.

Wypowiedzi przeanalizowano przy wykorzystaniu przyjętego w obecnej pracy modelu frazy intonacyjnej według schematów O’Connora i Jassema dla języka angielskiego (Jassem 1996a, Demenko, Jassem 1999). Jakkolwiek pewne zjawiska w zakresie intonacji są uniwersalne dla różnych języków (np. emocje objawiające się zwiększeniem interwałów częstotliwości podstawowej), to poszczególne języki wykazują często znaczne różnice zarówno w klasyfikacji typów intonacji rdzennych, jak i sposobie ich wykorzystania. Nawet w grupie języków nietonicznych, takich jak: polski, angielski, francuski, niemiecki istnieją znaczne różnice intonacyjne. Różnice te można określić następująco.

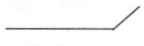
- 1. Różnice strukturalne.

Pewne typy intonacyjne istnieją tylko w określonym języku. Przykładowo intonacja rdzenna rosnąco-opadająca-rosnąca funkcjonująca w języku angielskim, w polskim nie jest wykorzystywana.

- 1. Różnice realizacyjne.

Istnieją w różnych językach podobne intonacje, które są inaczej realizowane. Przykładowo intonacja niska rosnąca LM jest w języku polskim inaczej realizowana niż w angielskim. W języku polskim wzrost częstotliwości podstawowej następuje

z reguły pod koniec frazy, w angielskim może rozpoczynać się od początku frazy, np.



Byłeś tu już?

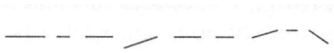
Have you been there?

- 1. Różnice semantyczne.

Intonacje mogą być strukturalnie takie same, ale posiadać całkiem inne znaczenie. Przykładowo intonacja pełna rosnąca LH nie oznacza w języku angielskim pytania. W wypowiedzi

It isn't as bad as all that.

intonacja LH oznacza sprzeciw. W języku polskim wypowiedź ta:



Wcale nie jest znowu taka zła.

zostałaby zrealizowana najprawdopodobniej z intonacją opadającą typu HL.

Tak więc przebiegi intonacyjne języka polskiego będą wykazywać różnice strukturalne, realizacyjne oraz semantyczne w porównaniu z innymi językami.

Dla eksperymentalnej weryfikacji modelu przedstawionego w rozdziałach 7, 8 i 9 przyjęto pewne ogólne założenia. Tzw. naiwni użytkownicy języka potrafią wskazać obecność akcentu. Potrafią także znacznie lepiej imitować (bez wcześniejszego treningu) te przebiegi intonacyjne, które występują w ich rodzimym języku, wreszcie w znacznie wyższym stopniu poprawnie przyporządkowują określone kontury intonacyjne właściwym intencjom mówcy języka rodzimego niż języka obcego.

7 DYSTYNKTYWNE CECHY AKCENTU W JĘZYKU POLSKIM

7.1. PERCEPCYJNO-AKUSTYCZNA OCENA AKCENTU

Percepcyjna analiza intonacji przeprowadzana dla różnych języków (w tym dla języka polskiego por. np. Steffen-Batogowa i Katulska 1984) wykazała, że słuchacze potrafią bez większych trudności wyróżnić w wypowiedzi sylaby akcentowane. Do przetestowania własnego materiału przyjęto następujące hipotezy:

1. Pewne sylaby w wypowiedzi są przez słuchaczy percypowane jako szczególnie wyróżnione i jako ważne dla poprawnego odbioru zawartej w sygnale informacji.

2. Wyodrębnione percepcyjnie sylaby można zdefiniować w terminach cech akustycznych sygnału.

3. Wyznaczniki akustyczne słuchowego wyróżnienia sylab mogą być zróżnicowane zależnie od fonetycznych i gramatyczno-semantycznych uwarunkowań wypowiedzi.

Dla weryfikacji wyżej wymienionych hipotez przeprowadzono doświadczenia akustyczno-percepcyjne. Wyniki eksperymentów obejmujących materiał zawierający kontrastywne próbki mowy poddano analizie statystycznej. Utworzono 4 pary krótkich wypowiedzi, w obrębie których miejsce akcentu decydowało o interpretacji gramatyczno-semantycznej wypowiedzi: *nawóz — na wóz, jajem — ja jem, poradnia — pora dnia, zbieraliście — zbiera liście.* Wyrazy z powyższych par, traktowane w dalszej części doświadczenia jako kluczowe, umieszczono w krótkich zrandomizowanych dialogach. Dialogi te — 16 wypowiedzi (por. zał. 2) zostały odczytane przez 15-osobową grupę studentów różnych specjalności. Osobom czytającym nie udzielono specjalnych instrukcji, zalecono jedynie swobodny sposób wypowiedzi. Z otrzymanych 15 replikacji każdego dialogu wycięto fragmenty sygnału mowy odpowiadające wyłącznie wyrazom kluczowym.

Izolowane wypowiedzi kluczowe zapisano w pliku dźwiękowym według schematu:

wyraz kluczowy 1 …1 sek. ciszy *…pierwsze powtórzenie wyrazu kluczowego 1* …1 sek. ciszy… *drugie powtórzenie wyrazu kluczowego 1 …* 3 sek. ciszy *wyraz, kluczowy 2* 1 sek. ciszy

wyraz kluczowy 16 1 sek. ciszy

przykładowo:

nawóz …1 sek ciszy….nawóz.… 1 sek. ciszy … *nawóz* … 3 sek. ciszy

Przerwy między replikacjami tej samej wypowiedzi kluczowej wynosiły 1 sek, przed kolejnymi, ustawionymi w losowym porządku wypowiedziami kluczowymi — 3 sek. W ten sposób przygotowany materiał oceniała percepcyjnie 20-osobowa, przypadkowa grupa studentów. Po trzykrotnym odsłuchaniu wyrazu należało przypisać mu jedno z dwóch znaczeń. Decyzję słuchacze zapisywali na formularzach:

1. jajem ja jem
2. pora dnia poradnia
3. nawóz na wóz
16. zbieraliście zbiera liście

Rezultaty testu odsłuchowego dla wszystkich porównywanych wypowiedzi zamieszczono w tabeli 1 (w załączniku 3). W kolumnach umieszczono wyniki odsłuchów wypowiedzi dla każdej z 15 osób. Wyniki przetestowano testem istotności X2. Wartość teoretyczna testu $\chi 2$ dla jednego stopnia swobody przy $a = 0{,}05$ wynosi 3,8, a przy $\alpha = 0{,}001$ wynosi 6,6. W kolejnych wierszach tabeli 1 umieszczono odpowiednio: rozpoznaną wypowiedź, wyniki testu $\chi 2$ oraz poziom istotności *a*. Przykładowo, w replikacji wyrazu *jajem* zrealizowanej przez głos MM, 18 osób rozpoznało znaczenie wypowiedzi prawidłowo, 2 osoby błędnie. Wartość statystyki $\chi 2 = 12{,}8$, co oznacza, że różnice nie są istotne. Dla większości przypadków hipoteza zerowa zakładająca brak różnic między liczebnościami oczekiwanymi i otrzymanymi została odrzucona na poziomie istotności $\alpha = 0{,}001$. Większość znaczeń wypowiedzi została rozpoznana przez słuchaczy poprawnie.

Na podstawie analizy przeprowadzonego doświadczenia wyodrębnić można 3 przypadki:

— wypowiedzi, których znaczenia zostały w 100 % poprawnie rozpoznane (między rozpoznaniami zachodziły tylko nieistotne statystycznie różnice),

— wypowiedzi, których znaczenia zostały tylko częściowo rozpoznane,

— wypowiedzi, o znaczeniu nierozpoznawalnym percepcyjnie.

W 100 % poprawnie rozpoznano znaczenia fragmentów wypowiedzi wyciętych z następujących kontekstów:

Mnie się nie spieszy. Ja jem.

Wiesz czym go obrzucili? Jajem.

Czym go obrzucili, pomidorem czy jajem?

W tamtych workach jest piasek, a w tych nawóz.

Co jest w tych workach, piasek czy nawóz ?

Błędne rozpoznanie niektórych replikacji wypowiedzi kluczowych (tych, które wykazały istotne statystycznie różnice w ocenie znaczeń) otrzymano z następujących kontekstów:

Podnieś te worki i wrzuć je na wóz.

Czy podnieść te worki i wrzucić je na wóz?

Tam są najlepsi lekarze i to jest najlepsza poradnia.

Było już dość jasno, więc to była najlepsza pora dnia.

Czy to była najlepsza pora dnia?

Czy klocki rozrzucaliście, czy zbieraliście ?

Czy Tomek rozrzuca, czy zbiera liście?

Nie rozrzucaliście klocków, tylko je zbieraliście.

Pozbawienie kontekstu nie pozwoliło na poprawną identyfikację fragmentów wypowiedzi z następujących zdań.

Czy tam są najlepsi lekarze i czy to jest najlepsza poradnia ?

Czy to wszystko wczoraj zbieraliście, czy dzisiaj?

To wszystko wczoraj zbieraliście, a nie dzisiaj.

W 100% poprawne rozpoznanie znaczenia uzyskano w tych przypadkach, w których wypowiedź kluczowa określała najważniejszą informację (narzuconą mówcy przez strukturę wypowiedzi).

W sytuacji, w której mówca posiadał pewną dowolność interpretacyjną wypowiedzi (np. we frazie: *to była najlepsza poradnia* uwydatnić można informację *najlepsza* lub *poradnia*) wystąpiły trudności ze stuprocentowym rozpoznaniem znaczenia wypowiedzi niektórych osób. Znaczenia wypowiedzi kluczowych, które nie są nośnikiem najważniejszej informacji (wypowiedzi: *czy to wszystko wczoraj zbieraliście, czy dzisiaj?·, to wszystko wczoraj zbieraliście, a nie dzisiaj)* nie zostały percepcyjnie rozpoznane (wartość statystyki $\chi 2$ poniżej 5). W celu wyjaśnienia błędnych rozpoznań znaczenia wyrazu *poradnia* szczegółowej analizie akustycznej poddano całe wypowiedzi zawierające ten wyraz kluczowy: *czy tam są najlepsi lekarze i czy to jest najlepsza poradnia?*

Dla sprawdzenia hipotezy, że po akcencie głównym może występować tylko akcent poboczny określony poprzez relacje iloczasowe, przeprowadzono dodatkowy

eksperyment w którym słuchacze odsłuchiwali wypowiedziane przez 20-oso- bową grupę studentów pary wypowiedzi: *mówiłem ci, żebyś nie kładł belek na wóz', mówiłem ci, że to jest kiepski nawóz.* Każdy z uczestników 20-osobowej grupy studentów powtarzał wymienione wypowiedzi po usłyszeniu wypowiedzi wzorcowej, w której mówca świadomie podkreślił informację *mówiłem.*

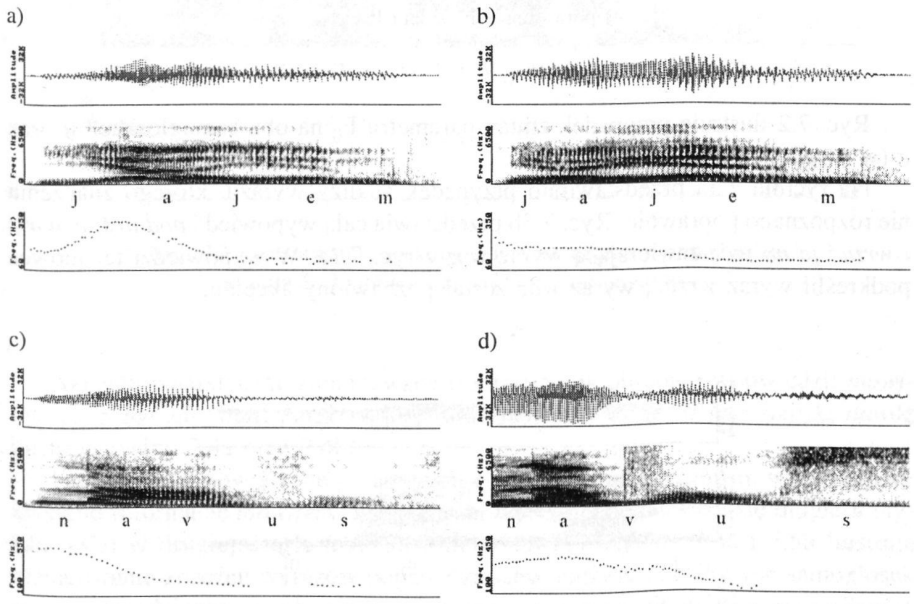
Analiza spektrograficzna wykazała możliwość udziału w akcentuacji poszczególnych wypowiedzi kluczowych następujących parametrów: częstotliwości podstawowej, czasu trwania samogłosek oraz poziomu sygnału. Przeprowadzono pomiary w zakresie 3 parametrów fizycznych.

1. Częstotliwości podstawowej. Ekstrakcji dokonano co 5 ms. W przypadkach wątpliwych przeprowadzono manualną korektę pomiarów na podstawie analizy widmowej.

2. Iloczasu. Segmentację sygnału przeprowadzono manualnie z dokładnością do 10 ms. Określono bezwzględny (wyrażony w ms) czas trwania samogłosek.

3. Poziomu sygnału. Przeprowadzono pomiar średniej poziomu w 20 ms interwałach czasowych.

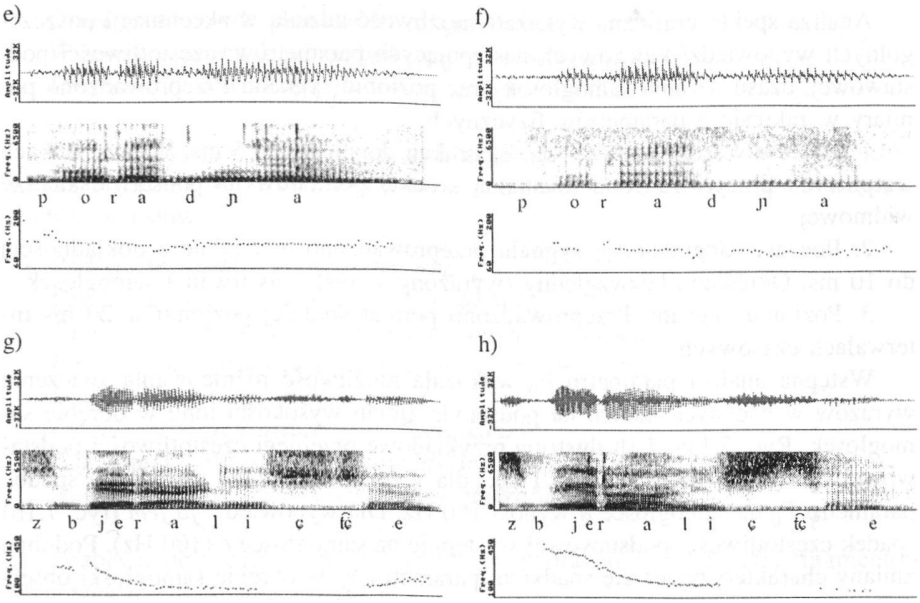
Wstępna analiza parametru F0 wykazała możliwość różnicowania znaczenia wyrazów w badanych parach na podstawie zmian wysokości tonu w obrębie samogłosek. Ryc. 7.1 a — 7.1h ilustrują przykładowe przebiegi częstotliwości podstawowej dla każdej badanej pary. I tak, dla wypowiedzi *jajem* (ryc. 7.1a) spadek parametru F0 na samogłosce *a* wynosi 150 Hz. Dla wypowiedzi *ja jem* (ryc. 7.1b) spadek częstotliwości podstawowej występuje na samogłosce *e* (100 Hz). Podobne zmiany charakteryzujące się spadkiem parametru F0 w obrębie samogłoski obserwuje się w pozostałych parach wyrazowych (ryc. 7.1c-7.1h).



Ryc. 7.1. Oscylogramy, spektrogramy i intonogramy wypowiedzi

a) jajem c) nawóz

b) ja jem d) na wóz



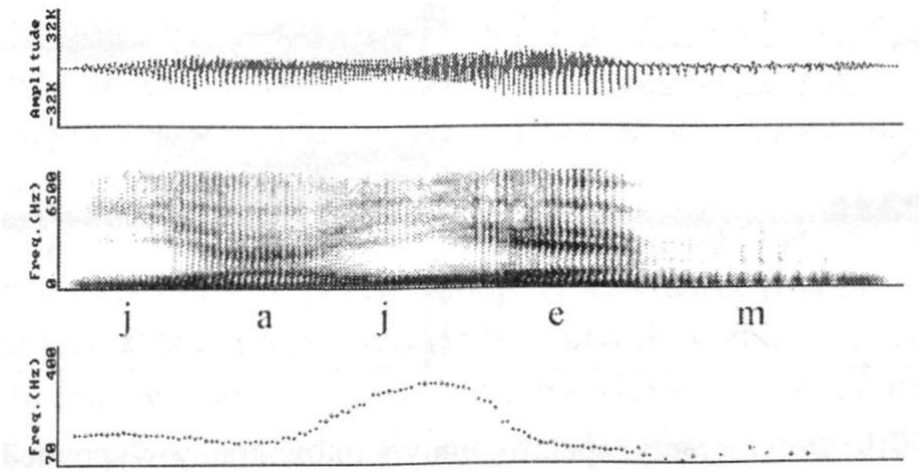
Ryc. 7.1. Oscylogramy, spektrogramy i intonogramy wypowiedzi

e) poradnia g) zbieraliście

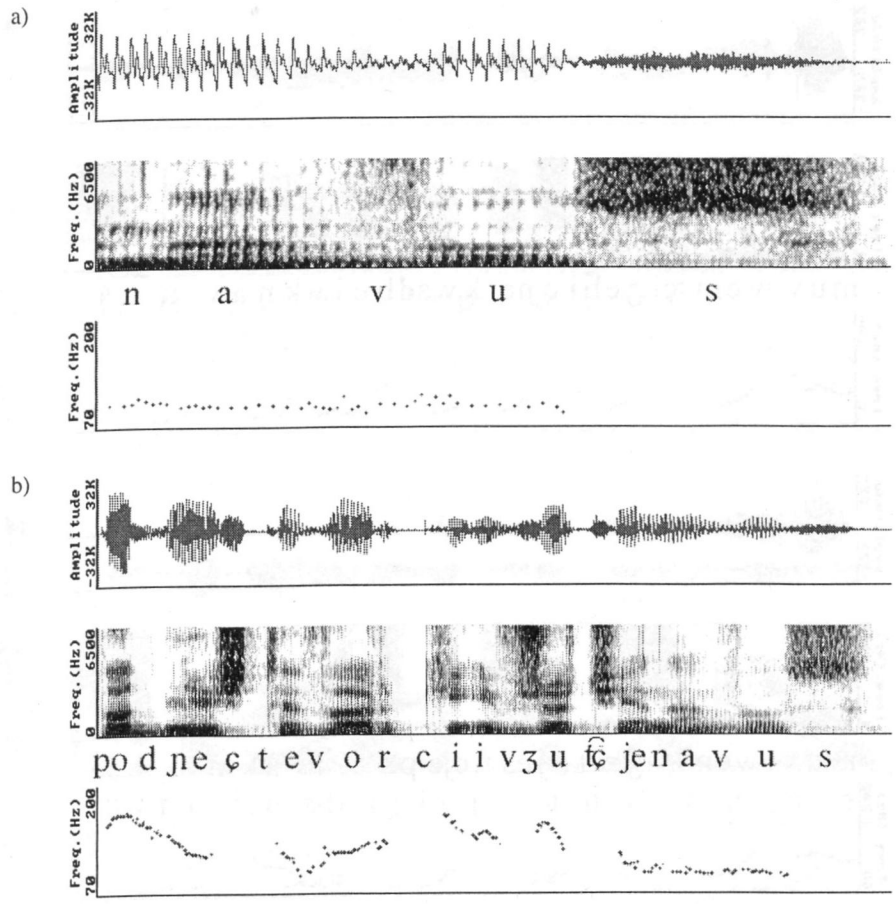
f) pora dnia h) zbiera liście

Ryc. 7.2 ilustruje przypadek zmian parametru F0 na obu samogłoskach *a* oraz *e* (w wypowiedzi *ja jem).*

Na rycinie 7.3a przedstawiono przypadek analizy wyrazu, którego znaczenia nie rozpoznano poprawnie. Ryc. 7.3b przedstawia całą wypowiedź *podnieś te worki i wrzuć je na wóz* zawierającą wyraz *wóz* z ryc. 7.3a. W wypowiedzi tej mówca podkreślił wyraz *wrzuć,* wyraz *wóz* został pozbawiony akcentu.



Ryc. 7.2. Oscylogram, spektrogram i intonogram wypowiedzi *ja jem*

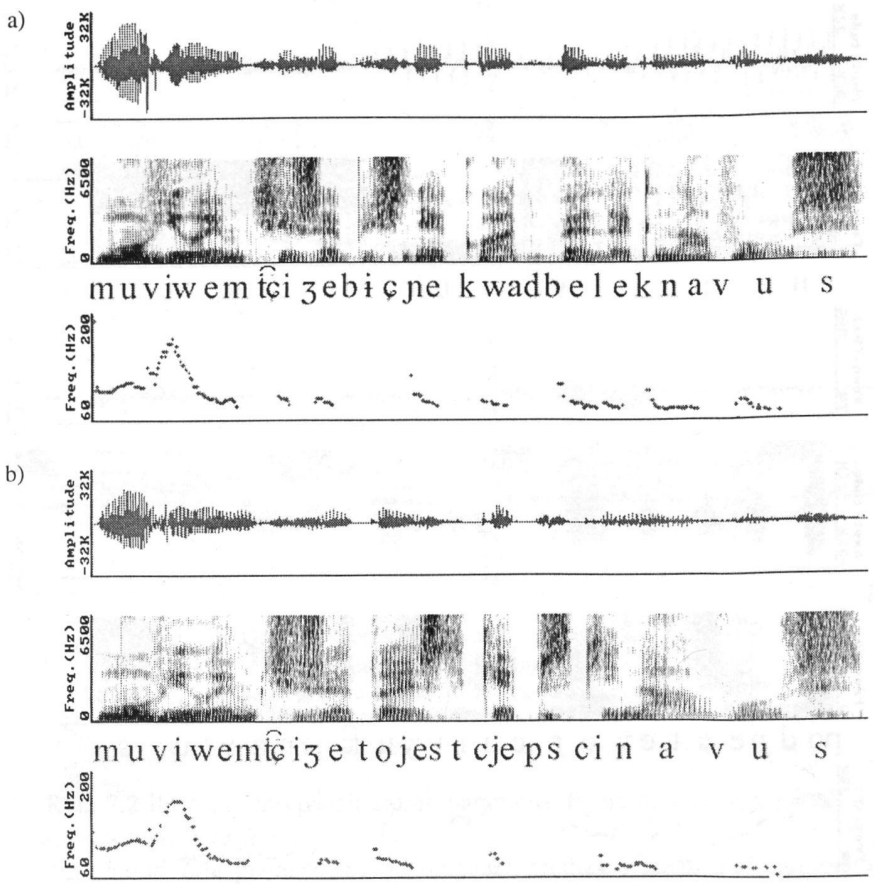


Ryc. 7.3. Oscylogramy, spektrogramy i intonogramy wypowiedzi

a) *na wóz,* b) *podnieś te worki i wrzuć je na wóz*

Akcenty postiktyczne zrealizowane iloczasowo w obrębie wypowiedzi: *mówiłem ci, żebyś nie kładł belek na wóz.* oraz *mówiłem ci że to jest kiepski nawóz* ilustrują ryciny 7.4a oraz 7.4b.

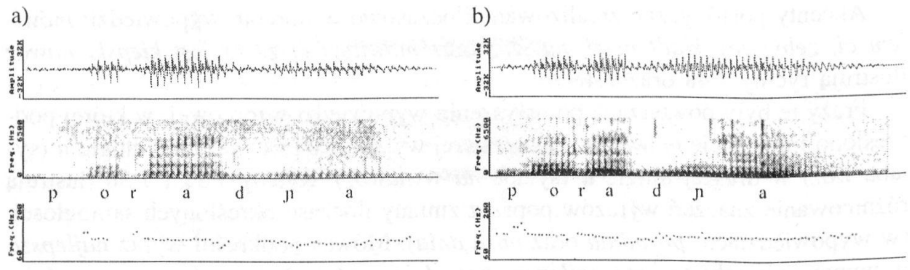
Frazy te były powtarzane po usłyszeniu wypowiedzi wzorcowej, w której podkreślono informację *mówiłem*. W pierwszej wypowiedzi zauważa się długie *u* (sylaba *wóz)* w drugiej długie *a* (sylaba *na* w *nawóz).* Ryciny 7.5a i 7.5b ilustrują różnicowanie znaczeń wyrazów poprzez zmiany iloczasu określonych samogłosek (w wypowiedziach: *poradnia* oraz *pora dnia).* Mówca podkreślił wyraz *najlepsza* w wypowiedziach: *to była najlepsza poradnia* oraz *to była najlepsza pora dnia*. W wyrazach *poradnia* oraz w *pora dnia* pozostał tylko akcent poboczny postiktyczny określony relacjami iloczasowymi.



Ryc. 7.4. Oscylogramy, spektrogramy i intonogramy wypowiedzi

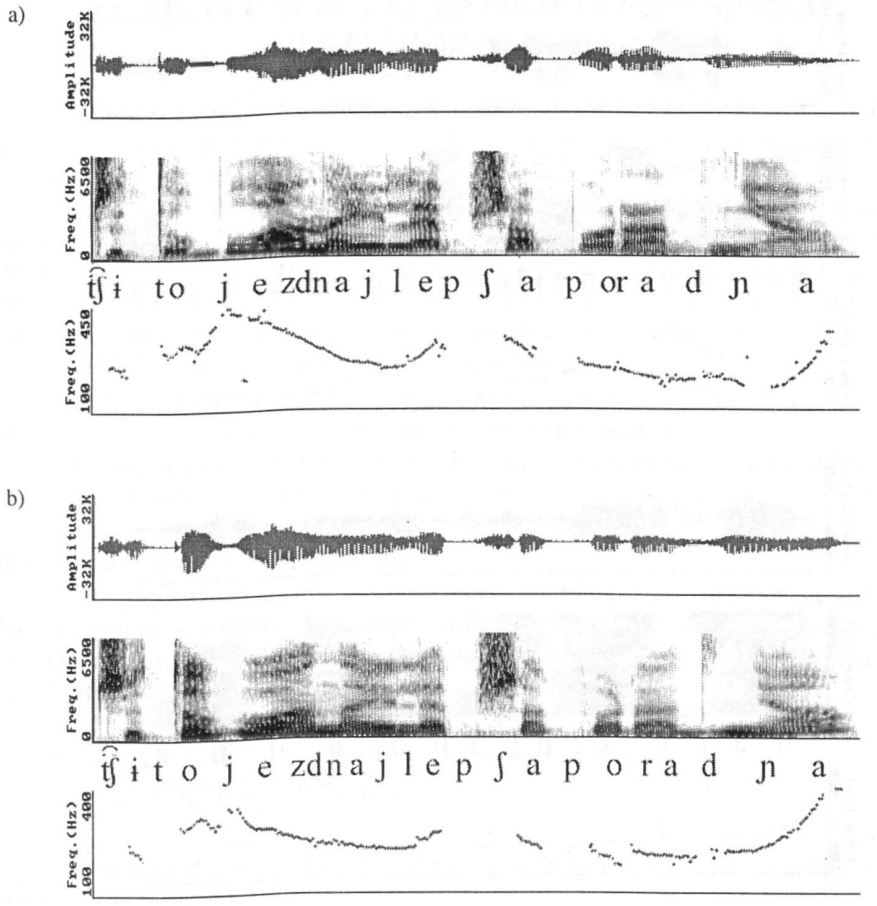
a) *mówiłem ci, żebyś nie kładł belek na wóz*

b) *mówiłem ci, że to jest kiepski nawóz*



Ryc. 7.5. Oscylogramy, spektrogramy i intonogramy wypowiedzi

a) *poradnia* b) *pora dnia*



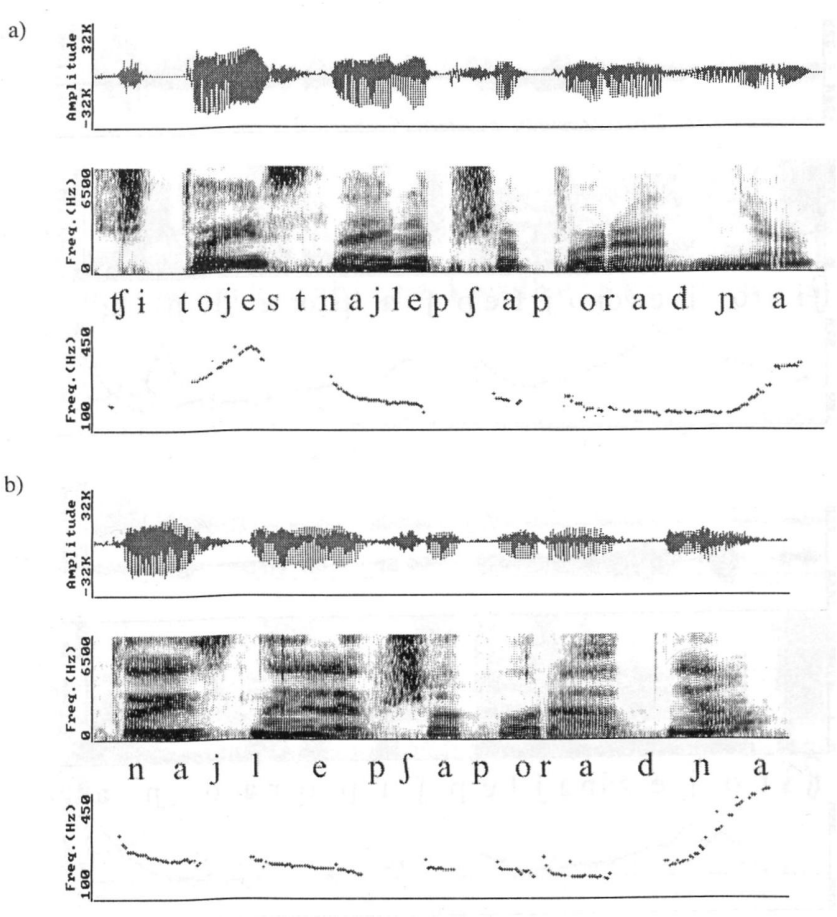
Ryc. 7.6. Oscylogramy, spektrogramy i intonogramy wypowiedzi

a) *czy to jest najlepsza poradnia?*

b) *czy to jest najlepsza pora dnia?*

Ryciny 7.6a i 7.6b ilustrują wyniki analiz dla wyrazów kluczowych *poradnia* oraz *pora dnia* umieszczonych w wypowiedziach pytajnych (czy *to jest najlepsza poradnia?* oraz *czy to jest najlepsza pora dnia?).*

Znaczenia tych wyrazów zostały dobrze percepcyjnie rozpoznane. Zauważa się dużą zmianę parametru F0 w obrębie sylaby *dnia* (ryc. 7.6a) oraz globalne minimum przebiegu parametru F0 na samogłosce *o* w wyrazie *pora* (ryc. 7.6b).



Ryc. 7.7. Oscylogramy, spektrogramy i intonogramy wypowiedzi

a) czy *to jest najlepsza poradnia?* b) *najlepsza poradnia*

Rycina 7.7a ilustruje przykład wypowiedzi, której znaczenie zostało nieprawidłowo rozpoznane (wyraz *poradnia* w wypowiedzi *czy to jest najlepsza poradnia?'). Przebieg parametru F0 charakteryzuje się minimum płaskim. Minimum występuje na samogłosce *o* oraz *a* w wyrazie *pora.* Duże trudności sprawiło słuchaczom rozpoznanie wyrazu *poradnia* w pytaniu: *czy tam są najlepsi lekarze* i *czy to jest najlepsza poradnia?* Tylko w jednym przypadku zilustrowanym na ryc. 7.7b słuchacze nie mieli trudności z rozpoznaniem wyrazu. Iloczas samogłoski *a* w wyrazie *poradnia* był prawdopodobnie istotnym czynnikiem wpływającym na decyzję słuchacza.

7.2. STATYSTYCZNA KLASYFIKACJA PARAMETRÓW SUPRASEGMENTALNYCH

Statystyczną analizę przeprowadzono oddzielnie w 2 grupach wypowiedzi kluczowych: dwusylabowe — *jajem, ja jem, nawóz, na wóz* oraz kilkusylabowe: *poradnia, pora dnia* i *zbieraliście, zbiera liście.*

Do przetestowania istotności akustycznych cech decydujących o percepcji akcentu opracowano zbiór kilkunastu parametrów opisujących zmiany: częstotliwości podstawowej (wartości ekstremalne, początkowe, końcowe, średnie), iloczasu

(względne zmiany długości samogłosek) i poziomu sygnału. Wyniki analizy wariancji wykazały, że efektywne w opisie akustycznym akcentu są następujące cechy:

1. Interwał zmian parametru F0 na samogłosce.
2. Względna zmiana parametru F0 na danej samogłosce, odniesiona do całego zakresu zmian częstotliwości podstawowej w wypowiedzi kluczowej.
3. Umiejscowienie przebiegu częstotliwości w danej samogłosce na indywidualnej skali częstotliwości mówcy (np. względem F0 min, F0 max).
4. Konfiguracja przebiegów częstotliwości podstawowej na samogłoskach w otoczeniu sąsiadujących ze sobą sylab.
5. W przypadku akcentu postiktycznego decydującą rolę odgrywają zmiany czasu trwania samogłosek w stosunku do czasu trwania samogłosek sąsiednich.

Dla wypowiedzi dwusylabowych wyznaczono następujące zbiory parametrów:

1. ΔDvi = Dvi-Dvs
2. ΔFi = ln(Fmaxi) - ln(Fmini)
3. ΔFr = AF - AFi
4. ΔFmax — ln(Fmax) — ln(Fmaxi)
5. ΔI = Isri-Isr

Dodatkowo, dla wyrazów kilkusylabowych przyjęto cechy uwzględniające sąsiedztwo samogłosek:

1. ΔdF1_2 = ΔFi-AFi+1
2. ΔdFkpi = ln Fki - ln Fpi+1

gdzie: Dvi — iloczas samogłoski,
Dvs — średni iloczas samogłosek w wypowiedzi,
Fmaxi Fmini — wartości ekstremalne parametru F0 w obrębie samogłoski,
Fmax — wartość maksymalna częstotliwości podstawowej w wypowiedzi,
ΔFi — określa różnicę logarytmu wartości maksymalnej i logarytmu wartości minimalnej parametru F0 na samogłosce,
ΔFi+1 — różnica logarytmu wartości maksymalnej i logarytmu wartości minimalnej parametru F0 na samogłosce następnej.
AF — określa różnicę logarytmu wartości maksymalnej i logarytmu wartości minimalnej parametru F0 w wypowiedzi,
Isri — średni poziom sygnału na samogłosce i,
Isr — średni poziom dla obu samogłosek,
Fki — końcowa wartość parametru F0 na samogłosce,
Fpi+1 — wartość początkowa parametru F0 na samogłosce następnej.
Dane znormalizowano poprzez transformację do wartości średniej równej zeru i jednostkowego odchylenia standardowego.

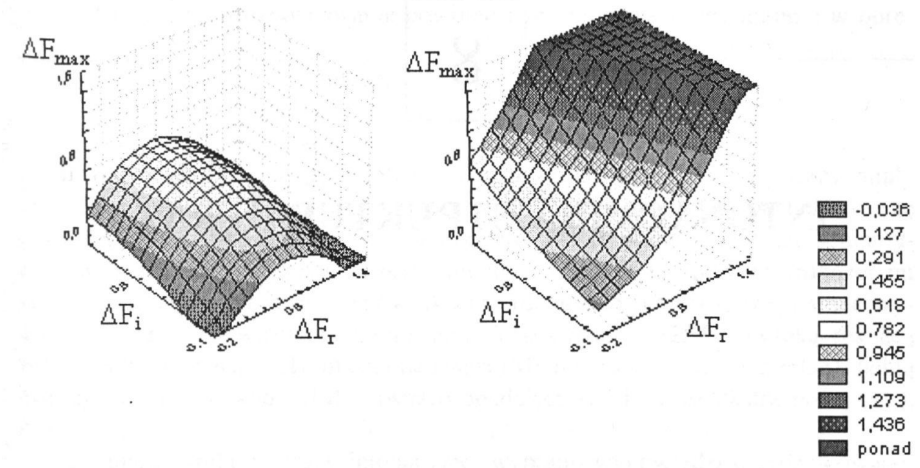
Na podstawie wyników analizy percepcyjnej każdą samogłoskę oznaczono jak akcentowaną lub nieakcentowaną. Dane poddano analizie dyskryminacyjnej. Jako zmienną niezależną przyjęto wyniki klasyfikacji percepcyjnej, zmienne zależne stanowiły wyżej wymienione cechy, określone dla każdej samogłoski. W tabeli 7.1 podano wyniki analizy dyskryminacyjnej samogłosek dla wypowiedzi dwusylabowych. Do klasyfikacji przyjęto tylko te wypowiedzi, których znaczenia zostały przez słuchaczy poprawnie rozpoznane. Z materiału zawierającego 7 wybranych do analizy wypowiedzi dwusylabowych (powtarzanych przez 15 osób) odrzucono 15 wyrazów. Do klasyfikacji pozostało więc 90 samogłosek akcentowanych

i 90 nieakcentowanych. Wysoki średni procent rozpoznania sylab akcentowanych i nieakcentowanych wynika z wyraźnych cech akustycznych analizowanych przykładów. Wypowiedzi dwusylabowe stanowiły centrum informacyjne zdania i były nośnikami akcentu rdzennego charakteryzującego się znacznymi zmianami parametru F0. W tabeli 2 (w załączniku 4) podano udział poszczególnych cech w klasyfikacji. Według testu Wilksa — w analizie wariancji — wartość statystyki F dla 3 parametrów opisujących zmiany częstotliwości podstawowej wynosi odpowiednio 41,04; 12,71 oraz 4,172, dla iloczasu 32,59, a dla poziomu sygnału 0,003. Z powyższych wartości wynika, że najważniejszą rolę w klasyfikacji odgrywają zmiany częstotliwości podstawowej. Iloczas spełnia funkcję pomocniczą i jak wykazały przedstawione przykłady charakteryzuje głównie akcent postiktyczny. Zmiany poziomu sygnału są nieistotne dla statystycznej analizy akcentu.

Tabela 7. 1. Wyniki rozpoznawania 90 samogłosek akcentowanych (grupa I) i 90 samogłosek nieakcentowanych (grupa II)

Grupa	Macierz klasyfikacji		Procent
	I	II	
I	85	5	94,44
II	7	83	92,22
Razem	92	88	średnia = 93,33

Wszystkie samogłoski akcentowane i nieakcentowane przedstawiono w układzie 3 współrzędnych opisujących zmienność częstotliwości podstawowej: AFi , AFr oraz AFmax (ryc. 7.8). Dla wypowiedzi kilkusylabowych, w których występowały zarówno akcenty rdzenne, jak i postiktyczne osiągnięto znacznie niższy procent klasyfikacji (w zakresie 68 - 74%). Z analiz wyłączono wypowiedzi pytające, ponieważ liczba ich replikacji nie była wystarczająco duża do oceny statystycznej. Wyniki eksperymentu odsłuchowego i przeprowadzona analiza dyskryminacyjna pozwoliły na wyciągnięcie następujących wniosków:



Ryc. 7.8. Samogłoski w układzie 3 współrzędnych: zmiany parametru F0 na samogłosce (AFi), względnej zmiany odniesionej do zakresu (AFr), względnego położenia samogłoski (AFmax) a) nieakcentowane, b) akcentowane

1. Akcenty mogą być rozpoznane bez wskazówek kontekstowych wyłącznie na podstawie cech akustycznych sygnału.

2. W określonych przypadkach może zachodzić zjawisko pozbawienia danej sylaby akcentu i przeniesienia go na inny fragment wypowiedzi.

3. Akcent realny główny związany jest zawsze ze zmianami wysokości tonu. Występuje w obrębie najważniejszej informacji w wypowiedzi.

4. W określonych przypadkach po akcencie realnym głównym może wystąpić akcent poboczny sygnalizowany iloczasem.

5. Wyniki klasyfikacji akustycznej akcentu/braku akcentu są podobne do rezultatów oceny słuchowej.

6. Najważniejszymi elementami sygnału mowy dla wyznaczenia parametrów akustycznych akcentów są samogłoski.

7. Cechy charakteryzujące zmienność parametru F0 powinny więc być w największym stopniu przydatne w klasyfikacji/rozpoznawaniu struktur akcentowych, natomiast iloczas jest istotny w detekcji akcentu postiktycznego i granic frazowych.

8 INTONACYJNA STRUKTURA FRAZY

8.1. PERCEPCYJNA ANALIZA STRUKTUR MELODYCZNYCH

Przeprowadzone prace w zakresie modelowania intonacji zarówno dla polskiego, jak i innych języków wykazały istnienie określonych wzorców intonacyjnych opisanych w terminach cech akustycznych sygnału mowy (problem percepcji kategorialnej i ustalenia wariantów intonacyjnych analizuje się również w muzyce, np. Rakowski 1999).

Jak dotychczas, nie ma jednak algorytmu przetwarzającego wartości częstotliwości podstawowej w jednostki tonu i/lub intonacji. Jest to taka sama sytuacja, jaka na segmentalnym poziomie analizy mowy dotyczy problemu przekształcania bezpośrednich próbek w ciąg fonemów, mimo że opis parametru F0 (w przeciwieństwie do opisu segmentalnego) jest pierwotnie jednozmienny. Układ przyjmujący sekwencje parametru F0 na wejściu i wytwarzający ciągi lingwistycznie kategorycznych lub quasi-kategorycznych wzorców intonacyjnych na wyjściu, wymaga ustalenia relacji między lingwistycznym (w tym fonetycznym) a akustycznym poziomem analizy. Do przetestowania przyjęto następujące hipotezy:

1. Jeżeli użytkownicy danego języka imitują przebiegi intonacyjne, to można dokonać obiektywnej decyzji, czy te imitacje są ekwiwalentne lingwistycznie lub paralingwistycznie, tzn. czy przenoszą tę samą informację, czy należą do tej samej klasy obiektów, pomimo występujących w imitacjach zróżnicowań osobniczych.

2. Jeżeli imitacje danego wzorca są poprawnie sklasyfikowane/rozpoznane, to jest oczywiste, że mówca przeprowadza proces klasyfikacji/rozpozna- wania przed imitacją danego wzorca. Interesujące jest zatem, czy i w jakim zakresie istnieje zgodność percepcyjnej i automatycznej klasyfikacji/roz- poznawania.

3. Klasyfikacja/rozpoznawanie powinno być realizowane co najmniej w obrębie:

    a. typu akcentu rdzennego,
    b. struktury intonacyjnej frazy,
    c. typu granicy frazowej.

Przeprowadzone eksperymenty w zakresie intonacyjnej struktury frazy miały na celu uzyskanie imitacji określonych wzorców intonacyjnych oraz ich ocenę percepcyjno-akustyczną. Jako wzorce przyjęto frazy o różnej strukturze semantyczno-gramatycznej, wymówione przez fonetyka z różnymi, typowymi intonacjami języka polskiego. Dla oceny typów akcentu rdzennego przyjęto krótką frazę wypowiedzianą z dziewięcioma intonacjami: niską rosnącą (LM), wysoką rosnącą (MH), pełną rosnącą (LH), niską opadającą (ML), wysoką opadającą (HM), pełną opadającą (HL), równą (MM), rosnąco-opadającą (LHL) oraz ekstra niską (xL), por. rozdz. 6.

Do analiz struktury frazy intonacyjnej wybrano wypowiedzi o zróżnicowanej długości:

- a. wyłącznie z akcentem rdzennym na początku lub na końcu frazy,

b. z akcentami preiktycznymi typu H oraz L.

Wszystkie frazy wymówione przez fonetyka (60 wypowiedzi umieszczonych w załączniku 5) zapisano w losowej kolejności w pliku dźwiękowym i traktowano w dalszej części doświadczenia jako wzorcowe. Wybrano losowo 26 studentów (10 głosów męskich i 16 kobiecych), którym postawiono zadanie jak najwierniejszego powtórzenia wzorców (trzykrotnego jedno- i dwusylabowych oraz dwukrotnego powtórzenia wielosylabowych). Po usłyszeniu wzorca student powtarzał go z zachowaniem swoich indywidualnych cech głosu (wysokości i tempa mowy). Osobom biorącym udział w doświadczeniu postawiono więc jako zadanie odtworzenie wzorców, a nie ich naśladownictwo. Odtworzenie poszczególnych intonacji miało na celu uzyskanie materiału porównawczego, w którym określona informacja językowa jest zakodowana przez różne osoby, a więc zawiera także dystynktywną informację osobniczą.

Dla sprawdzenia ekwiwalentności informacji językowej przeprowadzono doświadczenie percepcyjne, w którym inna 20-osobowa grupa studentów oceniała zgodność imitacji z danym wzorcem. Na specjalnych formularzach, według skali podobieństwa od 0 do 4, studenci oceniali stopień zgodności z wzorcem. Ocena 4 oznaczała całkowitą zgodność wzorca oraz imitacji, odpowiednio 0 — całkowitą niezgodność. Eksperyment (ocenę zgodności 3640 imitacji z 60. wzorcami) przeprowadzono w różnych odstępach czasu w ciągu miesiąca. Do jednorazowego odsłuchu podawano słuchaczom po kilkanaście par wypowiedzi (wzorzec: imitacja wzorca). W materiale umieszczono również kilka par testowych zawierających tę samą wypowiedź (np. wzorzec 1: wzorzec 1). Pary te umożliwiły kontrolę koncentracji uwagi słuchacza. Przykładowe oceny uzyskane dla osoby, której imitacje zostały najwyżej ocenione przez słuchaczy przedstawiono w tabeli 8.1.

Tabela 8.1. Fragment tabeli z ocenami imitacji wzorców dla głosu GI.

W poszczególnych kolumnach umieszczono liczbę słuchaczy, którzy ocenili imitację jako: 0 - zupełnie niezgodną, 1 - niezgodną, 2 - podobną, 3- bardzo podobną, 4 - taką samą

Frazy	Oceny					
	0	1	2	3	4	E
znów	1	1	5	8	5	2,75
znów	0	2	6	8	4	2,70
znowu	1	3	1	3	2	2,10
znowu on	0	2	2	14	2	2,80
znowu ona	0	0	4	9	7	3,15
znowu ten wariat	3	4	7	6	0	3,15
znów	0	0	6	5	9	3,15
znowu	0	1	4	9	6	3,00
znowu on	1	1	4	6	8	2,95
znowu ona	0	0	4	5	11	3,35
znowu ten wariat	0	0	1	10	9	3,40
znak	0	0	2	11	7	3,25
ten znak	0	1	5	7	7	3,00
jakiś znak	0	0	5	11	4	2,95
tu jest jakiś znak	0	0	2	10	8	3,30
znak	0	1	2	10	7	3,15
zły znak	0	1	3	10	6	3,05
bardzo zły znak	0	0	4	11	5	3,05
bardzo zły znak	0	0	2	9	9	3,35
znak	0	0	5	11	4	2,95
jakiś znak	0	0	4	10	6	3,10
tu jest jakiś znak	0	1	1	10	8	3,25
bardzo niedobry znak	0	0	4	8	8	3,20
to jest bardzo niedobry znak	1	0	9	8	2	2,50
bardzo zły człowiek	0	0	7	10	2	2,40
bardzo zły człowiek	1	0	7	9	3	2,65
co	0	1	6	7	6	2,90
proszę	0	1	6	6	7	2,95
co mówiłeś	0	0	2	10	8	3,30
co ona mówiła	0	0	2	7	11	3,45
to był całkiem niezły i uczciwy człowiek	0	0	4	9	7	3,15
to był całkiem niezły i uczciwy człowiek	0	1	2	14	3	2,95
Σ	8	21	140	281	189	
średnia	x=2,97					

Średnia ocen dla wszystkich wypowiedzi tej osoby wyniosła 3,2 (tylko dla fragmentu zamieszczonego w tabeli średnia wyliczona wynosi 2,97). Imitacje realizowane przez niektóre osoby zostały percepcyjnie ocenione jako bardzo dobre lub dobre, imitacje pochodzące od innych osób jako dostatecznie dobre (maksymalna średnia wyniosła 3,2, minimalna 2,8). Trudności występowały z powtórzeniem złożonej struktury melodycznej z wieloma akcentami (np. *to był całkiem niezły i uczciwy człowiek*) i związane były głównie z realizacją akcentu pobocznego. Wszystkie wypowiedzi, które uzyskały średnią ocenę podobieństwa powyżej 2 uznano jako zgodne z wzorcem.

Zgodność imitacji i wzorca uzyskano w 93,5%, co świadczy o dużej łatwości przeciętnego użytkownika języka w percepcyjnym rozpoznawaniu i klasyfikacji struktur melodycznych. Imitacje niezgodne z wzorcami (6,5%, a więc 235 wypowiedzi) zostały pominięte w dalszych analizach.

Materiał do dalszej analizy akustycznej i statystycznej podzielono następująco:

- a. wzorce i imitacje 9 typów akcentu rdzennego,

b. wypowiedzi wyłącznie z akcentem rdzennym (występującym na początku lub na końcu frazy),

c. wypowiedzi z akcentem rdzennym oraz jednym preiktycznym akcentem typu L lub H,

d. złożone melodycznie wypowiedzi z kilkoma akcentami pobocznymi.

8.2. SCHEMATY INTONACYJNE

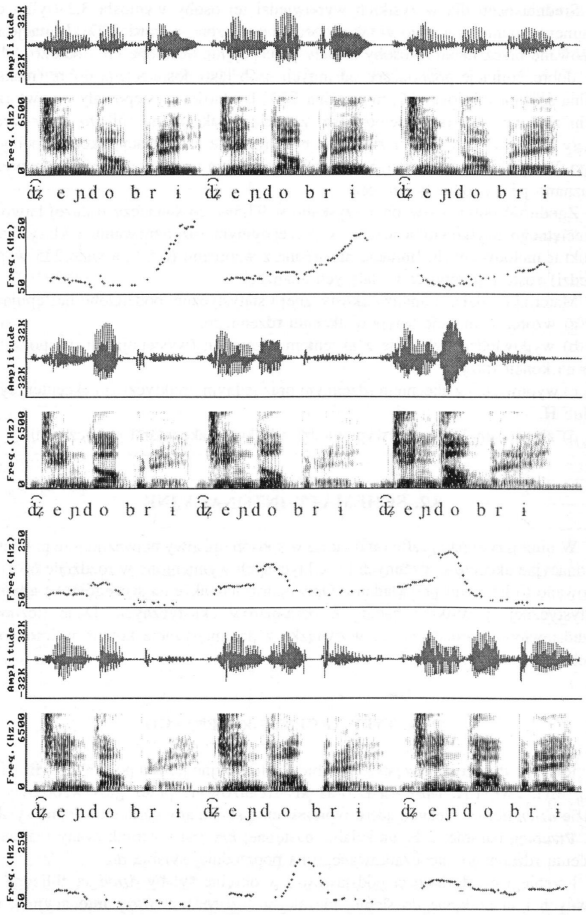
W niniejszym paragrafie omówiono w sposób opisowy najważniejsze przebiegi intonacyjne akcentów rdzennych i preiktycznych wymienione w rozdziale 6 i zilustrowano to licznymi przykładami. Opis oparto jednakże na szczegółowej analizie statystycznej wszystkich badanych parametrów akustycznych. Dane liczbowe przedstawiono w rozdziale 13 w związku z parametryzacją konturów intonacyjnych.

8.2.1. TYPY AKCENTU RDZENNEGO

Ryc. 8.1 ilustruje najczęstsze przebiegi intonacyjne języka polskiego: ML, HM, HL, MH, LM, LH, MM, LHL oraz xL zrealizowane przez głos wzorcowy na frazie *dzień dobry.* Decydujące o typie akcentu są zmiany parametru F0 na sylabie *do.* Przebieg parametru F0 na sylabie następnej *bry* jest uwarunkowany realizacją akcentu rdzennego zapoczątkowanego na poprzedniej sylabie *do.*

Przebiegi częstotliwości podstawowej w obrębie sylaby *dzień* są zbliżone do równych i są podobne do siebie, niezależnie od rodzaju frazy intonacyjnej, nie posiadają więc cech dystynktywnych odróżniających typy akcentu głównego.

Imitowanie powyższych wzorców intonacyjnych nie sprawiało trudności. W nie-



Ryc. 8.1. Oscylogramy, spektrogramy i intonogramy wypowiedzi: *dzień dobry*. 9 typów akcentu rdzennego: LH, MH, LM, ML, HM, HL, xL, LHL, MM

licznych przypadkach występowały problemy związane głównie z odtworzeniem właściwego zakresu zmian (np. LH, LM i MH). Ocena percepcyjną imitacji przeprowadzona przez 20 osób wykazała w 95% zgodność wzorców i imitacji. Dla stwierdzenia, czy przynajmniej niektóre spośród analizowanych przebiegów intonacyjnych wyróżniają się pod względem sytuacyjno-kontekstowym, przeprowadzono eksperyment polegający na przypisaniu każdej wypowiedzi pewnego znaczenia emocjonalnego. Doświadczenie przeprowadzono w 20-osobowej grupie słuchaczy. Po usłyszeniu wypowiedzi *dzień dobry* słuchacz proszony był o przypisanie wypowiedzi jednego z dziewięciu znaczeń według następującego klucza:

- a. wypowiedź rzeczowa, sucha,
  b. zaskoczenie, zdumienie,
  c. przymilność, pochlebstwo,
  d. zdziwienie, pytanie,
  e. uprzejmość, kontynuacja,
  f. znudzenie,
  g. opryskliwość, niechęć,
  h. wylewność,
  i. perswazja, uspokojenie.

Kategorie te ustalono arbitralnie na podstawie wstępnego przetestowania kilku słuchaczy i ich propozycji przypisania określonym wypowiedziom poszczególnych kontekstów sytuacyjnych i zawartości emocjonalnej. Każdy słuchacz indywidualnie przypisywał znaczenie słyszanych wzorców na podstawie wielokrotnego

odsłuchu wypowiedzi. Wyniki eksperymentu wykazały, że niektóre wzorce były bez trudności kojarzone z określoną sytuacją, np. dla wypowiedzi oznaczających „opryskliwość", „wylewność", „zaskoczenie" zgodność słuchaczy osiągnęła 98%.

Trudności wystąpiły z określeniem wypowiedzi oznaczających „przymilność" lub „uprzejmą kontynuację". W tych przypadkach zgodność słuchaczy wyniosła zaledwie 70%. W ogólnym wyniku doświadczenia słuchacze przypisali poszczególnym wzorcom intonacyjnym następujące znaczenia:

wypowiedź rzeczowa, sucha — ML, zaskoczenie, zdumienie — HL, przymilność, pochlebstwo — LM, zdziwienie, pytanie — LH, uprzejmość, kontynuacja — MH, znudzenie — MM, opryskliwość, niechęć — xL, wylewność — LHL, perswazja, uspokojenie — HM.

Analizowane przebiegi intonacyjne stanowią więc realizację funkcjonalnej jednostki intonacyjnej. Obszerne doświadczenie poświęcone możliwości identyfikacji zawartości emocjonalnej wypowiedzi przeprowadziła dla języka polskiego Steffen-Batogowa (1996). Autorka analizowała związki intonacji z zawartością emocjonalną określoną następującymi kategoriami:

a. zdziwienie, zniecierpliwienie, radość, niedowierzanie, przestrach,

b. ironia, ton wyzywający, irytacja, pobłażliwość, stanowczość,

c. ton protekcjonalny, uspokajający, rzeczowy, oburzenie, namysł.

Wyniki badań przeprowadzonych przez Steffen-Batogową (1966) wykazały, że jednoznaczna interpretacja wypowiedzi nie zawsze jest możliwa na skutek oddziaływania różnych czynników, takich jak kategoria gramatyczna zdania, kontekst sytuacyjny czy zabarwienie emocjonalne.

Przeprowadzony w niniejszej pracy eksperyment, polegający na przypisaniu znaczeń emocjonalnych dziewięciu wypowiedziom frazy *dzień dobry* miał wyłącznie charakter pomocniczy. Jego celem było ustalenie funkcjonowania podstawowych wzorców intonacyjnych wyraźnie różniących się miedzy sobą cechami akustycznymi sygnału.

Zgodność słuchaczy w zakresie 70 - 95% pozwala wysunąć wniosek, że przebiegi intonacyjne — zależnie od kontekstu sytuacyjnego — reprezentują poszczególne typy wzorców: pełna rosnąca (LH), pełna opadająca (HL), wysoka rosnąca (MH), wysoka opadająca (HM), niska rosnąca (LM), niska opadająca (ML), równa (MM), rosnąco-opadająca (LHL) oraz ekstra niska (xL). Wyniki te należy zweryfikować za pomocą analizy statystycznej, która może wykazać istnienie dystynktywnych cech akustycznych odróżniających poszczególne wzorce.

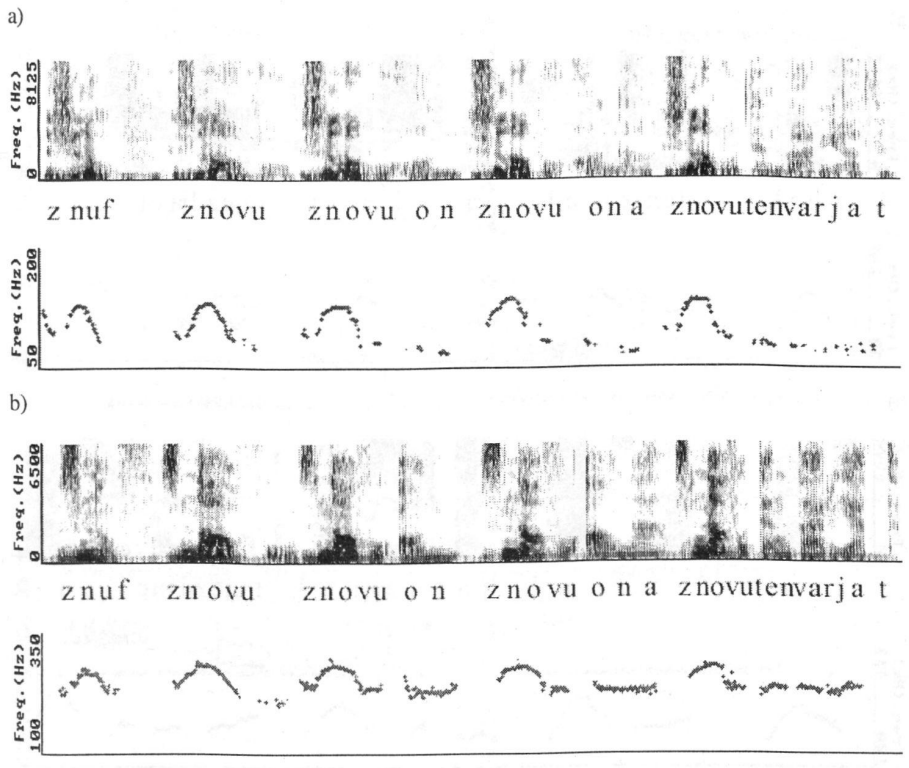
8.2.2. WYPOWIEDZI Z JEDNYM AKCENTEM RDZENNYM

8.2.2.1. Intonacje opadające

Na ryc. 8.2a przedstawiono intonację wzorcową typu ML zrealizowaną we frazach: *znów, znowu, znowu on, znowu ona, znowu ten wariat* wymówionych z akcentem rdzennym na pierwszej sylabie. Ryc. 8.2b ilustruje przykładowe imitacje poszczególnych intonacji w realizacji najlepszego imitatora.

Wizualna ocena przebiegów pozwala zauważyć wysokie podobieństwo zarówno między imitacjami i wypowiedziami wzorcowymi, jak i między frazami różniącymi się tylko długością zakończenia intonacyjnego. Istotne zmiany parametru F0 występują głównie na pierwszej sylabie wypowiedzi, tj. sylabie rdzennej. W analizowanych przypadkach są to zmiany rzędu kilkudziesięciu Hz. Przebieg tonu na sylabie rdzennej, zależnie od konkretnej realizacji, jest bardziej lub mniej rosnąco-opadający. W przypadku fraz dłuższych (np. *znowu ten wariat,* ryc. 8.2a oraz ryc. 8.2b fraza 5) efekt zmian parametru F0 można zaobserwować również na

sylabach następujących, głównie na sylabie bezpośrednio występującej po sylabie akcentowanej. Na dalszych sylabach frazy obserwuje się tylko znikomy wpływ zmian parametru F0 zapoczątkowanych na sylabie rdzennej. Podobne spostrzeżenia dotyczą intonacji typu ML z akcentem rdzennym na końcu frazy. Ryc. 8.3a i 8.3b ilustrują odpowiednio wzorce i imitacje wypowiedzi: *znak, ten znak, jakiś znak, tu jest jakiś znak.*



Ryc. 8.2. Spektrogramy i intonogramy fraz z akcentem rdzennym ML na początku wypowiedzi: *znów, znowu, znowu on, znowu ona, znowu ten wariat*

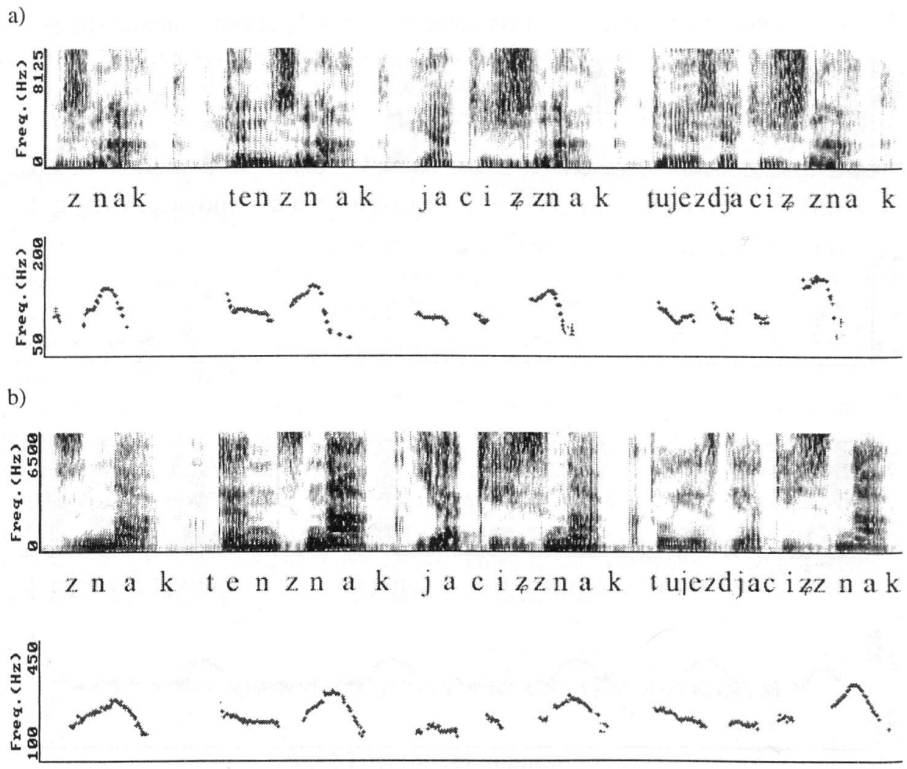
a) wzorce b) imitacje wzorców

Analogicznie jak w poprzednim przypadku zauważa się, że zmiany parametru F0 realizowane są na sylabie rdzennej. Zmiany te poprzedzone są przedrdzennym, prawie równym przebiegiem intonacyjnym. Całkowity spadek parametru F0 występuje wyłącznie na ostatniej sylabie. Sylaby przedrdzenne są umiejscowione w środkowym zakresie zmian częstotliwości. Również w tym przypadku wizualna ocena przebiegów pozwala zauważyć podobieństwa między wzorcami i ich imitacjami.

Interesująca jest również analiza zmian parametru F0 w obrębie sylaby rdzennej.

Ryc. 8.4 przedstawia frazę *znów* (typ intonacji ML, wypowiedź wzorcowa) i jej imitacje realizowane przez 9 osób. Oś czasu znacznie rozciągnięto, by bardziej szczegółowo móc prześledzić synchronizację maksimum konturu intonacyjnego z początkiem czasu trwania samogłoski.

We wszystkich przypadkach spadek parametru F0, decydujący o przynależności

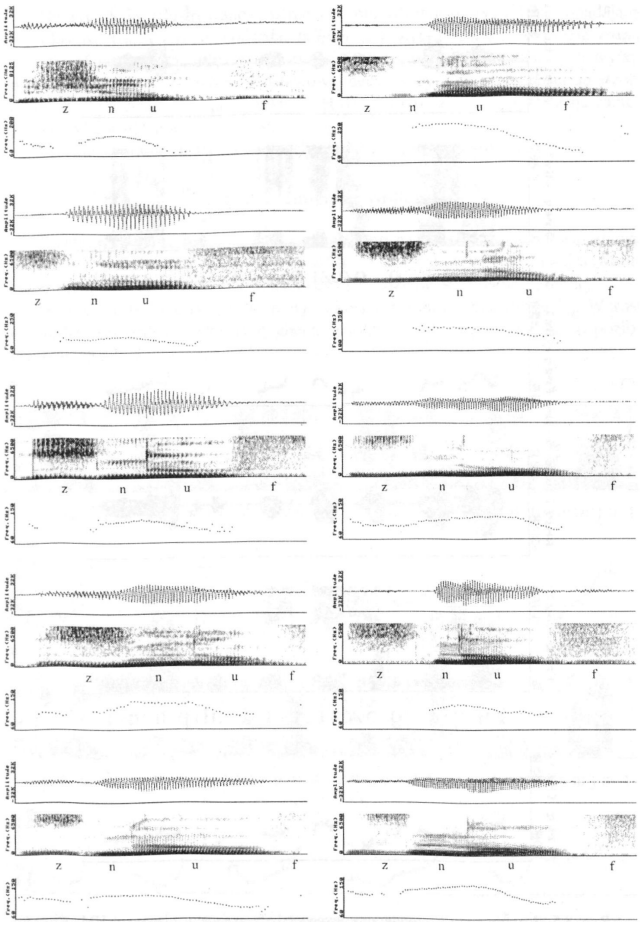


Ryc. 8.3. Spektrogramy i intonogramy fraz z akcentem rdzennym ML na końcu wypowiedzi: *znak, ten znak, jakiś znak, tu jest jakiś znak*

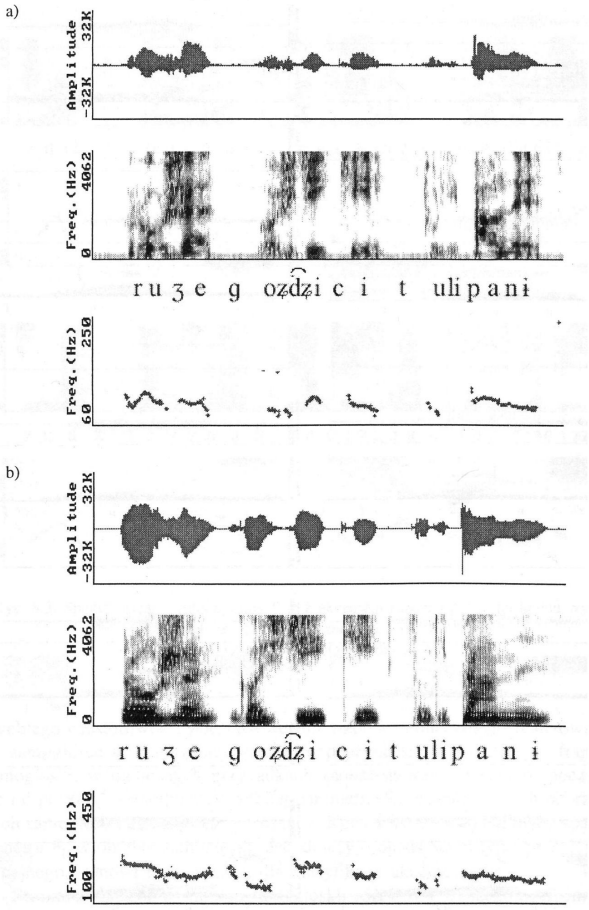
a) wzorce b) imitacje wzorców

przebiegu częstotliwości podstawowej do wzorca intonacyjnego realizowany jest na samogłosce *u*. Najczęściej spadek rozpoczyna się w środkowym fragmencie samogłoski, w nielicznych przypadkach zaobserwowano spadek rozpoczynający się od początku samogłoski. Przebieg parametru F0 na spółgłoskach poprzedzających samogłoskę jest najczęściej rosnący. Konsekwentnie występujący spadek parametru F0 w obrębie samogłoski, decydujący o opadającym typie przebiegu intonacyjnego, stanowi istotną cechę dla klasyfikacji akcentu.

Stromość spadku w obrębie samogłoski, różna w poszczególnych imitacjach wzorca, wydaje się nie mieć istotniejszego znaczenia dla percepcji typu akcentu (wszystkie imitacje zostały ocenione jako zgodne z wzorcem). Istotniejsza jest lokalizacja zmiany. Jeżeli spadek występuje od początku samogłoski, realizuje się na tej samogłosce w całości. Jeżeli spadek parametru F0 występuje pod koniec samogłoski, to realizuje się on w całości na sąsiadujących sylabach.



Ryc. 8.4. Wzorzec frazy *znów* (lewy górny rysunek) i jej imitacje przez 9 parlatorów. Akcent rdzenny typu ML



Ryc. 8.5. Spektrogramy i intonogramy wypowiedzi z akcentem rdzennym HM: *róże, goździki, tulipany*

a) wzorce b) imitacje wzorców

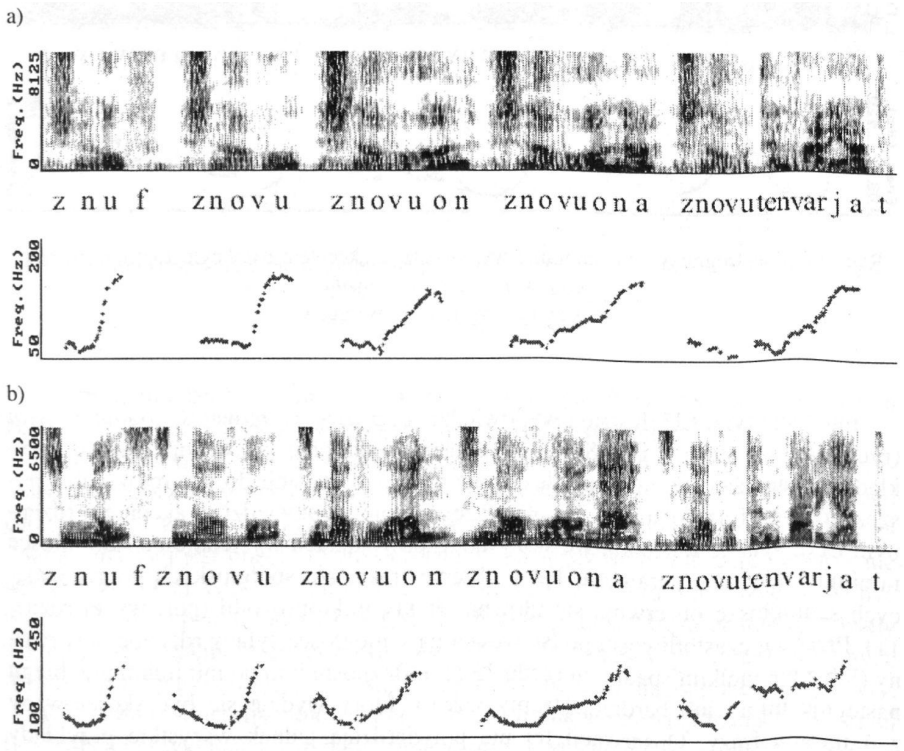
Na ryc. 8.5a zilustrowano wzorce akcentu rdzennego typu HM (opadający wysoki) we frazach *róże, goździki, tulipany* (w wypowiedzi: *w ogrodzie rosną kwiaty: róże, goździki, tulipany)* na ryc. 8.5b przedstawiono imitacje tych wypowiedzi. Spadek częstotliwości podstawowej decydujący o przynależności do wzorca intonacyjnego (w tym przypadku do HM) zrealizowany jest głównie na samogłosce sylaby rdzennej.

Przebieg parametru F0 na sylabach przedrdzennych jest prawie równy.

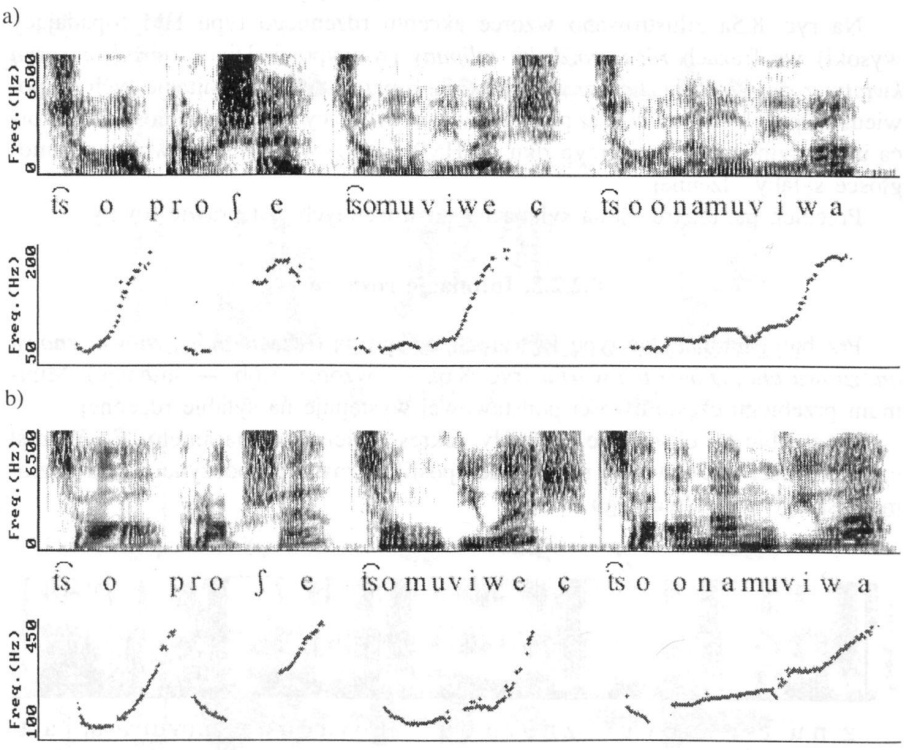
8.2.2.2. Intonacje rosnące

Przebiegi intonacyjne typu LM zrealizowano na frazach *znów, znowu, znowu on, znowu ona, znowu ten wariat* (ryc. 8.6a — wzorce, 8.6b — imitacje). Minimum przebiegu częstotliwości podstawowej występuje na sylabie rdzennej.

Na sylabie tej obserwuje się mały zakres zmienności parametru F0. Wzrost częstotliwości podstawowej na sylabach postiktycznych rozpoczyna się od poziomu niskiego L do średniego M.



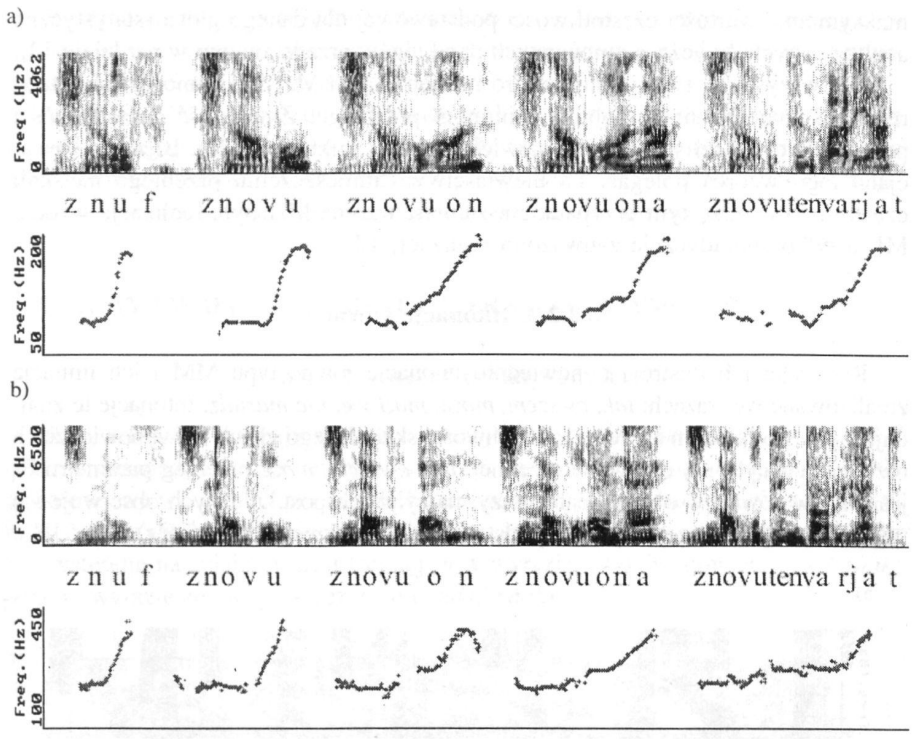
Ryc. 8.6. Spektrogramy i intonogramy wypowiedzi z akcentem rdzennym LM: *znów, znowu, znowu on, znowu ona, znowu ten wariat*a) wzorce b) imitacje wzorców



Ryc. 8.7. Spektrogramy i intonogramy wypowiedzi z akcentem rdzennym LH: *co, proszę, co mówiłeś, co ona mówiła*

a) wzorce b) imitacje wzorców

Intonacje typu LH ilustrują wypowiedzi *co, proszę, co mówiłeś, co ona mówiła* (ryc. 8.7a — wzorce, ryc. 8.7b — imitacje). Podobnie jak w poprzednich przykładach, zauważa się zarówno duże podobieństwo imitacji do wzorców, jak i poszczególnych fraz do siebie. Na pierwszej sylabie wypowiedzi (z akcentem rdzennym LH), sylabie rdzennej, we wszystkich przypadkach występuje na samogłosce minimum globalne parametru F0 w obrębie frazy. Na spółgłoskach poprzedzających samogłoskę obserwuje się głównie efekty mikroprozodii (por. wyżej rozdz. 11). Przebieg częstotliwości podstawowej na samogłosce sylaby rdzennej jest równy (lub z niewielkim spadkiem rzędu 15 Hz). Bezpośrednio po minimum przebiegu następuje mniej lub bardziej stromy wzrost, który wydaje się być skorelowany z długością frazy. Obserwacji tej nie potwierdzają jednak wszystkie przykłady (por. np. ryc. 8.7a fraza 5: *co ona mówiła?).* Na sylabie *co* zauważa się minimum przebiegu, później lekki wzrost parametru F0 na sylabach *ona mówi* i gwałtowny wzrost na ostatniej sylabie *ła.*



Ryc. 8.8. Spektrogramy i intonogramy wypowiedzi z akcentem rdzennym MH: znów, znowu, znowu on, znowu ona, znowu ten wariat a) wzorce b) imitacje wzorców

Różnicę między wzorcami LH i LM obserwuje się głównie w interwale zmian parametru F0 na sylabach występujących po sylabie rdzennej.

Szybkość zmiany nie jest w percepcji krytyczna, wszystkie imitacje zostały ocenione jako zgodne z wzorcami. Istotna natomiast jest lokalizacja minimum przebiegu parametru F0. W obydwu typach wzorców przebieg częstotliwości podstawowej na sylabie rdzennej jest prawie równy lub lekko opadający (interwał zmian wynosi nie więcej niż 15 Hz). Globalne minimum przebiegu na sylabie rdzennej wydaje się być istotne dla percepcji wzorca LH i LM. Nieliczne imitacje, które charakteryzowały się minimum płaskim (z umiejscowieniem kilku samogłosek w pobliżu minimalnych wartości przebiegu) zostały przez słuchaczy ocenione jako niepodobne do wzorców.

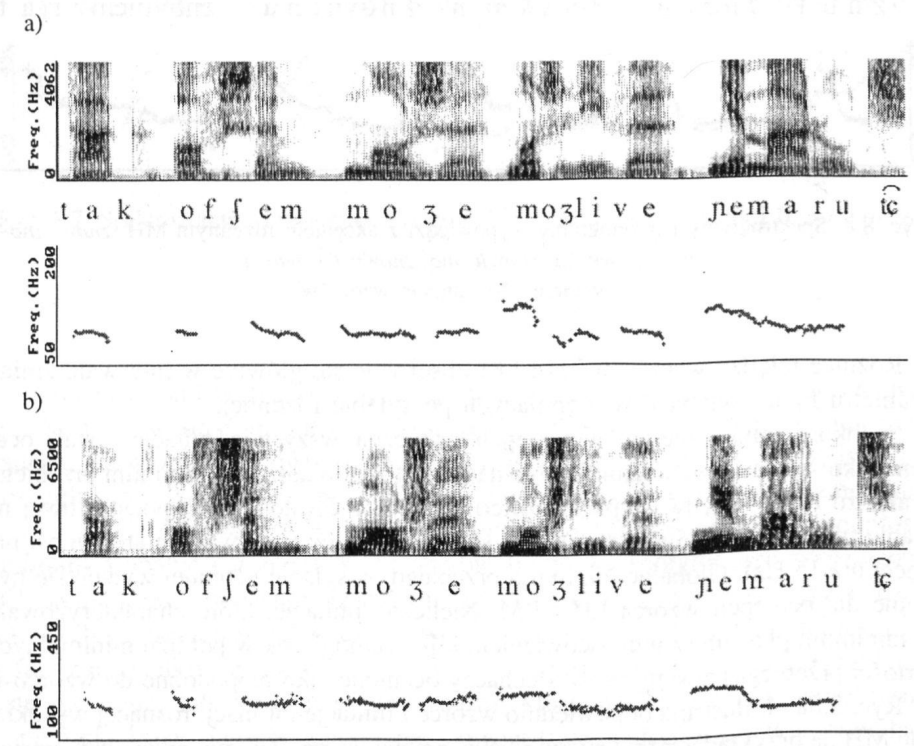
Ryc. 8.8a i b ilustrują odpowiednio wzorce i imitacje intonacji rosnącej wysokiej typu MH na przykładach fraz *znów, znowu, znowu on, znowu ona, znowu ten wariat.*

Minimum przebiegu parametru F0 występuje na samogłosce sylaby rdzennej (pierwszej sylaby w poszczególnych frazach). Minimum to występuje w środkowym zakresie zmian częstotliwości. Przebieg po sylabie rdzennej jest rosnący do maksymalnej wartości częstotliwości podstawowej dla danego głosu (statystyczną analizę danych dotyczącą omawianych przebiegów przedstawiono w rozdziale 13).

Dla wszystkich typów intonacji rosnących (LH, LM, MH) samogłoska sylaby rdzennej znajduje się w pobliżu minimum przebiegu. Zmienność częstotliwości podstawowej na tej sylabie jest niewielka (rzędu kilkunastu Hz). Błędy w imitacjach tego wzorca polegały na niewłaściwym umieszczeniu przebiegu na skali częstotliwości. Częstym przypadkiem zamiast wzorca MH była realizacja wzorca ML a tylko sporadycznie zauważono realizację LH.

8.2.2.3. Intonacje równe

Ryc. 8.9a i b ilustrują odpowiednio intonacje równe typu MM i ich imitacje zrealizowane we frazach: *tak, owszem, może, możliwe, nie marudź.* Intonacje te znajdują się w środkowym zakresie częstotliwości skali danego głosu. W wypowiedziach rozpoczynających się od sylaby rdzennej: *tak, owszem, może* przebieg parametru F0 jest prawie równy w obrębie całej frazy, na sylabach postiktycznych obserwuje się tylko lekki spadek częstotliwości podstawowej (w zakresie 10-15 Hz).



Ryc. 8.9. Spektrogramy i intonogramy wypowiedzi z akcentem rdzennym MM: *tak, owszem, może, możliwe, nie marudź* a) wzorce b) imitacje wzorców

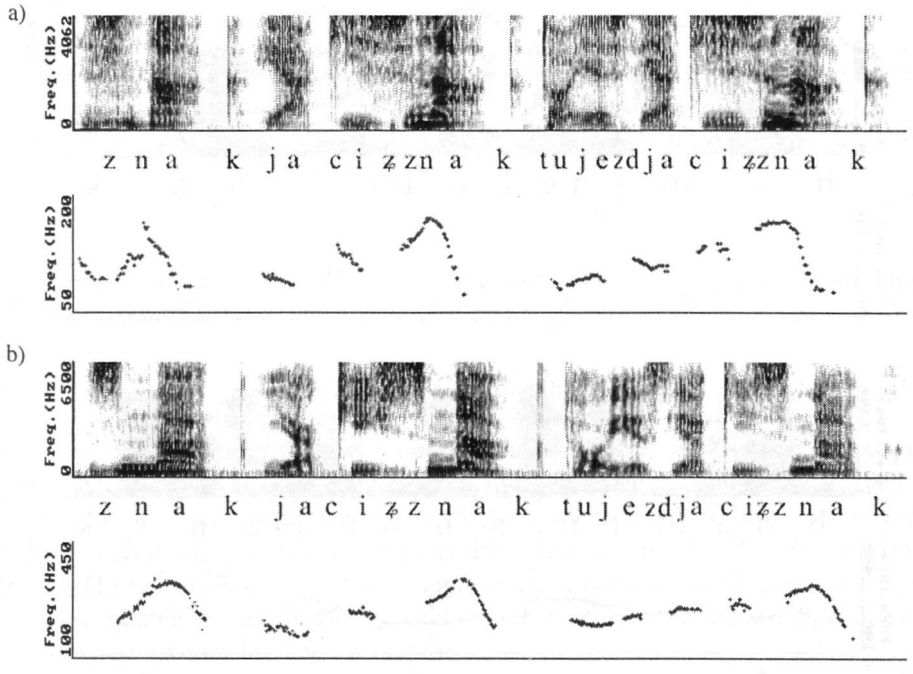
W wypowiedziach poprzedzonych anakruzą: *możliwe, nie marudź* przebieg częstotliwości na sylabach rdzennych (głównie na samogłoskach), decydujący o przynależności przebiegu do wzorca jest prawie równy.

Wzorzec MM należał do najłatwiejszych w imitacji. Niekiedy zdarzały się pomyłki polegające na realizacji zamiast MM wzorca intonacyjnego typu ML. Nie było również trudności w umieszeniu tego wzorca poprawnie na skali częstotliwości — w środkowym zakresie zmian typowym dla danego głosu.

8.2.3. WYPOWIEDZI Z AKCENTAMI PREIKTYCZNYMI TYPU L LUB H

8.2.3.1. Akcent preiktyczny L

Ryc. 8.10a i b ilustrują wzorce i imitacje frazy *znak, jakiś znak, tu jest jakiś znak,* wypowiedziane z akcentem rdzennym na końcowej sylabie frazy *znak*. Na sylabie tej we wszystkich przypadkach zrealizowany jest stromy spadek (typ akcentu HL) całkowicie w obrębie samogłoski *a* w sylabie *znak*. Akcent poboczny (typu L) występuje na sylabie przedrdzennej *ja* w wyrazie *jakiś* (wypowiedź *jakiś znak)* oraz w wyrazie *tu* (wypowiedź *tu jest jakiś znak).*



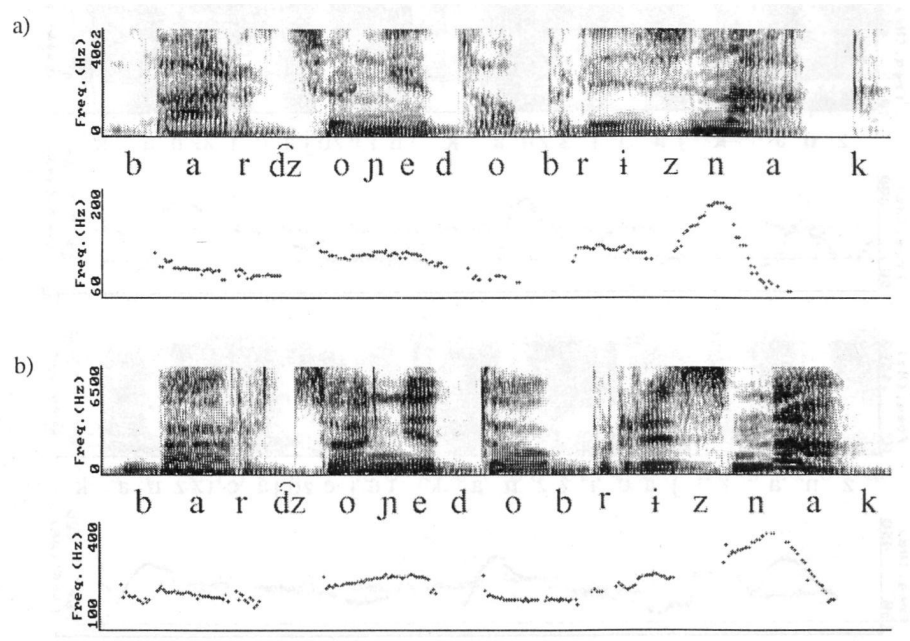
Ryc. 8.10. Spektrogramy i intonogramy wypowiedzi z akcentem rdzennym HL i akcentem pobocznym L: *znak, jakiś znak, tu jest jakiś znak* a) wzorce b) imitacje wzorców

Na sylabach akcentowanych pobocznie (typ L) zauważa się minimum lokalne przebiegu częstotliwości podstawowej. Np. sylaba *ja* jest położona znacznie niżej niż sylaba *kiś* (w wypowiedzi *jakiś znak),* analogicznie sylaba *tu* znajduje się poniżej sylaby *ja* (w wypowiedzi *tu jest jakiś znak).*

Ryc. 8.11 a i b ilustrują odpowiednio wzorce i imitacje wypowiedzi *bardzo niedobry znak* z dwoma akcentami pobocznymi L i akcentem rdzennym (typu HL) na sylabie *znak.* Akcenty poboczne (typu L) wyróżnić można na sylabie *bar* w wyrazie *bardzo i dob* w wyrazie *niedobry.* Przebieg częstotliwości podstawowej na samogłoskach z akcentem L jest opadający (do 15 Hz) i zawiera lokalne minimum. Analiza imitacji wykazała trudności w realizacji akcentów preiktycznych typu L. Najczęściej pomijano realizację tego akcentu lub zamiast 2 akcentów realizowano tylko jeden.

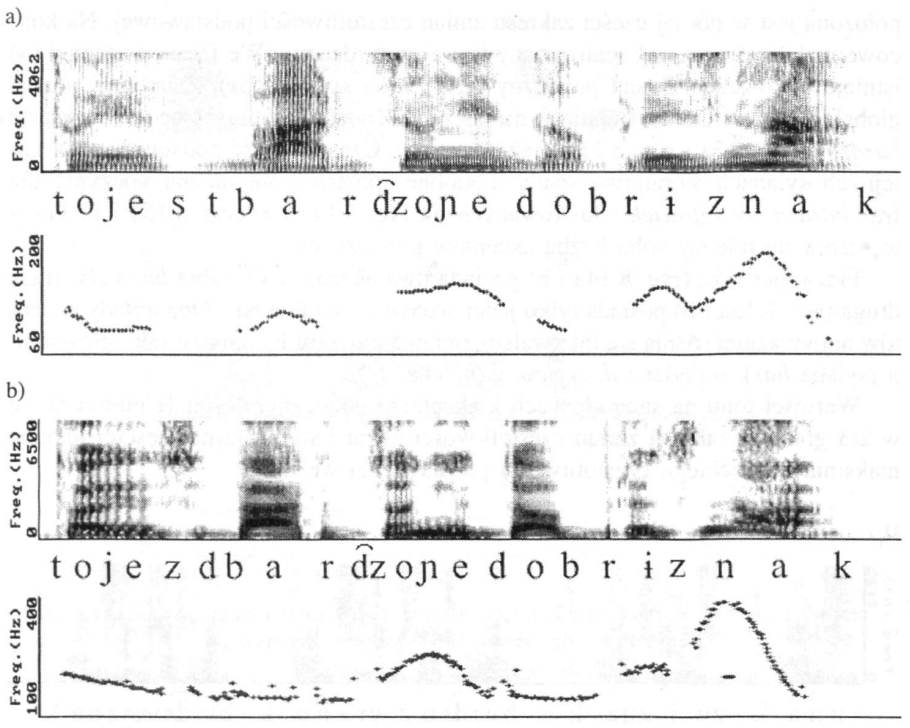
Ryc. 8.12a i b ilustrują odpowiednio wzorce i imitacje wypowiedzi *to jest bardzo niedobry znak* z dwoma akcentami pobocznymi i nieakcentowanymi sylabami (*to jest*) na początku frazy. Akcenty poboczne typu L wyróżnić można na sylabie *bar* w wyrazie *bardzo i dob* w wyrazie *niedobry.* Podobnie jak w poprzednim przypadku, przebieg częstotliwości podstawowej na samogłosce z akcentem L jest lekko opadający i zawiera lokalne minimum. Na sylabach nieakcentowanych zauważa się lekki spadek częstotliwości (15-20 Hz).

W obu przypadkach z ryc. 8.11 i 8.12 nie zauważono wpływu akcentów preiktycznych na akcent rdzenny.



Ryc. 8.11. Spektrogramy i intonogramy wypowiedzi z akcentem rdzennym HL i dwoma akcentami pobocznymi L: *bardzo niedobry znak*

a) wzorzec b) imitacja wzorca



Ryc. 8.12. Spektrogramy i intonogramy wypowiedzi z akcentem rdzennym HL oraz dwoma akcentami pobocznymi L: *to jest bardzo niedobry znak*

a) wzorzec b) imitacja wzorca

Wartości lokalnych minimów dla akcentów typu L leżą w dolnym zakresie zmian parametru F0 (w dolnej 1/3 charakterystycznego zakresu dla głosu mówcy).

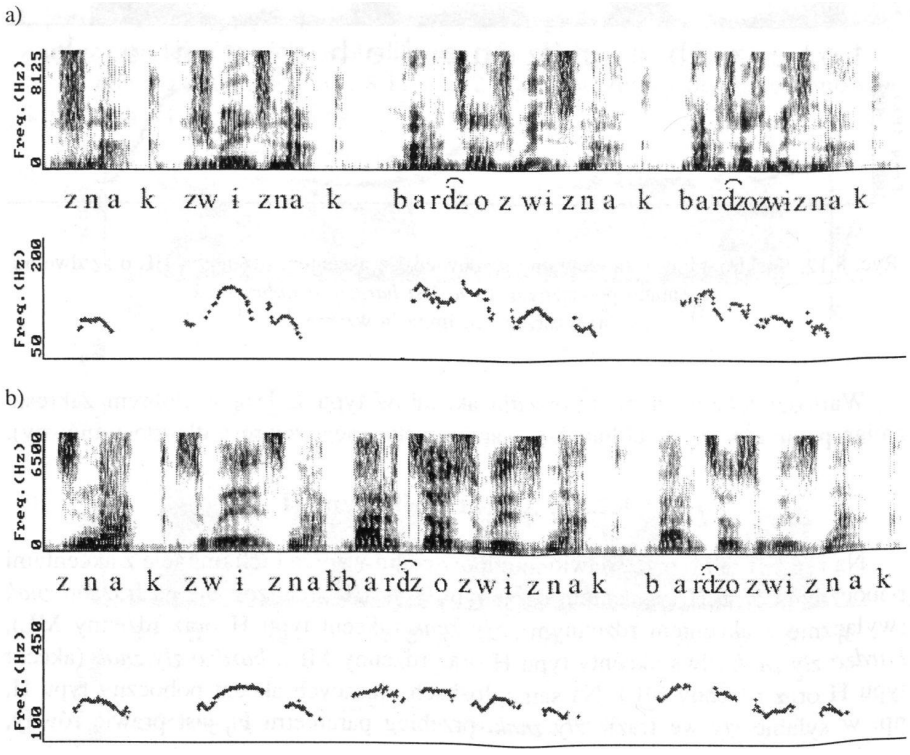
8.2.3.2. Akcent preiktyczny H

Na ryc. 8.13a i b przedstawiono odpowiednio wzorce i ich imitacje z akcentami pobocznymi typu H i akcentem rdzennym typu ML zrealizowane na frazach znak (wyłącznie z akcentem rdzennym), *zły znak* (akcent typu H oraz rdzenny ML), *bardzo zły znak* (dwa akcenty typu H oraz rdzenny ML), *bardzo zły znak* (akcent typu H oraz rdzenny ML). Na samogłoskach niosących akcent poboczny typu H, np. w sylabie *zły* we frazie *zły znak*, przebieg parametru F0 jest prawie równy, występuje na nim lokalne maksimum.

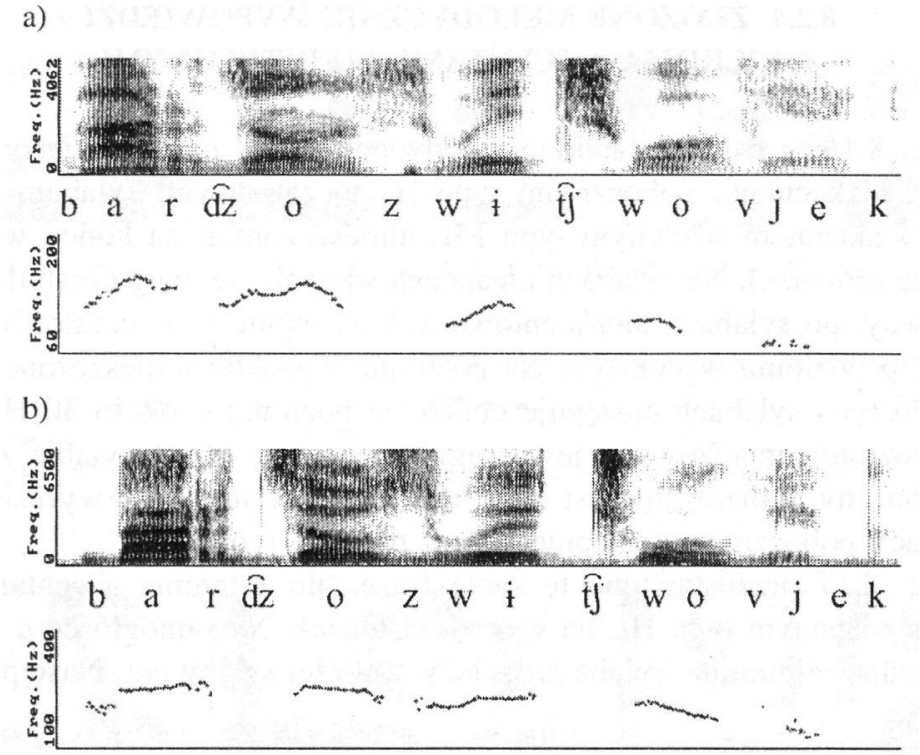
Dwie ostatnie frazy (3 i 4) *bardzo zły znak* różnią się między sobą liczbą akcentów. We frazie 3 dwie pierwsze samogłoski (w wyrazie *bardzo*) położone są wysoko, przebieg parametru F0 zbliżony jest do równego. Również sylaba *zły* położona jest w górnej części zakresu zmian częstotliwości podstawowej. Na końcowej sylabie frazy *znak* realizowany jest akcent rdzenny. We frazie następnej (4) istnieje tylko jeden akcent poboczny typu H (na sylabie *bar)*. Zauważyć można globalne maksimum występujące na początku frazy na samogłosce *a* w wyrazie *bardzo* (ryc. 8.13a i ryc. 8.13b, fraza ostatnia). Częstotliwość podstawowa na kolejnych sylabach stopniowo spada. Podobne spostrzeżenie można poczynić dla fraz *bardzo zły człowiek* zilustrowanych na ryc. 8.14a i b oraz 8.15a i b. Frazy te różnią się między sobą liczbą akcentów pobocznych.

Fraza pierwsza (ryc. 8.14a i b) posiada dwa akcenty na sylabie *bar* i *zły,* fraza druga (ryc. 8.15a i b) posiada tylko jeden akcent na sylabie *bar.* Oba układy akcentów najwyraźniej różnią się interwałem zmian parametru F0 między samogłoskami: *a* (sylaba *bar*), *o* (sylaba *dzo*) oraz *y* (sylaba *zły*).

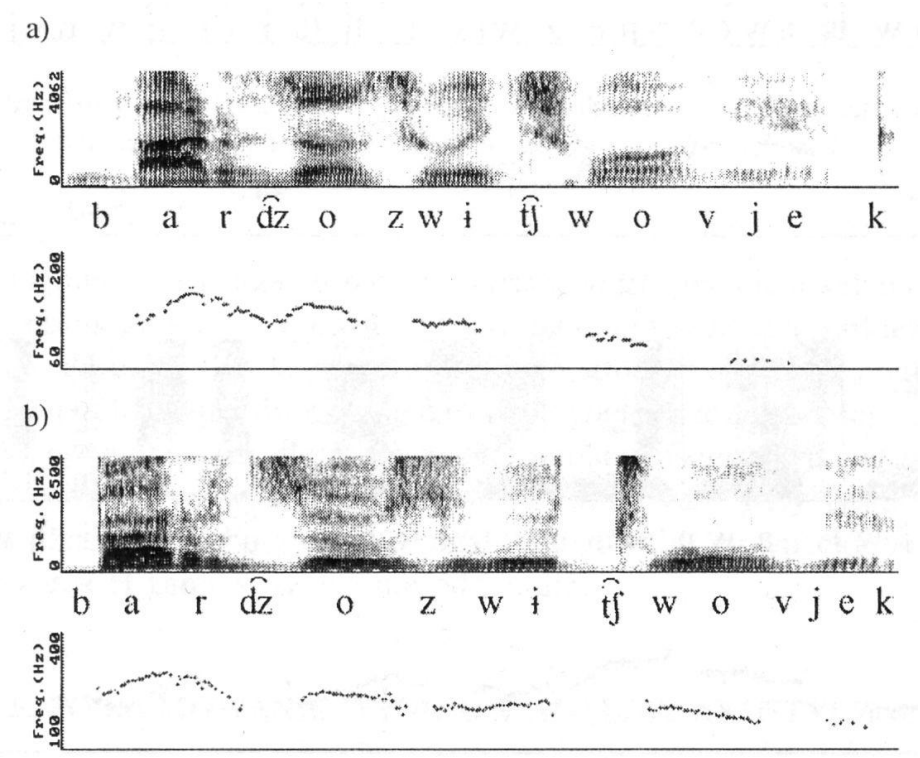
Wartości tonu na samogłoskach z akcentem pobocznym typu H mieszczą się w 2/3 górnego zakresu zmian częstotliwości głosu i stanowią najczęściej lokalne maksima w przebiegu częstotliwości podstawowej we frazie.



Ryc. 8.13. Spektrogramy i intonogramy wypowiedzi z akcentem rdzennym ML i akcentami pobocznymi H: *znak, zły znak, bardzo zły znak, bardzo zły znak* a) wzorce b) imitacje wzorców



Ryc. 8.14. Spektrogramy i intonogramy wypowiedzi z akcentem rdzennym ML i dwoma akcentami pobocznymi H: *bardzo zły człowiek* a) wzorzec b) imitacja wzorca



Ryc. 8.15. Spektrogramy i intonogramy wypowiedzi z akcentem rdzennym ML i akcentem pobocznym H: *bardzo zły człowiek*

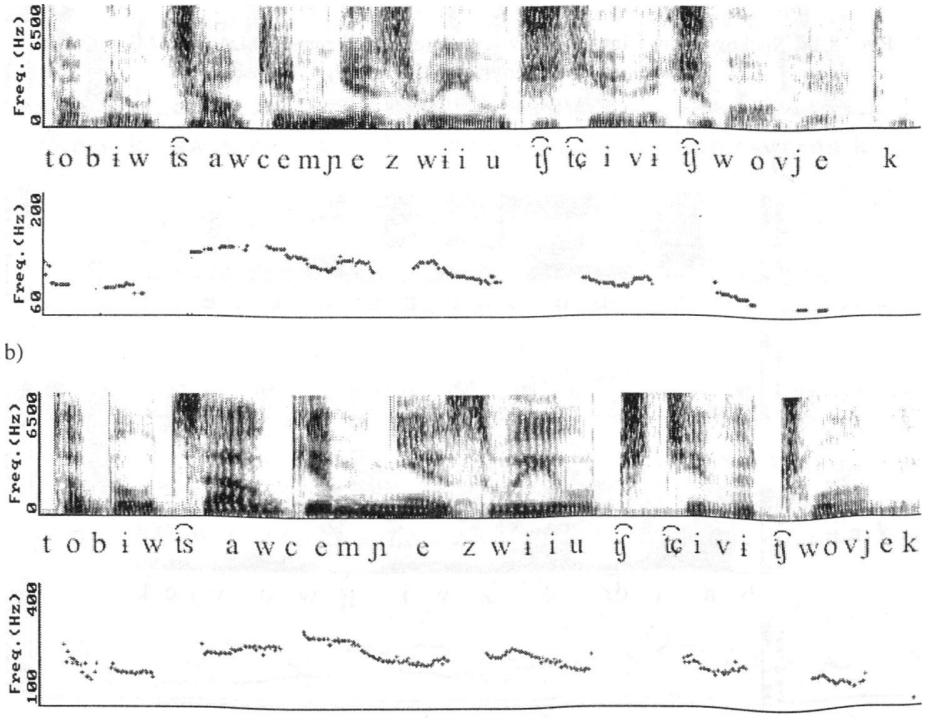
a) wzorzec b) imitacja wzorca

8.2.4. ZŁOŻONE MELODYCZNIE WYPOWIEDZI Z KILKOMA AKCENTAMI PREIKTYCZNYMI

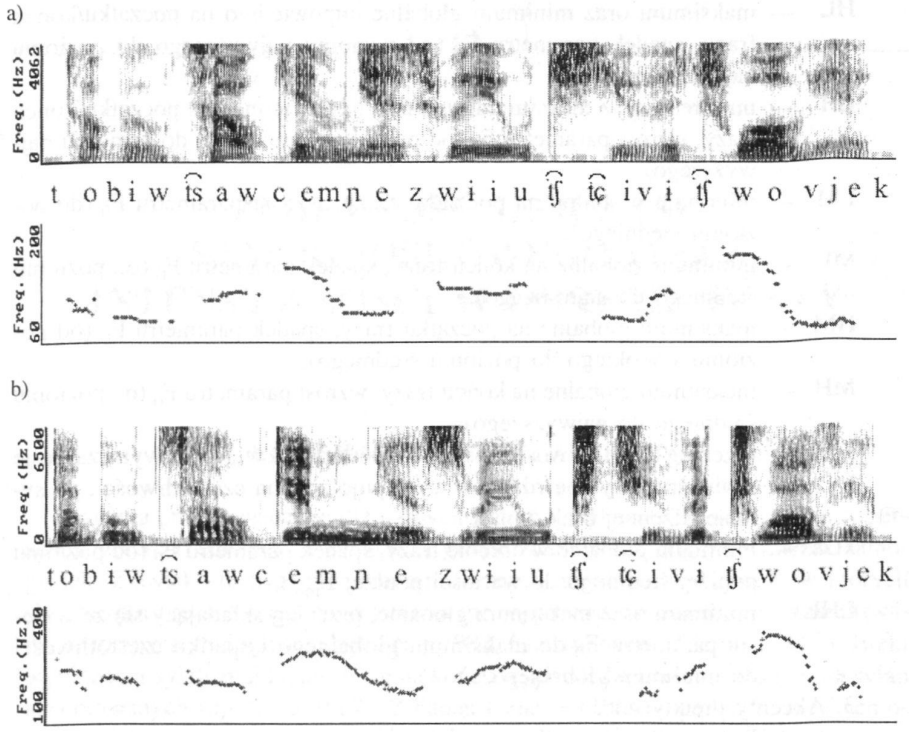
Na ryc. 8.16a i b przedstawiono wypowiedź *to był całkiem niezły i uczciwy człowiek* z 3 akcentami pobocznymi typu H, początkowymi sylabami nieakcentowanymi i akcentem rdzennym typu ML umieszczonym na końcu wypowiedzi (na wyrazie *człowiek).* Na sylabach nieakcentowanych przebieg częstotliwości jest prawie równy, po sylabach nieakcentowanych następuje skok częstotliwości podstawowej do poziomu wysokiego. Na poziomie wysokim umieszczone są sylaby *całkiem.* Po tych sylabach następuje obniżenie poziomu o około 30 Hz. Na tym niższym poziomie realizowany jest drugi akcent typu H na sylabie *nie.* Trzeci akcent poboczny realizowany jest jeszcze niżej na sylabie *ci* w wyrazie *uczciwy.* Po akcentach pobocznych następuje akcent rdzenny typu ML.

Na ryc. 8.17 przedstawiono tę samą frazę, ale z trzema akcentami typu L i akcentem rdzennym typu HL na wyrazie *człowiek.* Na samogłosce *a* w *cał* występuje lokalne minimum. Sylaba *kiem* leży powyżej sylaby *cał.* Następne lokalne

a)



Ryc. 8.16. Spektrogramy i intonogramy wypowiedzi z akcentem rdzennym ML i akcentami pobocznymi typu H: *to był całkiem niezły i uczciwy człowiek* a) wzorzec b) imitacja wzorca



Ryc. 8.17. Spektrogramy i intonogramy wypowiedzi z akcentem rdzennym HL i akcentami pobocznymi typu L: *to był całkiem niezły i uczciwy człowiek*

a) wzorzec b) imitacja wzorca

minimum znajduje się na samogłosce *e* w wyrazie *niezły,* trzecie lokalne mimimum występuje na samogłosce *i* w wyrazie *uczciwy.* Samogłoski te są nośnikiem akcentu typu L. Analiza ryc. 8.16 i 8.17 wykazuje, że samogłoski z akcentem preiktycznym położone są w pobliżu punktów ekstremalnych przebiegu częstotliwości. Ryc. 8.16b i 8.17b ukazują wypowiedzi dla osoby, której imitacje osiągnęły najwyższe oceny podobieństwa do wzorców. Z reguły imitacje realizowane były tylko z dwoma lub jednym akcentem pobocznym. Zdarzały się również przypadki pomyłki — realizacji wzorca H zamiast L (ale nie odwrotnie).

8.2.5. PODSUMOWANIE WYNIKÓW ANALIZY AKUSTYCZNEJ

1. Akustyczne cechy dystynktywne akcentu pobocznego oraz rdzennego związane są z samogłoską.

2. Akcenty rdzenne charakteryzują się następującymi cechami:

HL — maksimum oraz minimum globalne odpowiednio na początku/końcu frazy, spadek parametru F0 (od poziomu najwyższego do poziomu najniższego),

LH — minimum oraz maksimum globalne odpowiednio na początku/końcu frazy, wzrost parametru F0 (od poziomu najniższego do poziomu najwyższego),

LM — minimum globalne na początku frazy, wzrost parametru F0 (do poziomu średniego),

ML — minimum globalne na końcu frazy, spadek parametru F0 (od poziomu średniego do najniższego),

HM — maksimum globalne na początku frazy, spadek parametru F0 (od poziomu wysokiego do poziomu średniego),

MH — maksimum globalne na końcu frazy, wzrost parametru F0 (od poziomu średniego do najwyższego),

MM — przebieg częstotliwości równy, umieszczony w środkowym zakresie skali; przed sylabą rdzenną występuje zmiana częstotliwości, po sylabie rdzennej brak zmian,

xL — minimum globalne w obrębie frazy, spadek parametru F0 (od poziomu poniżej średniego do wartości poniżej Fmin),

LHL — minimum oraz maksimum globalne, przebieg składający się ze wzrostu parametru F0 do maksimum globalnego i spadku częstotliwości do minimum globalnego.

3. Akcenty preiktyczne:

typu H

— umieszczone są w 2/3 górnego zakresu zmian częstotliwości,

— na samogłosce występuje punkt ekstremalny przebiegu, zawierający ton wyższy niż poprzedzająca i następująca samogłoska,

— poprzedza samogłoskę o tonie równie wysokim, po której następuje spadek, poprzedzająca samogłoska ma ton niższy,

— na początku frazy samogłoska akcentowana poprzedza samogłoskę o tonie równie wysokim, po której następuje spadek częstotliwości,

— na początku frazy samogłoska poprzedza samogłoskę o tonie niższym, typu L

— umieszczone są w dolnej części 1/3 zakresu skali częstotliwości,

— na samogłosce występuje minimum lokalne, sylaby poprzedzająca i następująca znajdują się powyżej,

— na początku frazy samogłoska typu L poprzedza samogłoskę o tonie wyższym.

4. Na samogłoskach sylab rdzennych mogą wystąpić znaczne zmiany częstotliwości podstawowej (do kilkudziesięciu Hz).

5. Na samogłoskach w sylabach preiktycznych nieakcentowanych zmiany parametru F0 są z reguły niewielkie (rzędu kilkunastu Hz).

9 ZMIENNOŚĆ ILOCZASU SAMOGŁOSKOWEGO ORAZ INTENSYWNOŚCI W OBRĘBIE FRAZY

9.1. WPŁYW POZYCJI AKCENTU

Badania przeprowadzone dla różnych języków wykazały, że istotną rolę w percepcji granic frazowych, a zwłaszcza granic zdaniowych odgrywa iloczas samogłosek. Zjawisko to było przedmiotem licznych analiz percepcyjnych, w których słuchaczom prezentowano resyntetyzowane wypowiedzi o zmiennym czasie trwania wybranych samogłosek (np. Scott 1982, Gussenhoven i Rietveld 1992). Efekt wydłużenia sylab znajdujących się na końcu wypowiedzi badano głównie dla sylab akcentowanych (np. Scott 1982). Z badań Delattre (1966) wynika, że nieakcentowane sylaby na końcu wypowiedzi są tak samo długie lub dłuższe niż akcentowane w pozycji niekońcowej. Nakatani, O'Connor i Aston (1981) wykazali, że końcowe sylaby akcentowane oraz nieakcentowane ulegały w podobnym stopniu wydłużeniu. Omówienie literatury dotyczącej tego zagadnienia zawarto w pracach Berkovits (1993, 1994). Z jej badań wynika, że wydłużenie w 25% dotyczy sylab akcentowanych (w wyrazach dwusylabowych, z akcentem na pierwszej sylabie), a w 75% ostatniej sylaby nieakcentowanej (w 57% wydłużenie dotyczyło ostatniej samogłoski). Efekt wydłużenia na sylabie przedostatniej był większy, jeżeli sylaba ta była akcentowana.

Dotychczasowe badania iloczasu głosek języka polskiego uwzględniały głównie wpływ cech segmentalnych w wypowiedziach logatomowych (Frąckowiak-Richter 1973). Dla języka polskiego nie przeprowadzono do tej pory szczegółowych badań nad zależnością iloczasu od pozycji sylaby we frazie. W późnych latach 40. zagadnienie to badała eksperymentalnie Dłuska (1957). Wyniki pomiarów iloczasu samogłosek zamieszczone w pracy Frąckowiak-Richter (1973) pozwalają uznać za czynniki istotne tzw. iloczas właściwy związany ze stopniem otwarcia samogłoski, wpływ dźwięczności następującej spółgłoski oraz miejsce akcentu w wyrazie.

Dla ustalenia roli iloczasu i intensywności w wypowiedziach mowy ciągłej przeprowadzono akustyczną analizę przygotowanych specjalnych tekstów. Aby zmniejszyć liczbę czynników wpływających na strukturę suprasegmentalną wypowiedzi, przyjęto do analizy frazy o określonej budowie segmentalnej, intonacyjnej i rytmicznej. Ułożono 3 kilkuzdaniowe teksty, w które wstawiono słowa kluczowe — imiona własne: *Marek, Darek, Czarek*. Założono określoną strukturę segmentalną wyrazów: samogłoska akcentowana *a* zawsze występuje przed *r,* nieakcentowana *e* przed spółgłoską *k.* Samogłoska akcentowana występuje w 3 kontekstach: po spółgłosce nosowej *m,* zwartej dźwięcznej *d* oraz bezdźwięcznej zwarto-trącej *t*ʃ. Aby uniknąć zjawiska udźwięczniania końcowej spółgłoski wyrazu kluczowego, wyraz następujący po kluczowym rozpoczyna się od spółgłoski bezdźwięcznej. Wybrano występowanie wyrazu kluczowego w 5 różnych strukturach suprasegmentalnych określonych kontekstem wypowiedzi:

0 — nieokreślona pozycja akcentu rdzennego,
1 — niekońcowa pozycja we frazie, brak akcentu rdzennego,
2 — niekońcowa pozycja we frazie, obecność akcentu rdzennego,
3 — końcowa pozycja we frazie, brak akcentu rdzennego,

4 — końcowa pozycja we frazie, obecność akcentu rdzennego.

Przykładowo, w tekście pierwszym wyrazy kluczowe umieszczono w następującym porządku:

Tekst 1

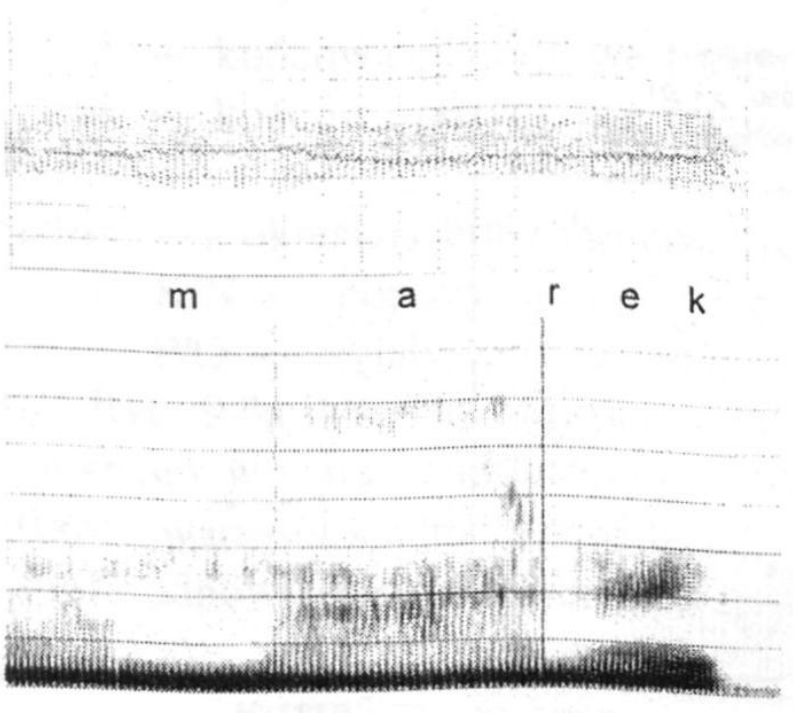
Zawsze przekręcasz moje słowa. Nie powiedziałem, że Marek (0) *sprzedaje warzywa. Powiedziałem, że teraz Marek* (1) *sprzedaje pieczywo. Wczoraj znalazłam stare zdjęcia. To jest chyba Czarek* (4). *Popatrz, jak się zmienił. Źle patrzysz. Nie w środku, tylko z boku jest widoczny Czarek* (3). *Tu jest Ewa Piotrowska, a tu Darek* (2) *Piotrowski.*

Pozostałe teksty zamieszczono w załączniku 6.

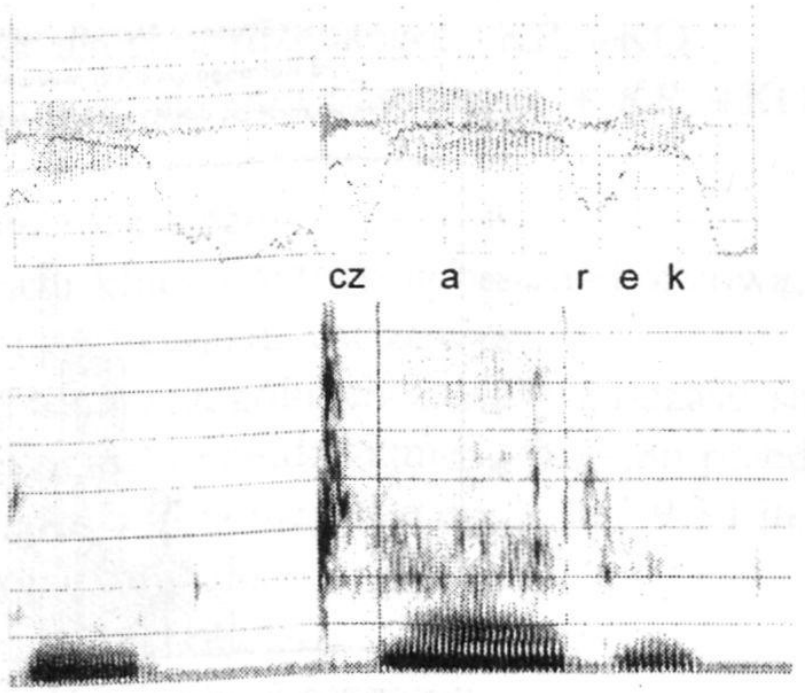
Do zbadania przyjęto następujące hipotezy:

- a. wydłużeniu ulegają dwie ostatnie sylaby we frazie, bez względu na to czy pada na nie akcent,
  b. główny efekt wydłużenia występuje na samogłoskach,
  c. efekt jest większy, jeżeli sylaba jest akcentowana.

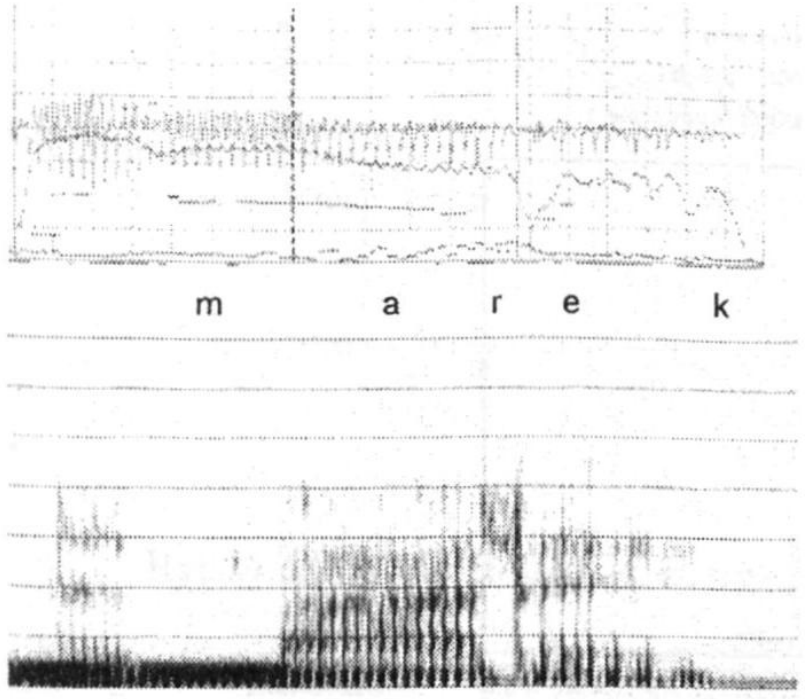
W eksperymencie wzięło udział 24 studentów wydziału fizyki. Ich zadaniem było przeczytanie 3 ułożonych do analizy tekstów. Nie udzielono im dodatkowych instrukcji dotyczących sposobu czytania ani kolejności wyboru poszczególnych tekstów. Zapisany cyfrowo materiał językowy odsłuchała osoba z przygotowaniem fonetycznym i wyeliminowała z dalszych badań (4 teksty) niepoprawnie przeczytane (mało płynnie lub z błędami). Korzystając ze spektrografu cyfrowego Kay 5500 przeprowadzono pomiary częstotliwości podstawowej, iloczasu oraz poziomu sygnału. Iloczas samogłosek określono na podstawie manualnej segmentacji widma sygnału. Dla samogłosek w wyrazach kluczowych (w imionach *Marek*, *Darek*, *Czarek*) określono następujące parametry: maksymalną wartość F0, czas trwania samogłosek i średni poziom sygnału (dB) w obrębie każdej samogłoski.



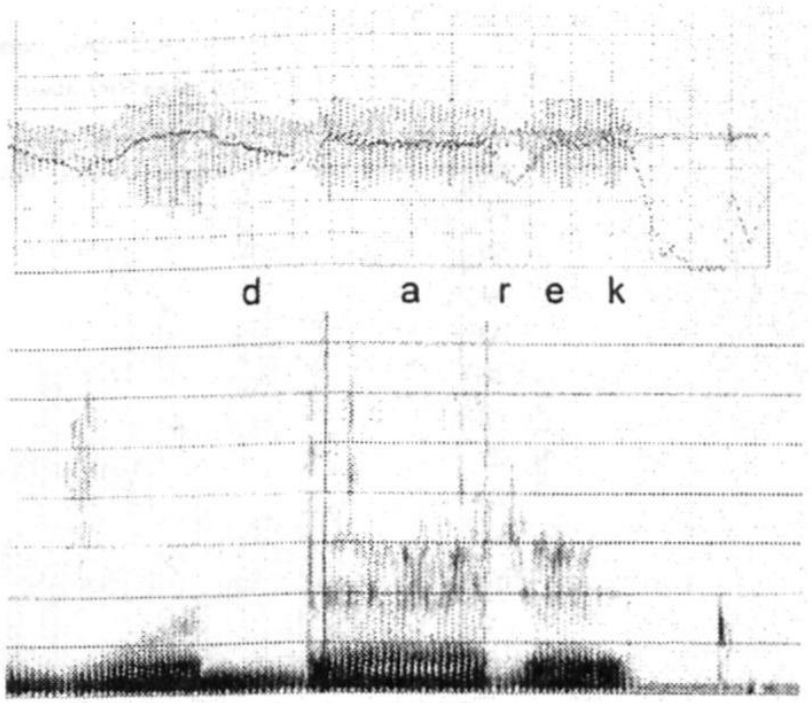
Ryc. 9.la. Oscylogram i spektrogram wyrazu *Marek*. Pozycja wyrazu: końcowa akcentowana



Ryc. 9.1b. Oscylogram i spektrogram wyrazu *Czarek*. Pozycja wyrazu: niekońcowa akcentowana

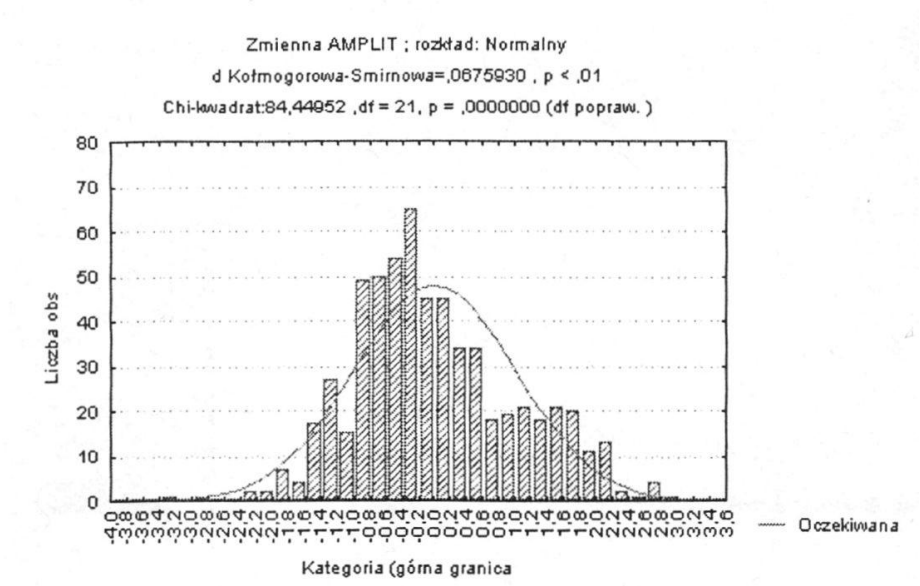


Ryc. 9.1c. Oscylogram i spektrogram wyrazu *Marek*. Pozycja wyrazu: końcowa nie-akcentowana

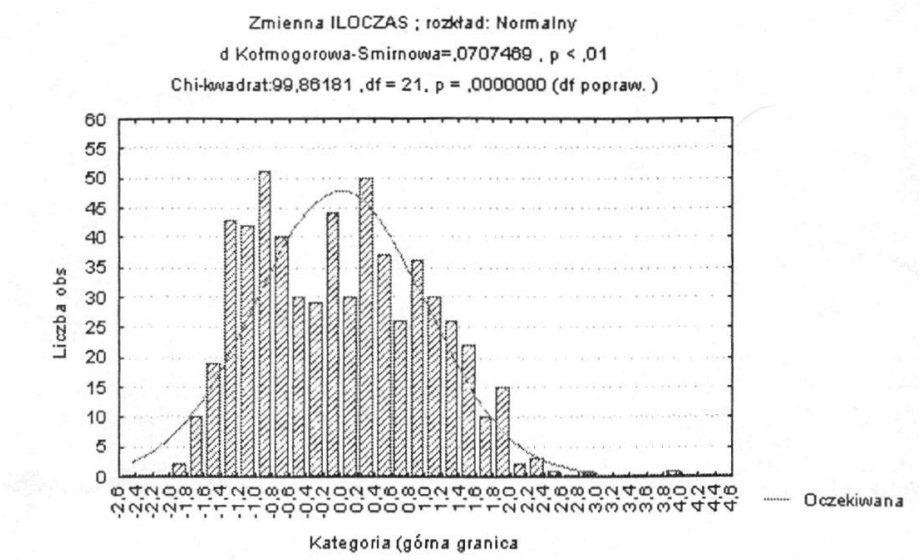


Ryc. 9.1d. Oscylogram i spektrogram wyrazu *Darek*. Pozycja wyrazu: niekońcowa nie- akcentowana

Przykładowo, na rycinach 9.1a-9.1d zilustrowano spektrogramy wypowiedzi *Marek, Czarek, Darek* w następujących kontekstach: końcowa pozycja we frazie, obecność akcentu rdzennego (ryc. 9.1a), niekońcowa pozycja we frazie, obecność akcentu rdzennego (ryc. 9.1b), końcowa pozycja we frazie, brak akcentu rdzennego (ryc. 9.1c) oraz niekońcowa pozycja we frazie, brak akcentu rdzennego (ryc. 9.1d). Zauważyć można wyraźne wydłużanie samogłosek w pozycji końcowej wyrazu kluczowego (zarówno ostatniej, jak i przedostatniej), bez względu na to, czy dotyczy ono sylaby akcentowanej.



Ryc. 9.2a. Rozkłady poziomu sygnału w wyrazach kluczowych



Ryc. 9.2b. Rozkłady iloczasu samogłosek w wyrazach kluczowych

Rozkłady poziomu sygnału i iloczasu przedstawiono na ryc.9.2 a i b. Wahania iloczasu mieszczą się w zakresie 36 - 290 ms, a średniego poziomu sygnału w przedziale 10-15 dB. Dane dla obu parametrów znormalizowano oddzielnie dla każdego głosu do wartości średniej równej zeru i odchylania standardowego równego 1.

Dla ustalenia statystycznej tendencji przeprowadzono analizę wariancji w zakresie iloczasu oraz energii dla 4 pozycji akcentu:

1 — niekońcowa pozycja we frazie, brak akcentu rdzennego: -NP, -NO

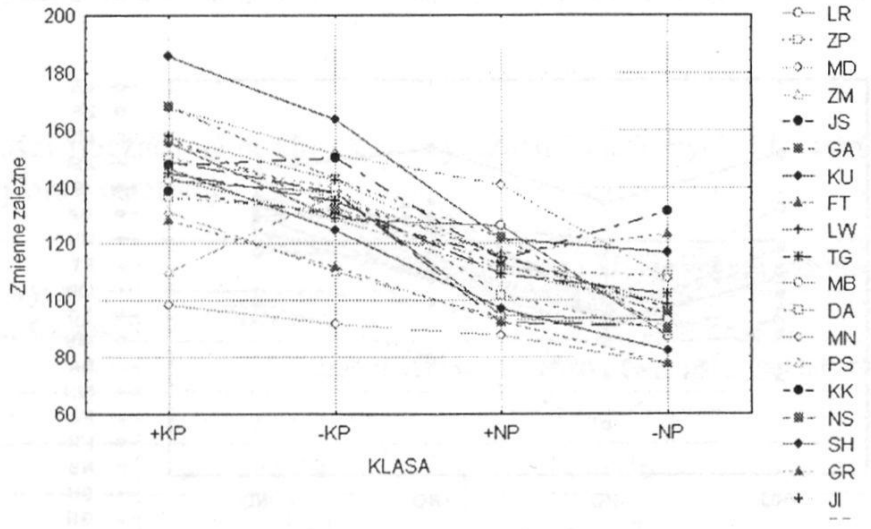
2 — niekońcowa pozycja we frazie, obecność akcentu rdzennego: + NP, +NO

3 — końcowa pozycja we frazie, brak akcentu rdzennego: -KP, -KO

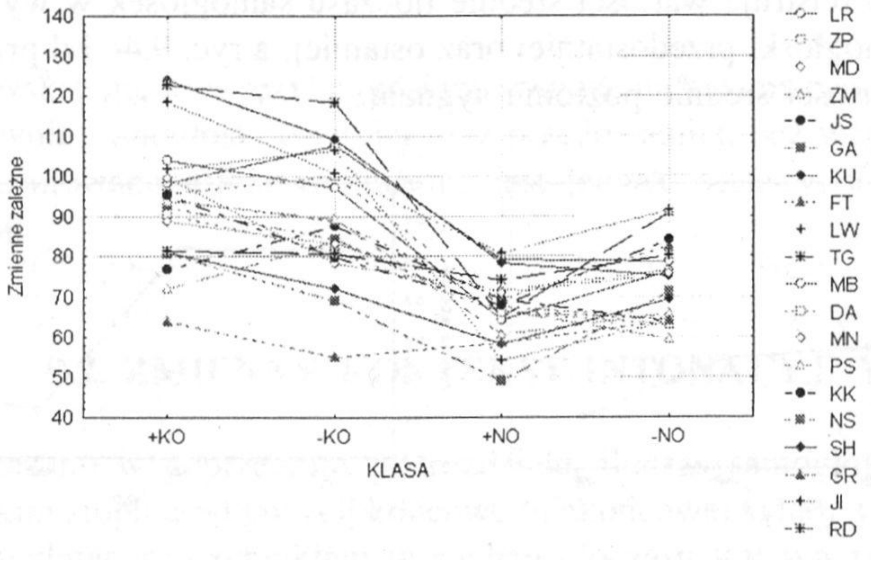
4 — końcowa pozycja we frazie, obecność akcentu rdzennego: + KP, +KO, gdzie: -/+ określają brak/obecność akcentu rdzennego,

K/N — odpowiednio pozycję wyrazu kluczowego końcową/niekońcową, P/O — sylabę przedostatnią/ostatnią w wyrazie kluczowym.

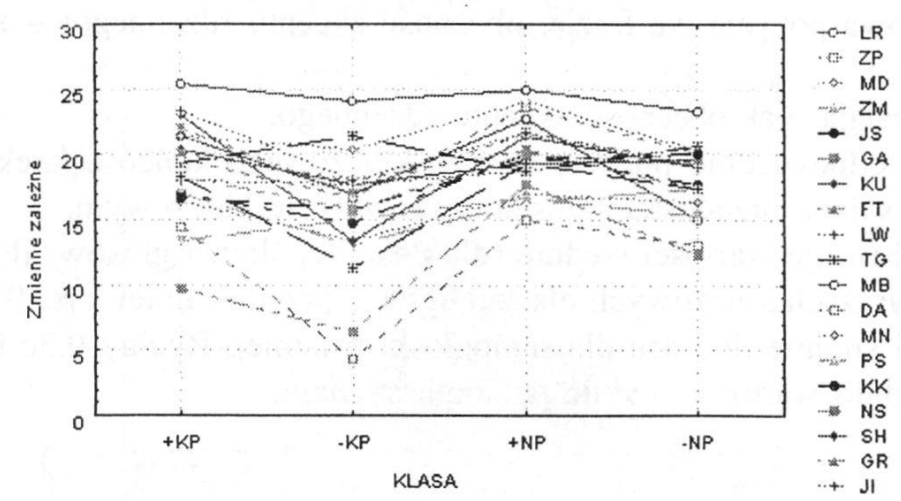
Ryc. 9.3a ilustruje wartości średnie (dla poszczególnych głosów) iloczasu samogłosek w wyrazach kluczowych dla samogłoski przedostatniej, ryc. 9.3b przedstawia wartości średnie iloczasu dla samogłoski ostatniej. Ryciny 9.3c i 9.3d ilustrują odpowiednio wartości średnie poziomu sygnału.



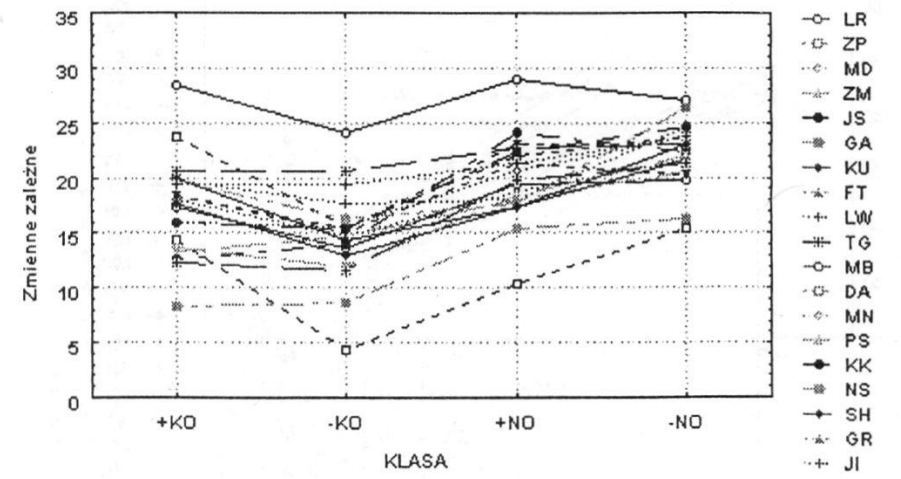
Ryc. 9.3a. Wartości średnie iloczasu dla samogłoski przedostatniej (20 głosów)



Ryc. 9.3b. Wartości średnie iloczasu dla samogłoski ostatniej (20 głosów)

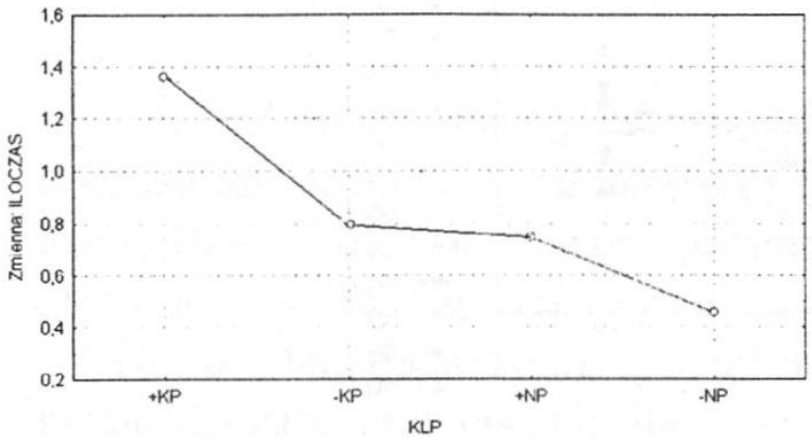


Ryc. 9.3c. Wartości średnie poziomu sygnału dla samogłoski przedostatniej (20 głosów)

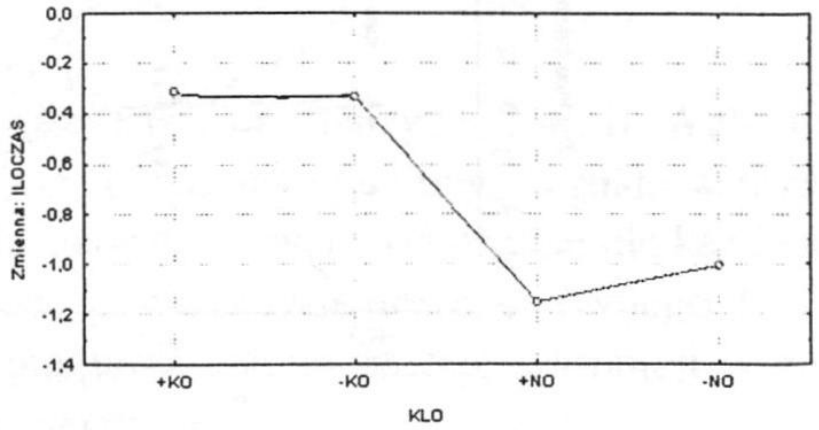


Ryc. 9.3d. Wartości średnie poziomu sygnału dla samogłoski ostatniej (20 głosów)

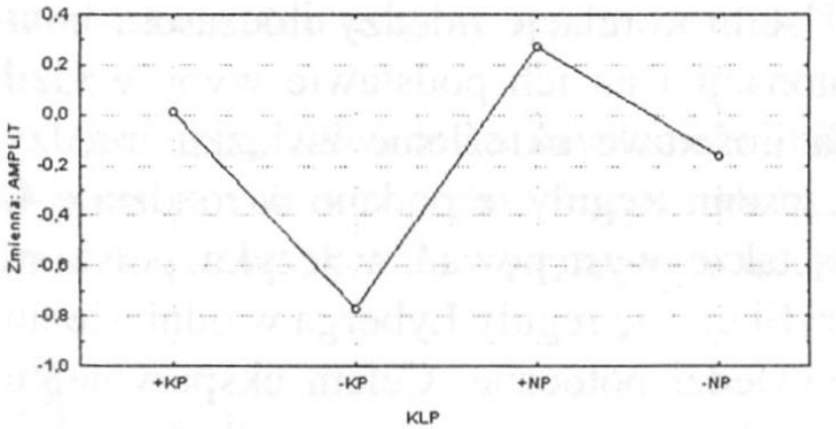
Ryc. 9.4a i b ilustrują wartości średnie iloczasu samogłosek w wyrazach kluczowych dla samogłoski przedostatniej oraz ostatniej, a ryc. 9.4c i d przedstawiają odpowiednio wartości średnie poziomu sygnału.



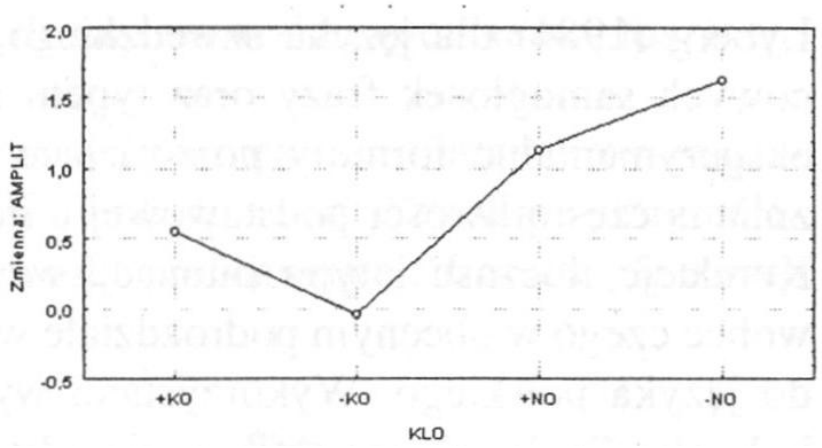
Ryc. 9.4a. Wartości średnie iloczasu dla samogłosek przedostatnich — 4 pozycje wyrazu kluczowego



Ryc. 9.4b. Wartości średnie iloczasu dla samogłosek ostatnich — 4 pozycje wyrazu kluczowego



Ryc. 9.4c. Wartości średnie poziomu sygnału dla samogłosek przedostatnich — 4 pozycje wyrazu kluczowego



Ryc. 9.4d. Wartości średnie poziomu sygnału dla samogłosek ostatnich — 4 pozycje wyrazu kluczowego

Efekt wydłużenia samogłosek końcowych przedstawia poniższe zestawienie:

Tabela 9.1 Średnie wartości iloczasu samogłosek w wyrazach kluczowych dla samogłoski przedostatniej oraz ostatniej

Kluczowy wyraz Marek, Darek, Czarek	Pozycja wyrazu kluczowego w zdaniu		
	Niekońcowa	Końcowa	Różnica
	czas trwania samogłoski (wartości standaryzowane)		
Akcent			
Samogłoska a	0,74	1,36	0,62
Samogłoska e	-1,14	-0,40	0,74
Brak akcentu			
Samogłoska a	0,40	0,80	0,40
Samogłoska e	-0,99	-0,30	0,69

Wyniki wskazują, że w języku polskim wydłużenie końcowego fragmentu frazy dotyczy głównie samogłoski ostatniej oraz przedostatniej, bez względu na to, czy jest ona akcentowana. Efekt wydłużenia jest jednak większy, jeżeli sylaba jest akcentowana.

9.2. ZMIENNY KONTEKST INTONACYJNY

Jak wykazano w poprzednim podrozdziale, iloczas samogłosek uzależniony jest w wysokim stopniu od pozycji końcowej/niekońcowej sylaby we frazie. Innym często uwzględnianym czynnikiem w analizie iloczasu jest typ zmiany częstotliwości podstawowej. Szczegółowe doświadczenia z tego zakresu przeprowadził np. Lyberg (1984) dla języka szwedzkiego. Ustalił korelacje między iloczasem końcowych samogłosek frazy oraz typem intonacji i na ich podstawie wyprowadził eksperymentalne formuły pozwalające na ilościowe określenie związku między zmianą częstotliwości podstawowej a iloczasem. Reguły te podano w rozdziale 4. Korelacje iloczasu i typu intonacji mogą także występować w języku polskim, wobec czego w obecnym podrozdziale weryfikuje się reguły Lyberga w odniesieniu do języka polskiego. Wykorzystano wypowiedzi potoczne. Celem eksperymentu było określenie roli częstotliwości podstawowej, iloczasu oraz intensywności w budowie melodycznej frazy.

Przygotowano krótkie, zróżnicowane treściowo i intonacyjnie dialogi. Umieszczono wyraz kluczowy *dobrze* w 6 kontekstach intonacyjnych związanych z rolą następujących znaków interpunkcyjnych: .,-!?() na końcu frazy[6]. Kilka wyrazów kluczowych umieszczono dla porównania w innej niż końcowa pozycji frazy. Przykładowy analizowany dialog ilustrują poniższe wypowiedzi (pozostały tekst podano w załączniku 7).

Głos A. *Jak wygląda Janek po chorobie?*

Głos B. *Janek wygląda dobrze, chociaż wydaje mi się, że zeszczuplał.*

Głos A. *A ty jak uważasz ? Janek wygląda dobrze?*

Głos B. *Myślę, że tak. Wczoraj zdałam trudny egzamin.*

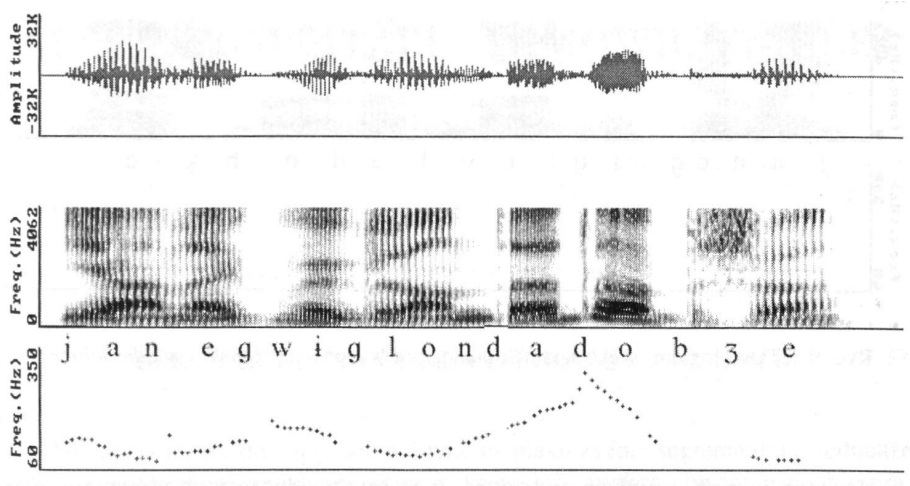
*Uzyskana ocena (**dobrze**) podniosła mnie na duchu.*

W doświadczeniu każdorazowo brały udział 2 osoby, które miały za zadanie przeprowadzenie między sobą dialogów, opierając się na dostarczonym tekście, w sposób jak najbardziej spontaniczny. W całym eksperymencie uczestniczyła grupa 24 studentów. Zapisany cyfrowo materiał językowy zweryfikował fonetyk, który eliminował z dalszej analizy wypowiedzi z błędami językowymi. Przeprowadzono analizę akustyczną wyrazów kluczowych w zakresie częstotliwości podstawowej, iloczasu samogłoskowego oraz intensywności. Na końcu frazy we wszystkich analizowanych przypadkach obserwowano znaczne zmniejszenie amplitudy sygnału, niezależnie od typu intonacji. Dla końcowych samogłosek we frazie zmiany intensywności były rzędu kilku dB, pominięto więc analizę statystyczną tego parametru.

Pomiarowi poddano iloczasy: samogłoski ostatniej i przedostatniej w wyrazie kluczowym *o* oraz *e*. Przeprowadzono analizę wariancji dla 20 powtórzeń, każdego z 6 typów intonacji. W zależności od głosu różnice iloczasów samogłosek pomiędzy poszczególnymi typami intonacji mieściły się w zakresie 0 - 20% i okazały się nieistotne statystycznie.

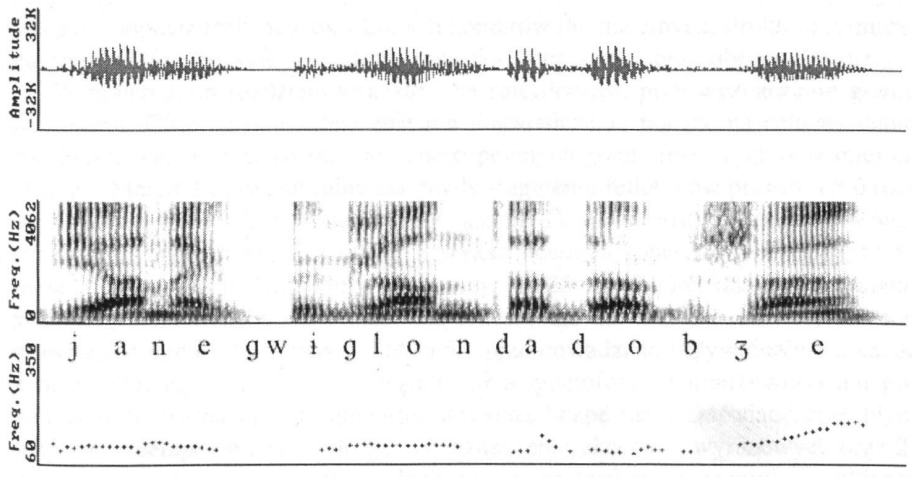
Na ryc. 9.5 a, b i c zilustrowano przykładowe zmiany częstotliwości podstawowej, w wypowiedziach: *Janek wygląda dobrze!, Janek wygląda dobrze, Janek wygląda dobrze?* Nawet wizualna ocena przykładów nie pozwala na stwierdzenie korelacji między typem intonacji i iloczasem końcowych samogłosek we frazie.

Oceniając całościowo uzyskane wyniki analizy zmienności parametrów suprasegmentalnych w obrębie frazy, należy potwierdzić, że w języku polskim relewantną cechą akcentu są zmiany częstotliwości podstawowej. Iloczas głoskowy

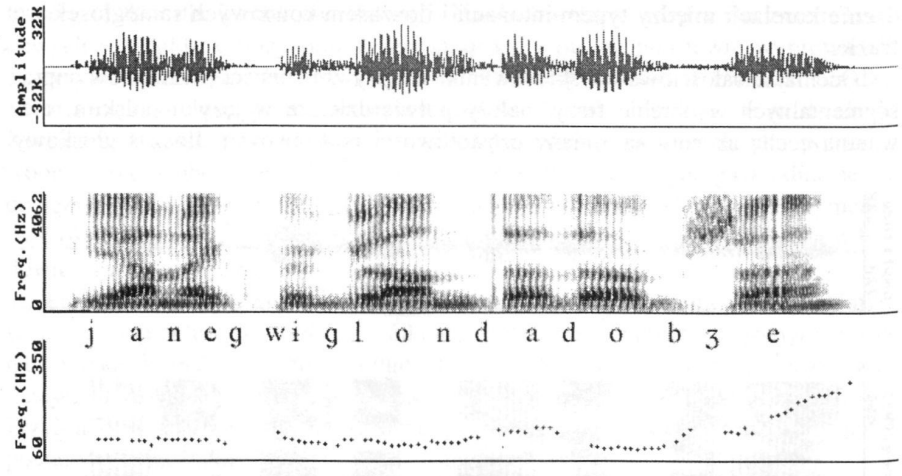


Ryc. 9.5a. Oscylogram, spektrogram i intonogram wypowiedzi *Janek wygląda dobrze!*

[6] Jedynym znakiem interpunkcyjnym w niekońcowej pozycji frazy jest oczywiście otwarcie nawiasu (.



Ryc. 9.5b. Oscylogram, spektrogram i intonogram wypowiedzi *Janek wygląda dobrze*.



Ryc. 9.5c. Oscylogram, spektrogram i intonogram wypowiedzi *Janek wygląda dobrze?*

w akcentuacji spełnia funkcję podrzędną. Jego rola uwidacznia się głównie w wyznaczaniu granicy frazowej. Zmiany intensywności są w dużym stopniu uzależnione od cech segmentalnych mowy. Rola tego parametru w sygnalizacji akcentu jest nieznaczna. Podobne rezultaty wykazano dla języka szwedzkiego, charakteryzującego się akcentem tonicznym.

10 SUPRASEGMENTALIA W MOWIE CIĄGŁEJ

10.1. PERCEPCYJNA KLASYFIKACJA AKCENTU

Percepcja mowy dokonuje się głównie na płaszczyźnie segmentalnej. Jednakże poprawne rozumienie mowy związane jest również z funkcjonowaniem cech suprasegmentalnych, które pozwalają dokonać podziału tekstu na sekwencje wyrazów stanowiących spójną całość pod względem syntaktycznym lub semantycznym. W tekście pisanym funkcję taką pełnią znaki interpunkcyjne, których obecność pozwala odbiorcy przeprowadzać podział tekstu na jednostki informacji zgodnie z intencją nadawcy. W tekście mówionym wyodrębnianie jednostek zwanych frazami jest osiągane poprzez realizację określonych konturów intonacyjnych, strukturę rytmiczną wypowiedzi (rozkład akcentów, zjawiska iloczasowe) oraz obecność pauz.

W poprzednim rozdziale wskazano na rolę iloczasu przy wyznaczaniu granic frazowych. Obecnie omówione zostanie doświadczenie mające na celu ustalenie zgodności słuchaczy w wyznaczaniu percepcyjnych granic frazowych oraz miejsca akcentu. Materiał doświadczalny stanowiły fragmenty felietonów prasowych o rozciągłości czasowej ok. 6 minut (por. załącznik 1). Tekst został odczytany w warunkach studyjnych przez 3 osoby z wykształceniem fonetycznym. W eksperymencie brały udział dwie grupy słuchaczy: 1) 25 studentów, jako tzw. „naiwni słuchacze", czyli osoby niewyszkolone fonetycznie ani lingwistycznie, 2) 5 osób z wykształceniem fonetycznym. Odsłuchy przeprowadzano indywidualnie z każdą z osób. Słuchacz samodzielnie obsługiwał magnetofon, co umożliwiało mu powracanie do wątpliwych fragmentów dowolną liczbę razy. Doświadczenie przebiegało dwuetapowo i polegało na: 1) zaznaczeniu akcentów wyrazowych oraz 2) wyznaczeniu granic frazowych. Słuchacz otrzymywał tekst nagrania, z którego usunięto wszystkie znaki przestankowe (kropki, przecinki, myślniki itd.), a duże litery, sygnalizujące początek zdania, zastąpiono małymi. Tekst nie zawierał więc żadnych informacji graficznych związanych z jego strukturą składniową. Zgodnie z podaną na wstępie instrukcją słuchacz zaznaczał:

1) Miejsce, w którym według niego przypadała granica oddzielająca dwie frazy. Frazę zdefiniowano jako odcinek tekstu dający się wyróżnić jako pewnego rodzaju całość, której elementy są ściśle powiązane ze sobą, zaś odróżniającą się od sąsiednich fragmentów tekstu. Jest ona wyznaczona przez określony przebieg intonacji, niekiedy zauważalne wydłużenie sylab końcowych lub obecność pauzy.

2) Sylaby akcentowane, to znaczy te szczególnie uwydatnione za pomocą środków prozodycznych.

Wyniki uzyskane od „naiwnych" słuchaczy poddano wstępnej ocenie, eliminując zawierające ewidentne błędy i skrajnie odbiegające od pozostałych.

Przyjęto hipotezę, że oceny stopnia zaakcentowania sylab (określonego przez liczbę słuchaczy, którzy uznali daną sylabę za akcentowaną), jak również siły granic frazowych różnią się między sobą. Wyniki przetestowano testem istotności $\chi 2$. Wartość teoretyczna testu $\chi 2$ dla jednego stopnia swobody przy $p = 0{,}05$ wynosi 3,8. Wartość ta stanowiła podstawę następującego podziału: a) jako słabo akcentowaną przyjęto sylabę ocenioną przez mniej niż 40% słuchaczy, b) sylaby, dla których uzyskano oceny w granicach 40 - 75%, uznano za średnio akcentowane, c) sylaby, które uzyskały więcej niż 75% ocen słuchaczy, uznano za silnie zaakcentowane. Podobną klasyfikację przeprowadzono dla fraz — słabą granicę określały

odpowiedzi poniżej 40%, średnią granicę wyznaczyły oceny w zakresie 40 - -75%, silną granicę oceny powyżej 75%.

Dokonano analizy ocen granic frazowych uzyskanych w grupie fonetyków. Okazało się, iż wszystkie granice uznane za silne (łącznie 75 - 100% ocen naiwnych słuchaczy) zostały jako takie wskazane również przez fonetyków. Fonetycy ocenili jako średnie 88% średnich granic w głosie MG, 95% średnich granic w głosie JI oraz odpowiednio 80% w głosie LR. Natomiast w grupie granic słabych (poniżej 40% łącznych ocen) zaledwie 45% zostało wskazanych przez fonetyków. W tej sytuacji można przyjąć, iż oceny pochodzące od słuchaczy naiwnych są wiarygodne dla granic silnych oraz średnich, a więc takich, które uzyskały ponad 40% ocen.

Dla określenia stopnia zgodności ocen podanych przez słuchaczy naiwnych oraz fonetyków posłużono się współczynnikiem zgodności ZG wyznaczonym według wzoru (por. np. Möbius 1993):

$$Z_G = \frac{2Z_{NF}}{Z_N + Z_F}$$

gdzie: Z_N — oznacza liczbę ocen uzyskanych od co najmniej 40% słuchaczy naiwnych,
Z_F — oznacza liczbę ocen uzyskanych od co najmniej 3 fonetyków,
Z_{NF} — liczbę ocen uzyskanych równocześnie od słuchaczy określonych jako N oraz jako F.

W tabeli 10.1 przedstawiono wartości współczynnika zgodności.

Tabela 10.1 Współczynniki zgodności ocen fonetyków oraz naiwnych słuchaczy

Jednostki klasyfikowane	Głos JI	Głos MG	Głos LR
Granice fraz	0,92	0,83	0,82
Akcenty	0,80	0,78	0,74

Generalnie należy stwierdzić wysoką zgodność ocen dla obu grup słuchaczy, przy czym lepszą zgodność wykazują wyniki dotyczące granic frazowych. Szczególnie wysoka wartość pojawiła się dla głosu JI. Związane jest to niewątpliwie z faktem, że podział na frazy w tym głosie okazał się bardziej wyrazisty — 56% wszystkich granic zalicza się do silnych, a tylko 23% do słabych, podczas gdy np. w głosie MG rozkład granic silnych i słabych jest niemal identyczny: 37% i 35%.

Analiza wyników dotyczących podziału na frazy na podstawie zjawisk prozodycznych pozwala stwierdzić, że słuchacze ściśle wiążą go ze strukturą syntaktyczną tekstu. W całym materiale bezbłędnie percypowano koniec zdania — wszędzie tam zaznaczono granicę silną. Poza tym granica silna oddziela od siebie zdania współrzędnie złożone, zdania nadrzędne od podrzędnych, zdania wtrącone lub równoważniki zdań, pełniąc tym samym funkcję składniową. Stwierdzono również silną zależność pomiędzy podziałem na frazy a obecnością znaków interpunkcyjnych. Miejsca, w których w tekście pisanym wystąpiły przecinki, myślniki, cudzysłowy, dwukropki i nawiasy mówcy przenieśli w jakiś sposób do odczytywanego tekstu, ponieważ słuchacze zaznaczyli tam granice silne lub średnie. Słuchacze naiwni sugerowali się niekiedy tak dalece regułami interpunkcyjnymi, że wprowadzali granicę w miejscu, w którym powinien wystąpić przecinek oddzielający zdanie podrzędne od nadrzędnego, gdy tymczasem fonetycy nie stwierdzali granicy w tym miejscu. Podział na frazy wynikający z subiektywnej interpretacji tekstu przez mówiącego, a nieuzasadniony syntaktycznie, zdarzał się wyłącznie w grupie granic słabych, a więc zaznaczanych najrzadziej. Wyniki pozwalają wnioskować, iż podział na frazy znajduje swe odbicie również w percepcji akcentu. Akcenty uznane za silne (75 - 100% łącznych ocen w obu grupach odsłuchujących) przypadają głównie na ostatni zestrój akcentowy we frazie. W głosie MG jest to 81%, a w głosie JI 89% wszystkich akcentów silnych, przy czym większość z nich przypada przed granicą silną. Przed granicą słabą, która jest często wątpliwa lub słabo

percypowana przez słuchaczy, tylko sporadycznie pojawia się akcent silny. Z kolei akcenty słabe (poniżej 40% łącznych ocen), a więc najgorzej percypowane przez słuchaczy, pojawiają się bardzo rzadko w ostatnim wyrazie frazy: np. w głosie LR — 6%, w głosie JI — 3% wszystkich akcentów słabych, przy czym nigdy nie występują przed silną granicą. Tak więc wyznacznikiem końca frazy okazuje się być między innymi silny lub średni akcent padający na ostatni wyraz we frazie. W załączniku 1 podano rozkład wyznaczonych granic frazowych dla głosu JI.

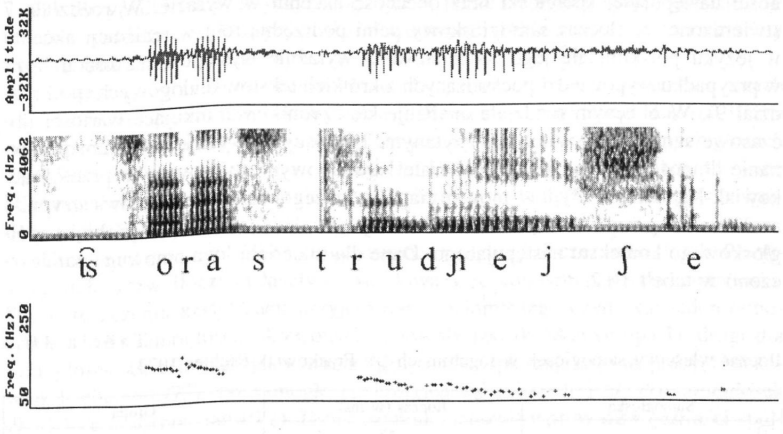
10.2. ANALIZA AKUSTYCZNA STRUKTUR MELODYCZNYCH

10.2.1. WYZNACZNIKI GRANICY FRAZOWEJ

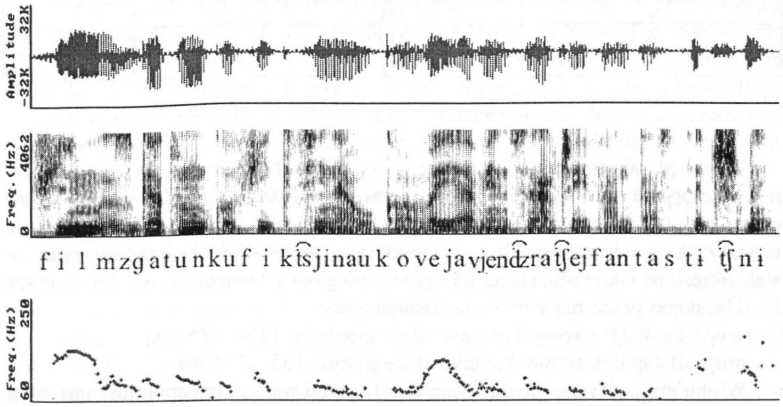
Z analiz percepcyjnych wynika, że słuchacze posiadają indywidualne progi percepcji i identyfikują cechy akustyczne wypowiedzi w różny sposób. Trudno więc spodziewać się kategorycznych podziałów w klasyfikacji percepcyjnej i przypuszczalnie w akustycznej. Granice frazowe tworzą pewne kontinuum, poczynając od najbardziej wyraźnej sytuacji wydłużenia sylab końcowych (które może być rzędu nawet 50%), charakterystycznej zmiany wartości częstotliwości podstawowej na końcu frazy (np. szybki wzrost lub spadek ponad oktawę) oraz pauzy akustycznej, do braku zmiany parametrów akustycznych. Istotność poszczególnych parametrów wpływających na percepcyjną segmentację wypowiedzi może być specyficzna dla różnych języków. Dla języka hebrajskiego stwierdzono (Berkovits 1994), że najistotniejszym czynnikiem wpływającym na podział na frazy jest wydłużenie sylab znajdujących się bezpośrednio przed granicą frazy oraz przebieg parametru F0. Dla języka angielskiego Lea (1979) wykazał, że 95% pauz równych lub dłuższych od 350 ms to faktyczne granice między zdaniami lub frazami.

Wstępna analiza akustyczna wybranych do analizy tekstów (załącznik 1) polegała na wyznaczeniu rozkładu pauz akustycznych (przerw dłuższych od 350 ms) i zbadaniu ich korelacji z granicami frazowymi. W analizowanym materiale w ponad 50% przypadków stwierdzono wystąpienie granicy silnej, mimo braku przerwy akustycznej. We wszystkich przypadkach, w których po frazie wystąpiła pauza, słuchacze zaznaczyli występującą granicę jako silną. Na ryc. 10.1 zilustrowano granicę frazową, przed którą obserwuje się znaczne zmniejszenie poziomu sygnału (ponad 20 dB) oraz zanik akustyczny ostatniej sylaby (na końcu frazy *..coraz trudniejsze…* zauważa się tylko krótki fragment samogłoski *e)*. Granicę tę słuchacze oznaczyli również jako silną.

Na ryc. 10.2 przedstawiono przykładowo oscylogram, spektrogram i intonogram wypowiedzi .. *film z gatunku fikcji naukowej, a więc raczej fantastyczny,* w której słuchacze wyznaczyli granicę po sylabie *wej*. Między sylabą *wej* i sylabą *a* brak jest jakiejkolwiek pauzy. Również w przebiegu intensywności obserwuje się małą zmienność. Na sylabie *ko* występuje spadek częstotliwości do Fmin (90 Hz). Na samogłosce *e* w sylabie *wej* zauważa się wzrost częstotliwości (w zakresie 115- 170 Hz), na spółgłosce *j* występuje spadek częstotliwości (w zakresie 170- - 120 Hz). Na samogłosce *a* (po granicy frazowej) wartość częstotliwości zmienia się w zakresie 120- 110 Hz. Obserwuje się tutaj typową granicę utworzoną przez przebieg parametru F0, tzw. wzrost kontynuacyjny („continuation rise”). Spodziewać się więc można, że w przypadku braku pauzy istotną rolę w podziale wypowiedzi na frazy odgrywają również inne cechy akustyczne (np. przebieg częstotliwości podstawowej oraz iloczas).



Ryc. 10.1. Oscylogram, spektrogram i intonogram wypowiedzi *coraz trudniejsze*



Ryc. 10.2. Oscylogram, spektrogram i intonogram wypowiedzi *film z gatunku fikcji naukowej, a więc raczej fantastyczny*

10.2.2. ILOCZAS SAMOGŁOSKOWY

Wyniki pomiarów iloczasu samogłosek polskich dla wyrazów izolowanych zamieszczone w pracy Frąckowiak-Richter (1973) pozwalają uznać za czynniki istotne tzw. iloczas właściwy związany ze stopniem otwarcia samogłoski, wpływ dźwięczności następującej spółgłoski oraz obecność akcentu w wyrazie. W rozdziale 7 stwierdzono, że iloczas samogłoskowy pełni podrzędną rolę w realizacji akcentu w języku polskim, ale jego zróżnicowania wyraźnie występują na końcu frazy w przypadku wypowiedzi pochodzących z krótkich tekstów dialogowych (por. rozdział 9). W obecnym rozdziale analizuje się czynniki warunkujące wartości iloczasowe samogłosek w tekście czytanym. Interesujące wydaje się zatem porównanie długości samogłosek w materiale logatomowym opracowanym przez Frąckowiak-Richter (1973) i w mowie ciągłej. Szczególnie istotne są dwa czynniki analizowane przez autorkę: iloczas właściwy oraz dźwięczność (lub jej brak) spółgłoskowego kontekstu następującego. Dane dla materiału logatomowego zamieszczono w tabeli 10.2.

Tabela 10.2 Iloczas właściwy samogłosek w logatomach (za Frąckowiak-Richter 1973)

Samogłoska	Iloczas (w ms)	Grupa
i	78	1
ɨ	90	1
u	88	1
a	124	2
o	110	2
e	111	2

Tak więc zakres zmian czasu trwania samogłosek w grupie 1 wynosi 78 - 90 ms, w grupie 2 wynosi 110- 124 ms. Frąckowiak-Richter (1973) jako następny czynnik badała wpływ dźwięczności spółgłoski następującej. Realizowany jest on również na dwóch poziomach — samogłoska przed spółgłoską bezdźwięczną została określona jako krótka (grupa I), przed spółgłoską dźwięczną jako długa (grupa II). Uzyskane przez nią wyniki są następujące:

grupa I (spółgłoskowy kontekst bezdźwięczny): 115-125 ms,

grupa II (spółgłoskowy kontekst dźwięczny): 135- 175 ms.

W obu grupach samogłoski występowały przed wszystkimi spółgłoskami związanymi z określonym miejscem artykulacji: dwuwargowymi, wargowo-zębowymi, przednio-językowo zębowymi, przednio-językowo dziąsłowymi, środkowo-językowymi oraz tylnojęzykowymi.

Pomiary czasu trwania wszystkich samogłosek w materiale własnym, występujących w percepcyjnie ocenionych tekstach, przeprowadzono z wykorzystaniem spektrografu cyfrowego KAY 5500 oraz karty dźwiękowej typu Sound Blaster. Operator — fonetyk wyznaczał ręcznie granicę głoski na podstawie obserwacji przebiegu czasowego z równoczesnym odtwarzaniem dźwiękowym wybranego fragmentu. Dokładność pomiaru wynosiła 1 ms. Przy niektórych połączeniach głoskowych zdarzało się, iż ustalenie granicy było praktycznie niemożliwe (por. np. Kvale 1993). Przypadki takie dotyczyły najczęściej połączeń samogłoski z głoskami /j/ lub /w/, niekiedy z /1/ lub spółgłoskami nosowymi. Takie nierozdzielne fragmenty sygnału potraktowano jako dyftongi (por. np. Jassem 1973).

Szczególny przypadek stanowiły nie dające się rozdzielić połączenia trzech głosek, np. /eja/, /owo/ itp., które zaklasyfikowano jako tryftongi. Łączna liczba uzyskanych z tekstu segmentów wokalicznych wyniosła w poszczególnych głosach 599, 595 i 537 (dane dla tekstu z załącznika 1). Różnice liczebności spowodowane są różnymi relacjami pomiędzy liczbą monoftongów, dyftongów i tryftongów w każdym z głosów.

Uwzględniono cztery czynniki wywierające wpływ na iloczas samogłosek. Pierwszy z nich to tzw. **iloczas właściwy,** skorelowany ze stopniem otwarcia samogłoski. Dla uproszczenia klasyfikacji przyjęto dwa poziomy tego czynnika: jeden odnoszący się do samogłosek określonych umownie jako krótkie (grupa I), drugi dla samogłosek określonych jako długie (grupa II). Opierając się na pracy Frąckowiak-Richter (1973), do samogłosek krótkich zaliczono /i i u/, do samogłosek długich /e a o/ (por. tab. 10.2). Drugi czynnik stanowił **wpływ dźwięczności** spółgłoski następującej. Realizowany on jest również na dwóch poziomach I i II, zgodnie z koncepcją Frąckowiak-Richter.

Trzeci czynnik stanowił **wpływ akcentu** na iloczas samogłosek przejawiający się względnie dłuższym czasem trwania w pozycji akcentowanej w porównaniu z pozycją nieakcentowaną. W związku z tym wprowadzono dwa poziomy tego czynnika: 1 — samogłoska krótka — nieakcentowana, 2 — samogłoska długa — akcentowana. Określenie samogłoski jako akcentowanej lub nieakcentowanej przyjęto na podstawie wyników odsłuchów. Czwarty z uwzględnionych czynników stanowiła **pozycja we frazie:** I — niekońcowa, 2 — przedostatnia oraz 3 — ostatnia. Granice fraz, tak samo jak miejsce akcentu, zostały przyjęte na podstawie odsłuchów.

Wstępna analiza statystyczna przeprowadzona dla jednego głosu (LR) wykazała, że drugi z czynników — wpływ dźwięczności następującej spółgłoski, okazał się całkowicie nieistotny. Średni iloczas samogłoski przed spółgłoską bezdźwięczną wyniósł 97 ms, zaś przed spółgłoską dźwięczną 97,8 ms. Wartość statystyki F uzyskana w analizie wariancji wyniosła 0,04817 przy $p < 0,83$. Uwzględnianie tego czynnika dla pozostałego materiału uznano za bezcelowe. Istotne natomiast okazało się uwzględnienie dodatkowego czynnika: **budowy sylaby końcowej we frazie.**

Ostatecznie przyjęta klasyfikacja segmentów dla całego materiału doświadczalnego (trzy głosy) opiera się na uwzględnieniu czterech czynników realizowanych na dwóch, trzech lub czterech poziomach, a mianowicie:

Czynnik I: iloczas właściwy segmentu fonetycznego

poziomy: 1 — samogłoska krótka, tj. i ɨ u,

2 — samogłoska długa, tj. e a o,

3 — dyftong,

4 — tryftong.

Czynnik II: miejsce akcentu

poziomy: 1 — samogłoska nieakcentowana,

2 — samogłoska akcentowana.

Czynnik III: pozycja sylaby we frazie

poziomy: 1 — sylaba nie będąca końcową ani przedostatnią we frazie,

2 — akcentowana przedostatnia sylaba we frazie,

3 — końcowa sylaba frazy.

Trzy poziomy czynnika III odnosiły się do segmentów monoftongicznych. W przypadku dyftongów i tryftongów przyjęto dwa poziomy — poziom 2 dotyczył zarówno sylaby końcowej, jak i przedostatniej we frazie.

Czynnik IV: budowa sylaby końcowej we frazie

poziomy: 1 — sylaba zamknięta (zakończona spółgłoską),

2 — sylaba otwarta (zakończona samogłoską).

Czynniki I - III odnoszą się do samogłosek występujących w dowolnym miejscu frazy, czynnik IV dotyczy wyłącznie sylaby wygłosowej we frazie. Czynniki II i III związane są z oceną subiektywną dokonaną przez słuchaczy uczestniczących w doświadczeniu. Ponieważ analiza zgodności ocen pozwoliła przyjąć trzy stopnie dla określenia siły akcentu i granicy frazowej, ustalenia te wykorzystano przy klasyfikacji samogłosek. Analizowaną sylabę uznawano za akcentowaną, jeśli w przyjętej trzystopniowej skali została określona jako posiadająca akcent silny (od 75% do 100% zgodnych ocen) lub średni (od 40% do 74% zgodnych ocen). Samogłoski, które uzyskały poniżej 40% wskazań (akcent słaby) zostały zaliczone do nieakcentowanych. Podobne zasady obowiązywały dla granic frazowych. Przyjęty sposób klasyfikacji pozwolił wyłonić 20 klas segmentów wokalicznych występujących w badanym materiale (tab. 10.3).

Tabela 10.3 Klasy segmentów wokalicznych

Nr klasy	Samogłoska	Sylaba
1	2	3
1	krótka nieakcentowana	niekońcowa i nieprzedostatnia we frazie
2	długa nieakcentowana	niekońcowa i nieprzedostatnia we frazie
3	krótka akcentowana	niekońcowa i nieprzedostatnia we frazie
4	długa akcentowana	niekońcowa i nieprzedostatnia we frazie
5	długa akcentowana	końcowa, zamknięta
6	długa nieakcentowana	końcowa, otwarta
7	krótka akcentowana	końcowa, zamknięta
8	krótka nieakcentowana	końcowa, otwarta
9	długa nieakcentowana	końcowa, zamknięta
10	krótka nieakcentowana	końcowa, zamknięta

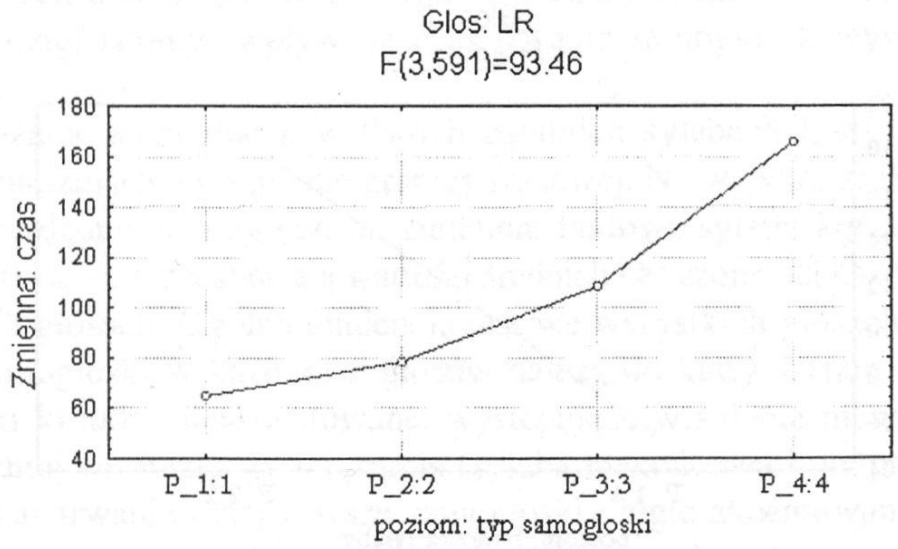
1	2	3
II	krótka, akcentowana	przedostatnia
12	długa, akcentowana	przedostatnia
13	dyftong akcentowany	przedostatnia
14	dyftong akcentowany	niekońcowa i nieprzedostatnia
15	dyftong nieakcentowany	końcowa
16	dyftong nieakcentowany	niekońcowa i nieprzedostatnia
17	tryftong akcentowany	przedostatnia
18	tryftong akcentowany	niekońcowa i nieprzedostatnia
19	tryftong nieakcentowany	końcowa
20	tryftong nieakcentowany	niekońcowa i nieprzedostatnia

Rozkłady iloczasu samogłosek w analizowanych tekstach są zbliżone do normalnych. Dla każdego głosu i każdego czynnika zmienności przeprowadzono analizę wariancji. Uzyskano następujące wyniki:

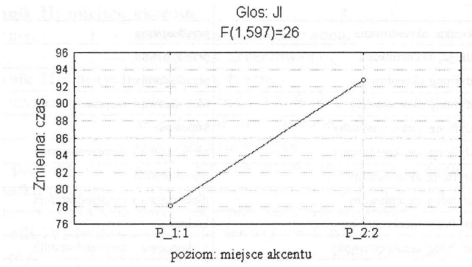
1. Zmienna I: iloczas właściwy

Ryc. 10.3 przedstawia średnie wartości uzyskane dla poszczególnych poziomów czynnika w głosie LR. Taka sama tendencja zaznaczyła się w pozostałych głosach.

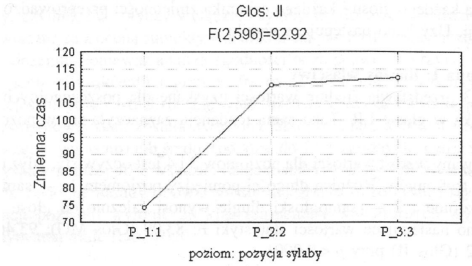
Zdecydowany wzrost wartości dla poziomów 3 i 4 jest oczywisty, gdyż dotyczy dyftongów i tryftongów. Różnica długości pomiędzy poziomem 1 — samogłoski krótkie i poziomem 2 — samogłoski długie wynosi zależnie od głosu 20 - 30 ms. Uzyskano następujące wartości statystyki F: 83,09 (Głos MG); 93,46 (Głos LR) i 216,62 (Głos JI) przy $p< 0,0001$.



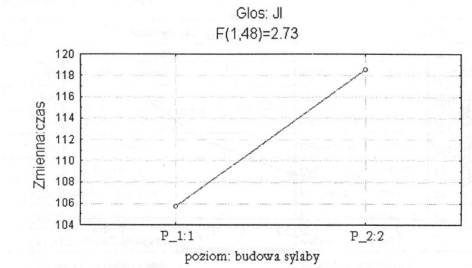
Ryc. 10.3. Wyniki analizy wariancji. Poziom: typ samogłoski



Ryc. 10.4. Wyniki analizy wariancji. Poziom: miejsce akcentu



Ryc. 10.5. Wyniki analizy wariancji. Poziom: pozycja sylaby



Ryc. 10.6. Wyniki analizy wariancji. Poziom: budowa sylaby

2. Zmienna II: miejsce akcentu

Ryc. 10.4 przedstawia średnie wartości uzyskane dla dwóch poziomów czynnika w głosie JI. W dwóch głosach samogłoski akcentowane dłuższe są od nieak-

centowanych o 15 ms, w jednym głosie o 17 ms. Uzyskane wartości F wynoszą: 26,55 (Głos MG); 26,75 (Głos JI) i 40,24 (Głos LR) przy $p < 0{,}0001$.

3. Zmienna III: pozycja sylaby we frazie

We wszystkich głosach widoczne jest znaczne wydłużanie samogłosek w sylabie przedostatniej oraz ostatniej we frazie (pozycja 2 i 3) w porównaniu z samogłoskami w sylabach wcześniejszych (pozycja 1). Różnice wartości średnich pomiędzy grupami 1 i 2 wynoszące 35 ms (Głos MG), 38 ms (Głos JI), 50 ms (Głos LR) znacznie przewyższają różnice średnich dla zmiennej miejsca akcentu. Dane te świadczą o silnym wpływie granicy frazowej na długość samogłosek. Średnia wartość iloczasu samogłoskowego w sylabie końcowej frazy (pozycja 3) jest zbliżona do średniej w sylabie przedostatniej lub ją nieznacznie przekracza (np. w głosie JI ryc. 10.5). Występujące we wszystkich głosach zdecydowane wydłużanie przedostatniej oraz ostatniej samogłoski we frazie niewątpliwie wpływało na decyzję słuchaczy przy wyznaczaniu granic frazowych. Uzyskane wartości F dla zmiennej III wynoszą: 39,13; 92,92 i 142,94 przy $p< 0{,}0001$.

4. Zmienna IV: budowa sylaby wygłosowej we frazie

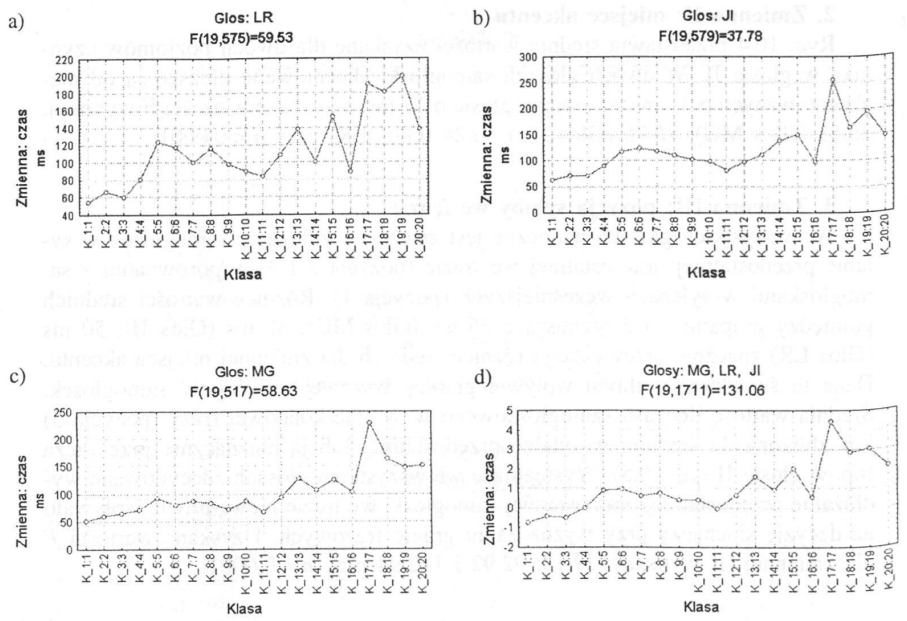
W głosach JI (ryc. 10.6) i LR samogłoska w sylabie otwartej (typ 2) wykazuje dłuższy czas trwania niż w sylabie zamkniętej (typ 1) średnio o 12 ms i 15 ms, lecz statystycznie jest to różnica nieistotna w przypadku głosu JI, zaś na granicy istotności w przypadku głosu LR. Dla JI wartość F wynosi 2,73 przy $p < 0{,}0151$, dla LR $F = 5{,}94$ przy $p < 0{,}0185$. Całkowicie nieistotny okazał się ten czynnik dla głosu MG, w którym różnica pomiędzy średnimi wyniosła 0,5 ms, a $F = 0$ przy $p < 0{,}9475$.

Z czterech uwzględnionych czynników zdecydowanie największy (statystycznie najbardziej istotny) wpływ na czas trwania samogłosek wywiera jej pozycja we frazie.

Wydłużanie samogłosek w dwóch ostatnich sylabach frazy należy uznać za cechę sygnalizującą wystąpienie granicy frazowej. Największe zróżnicowanie osobnicze stwierdzono ze względu na zmienną: budowa sylaby końcowej we frazie.

Ryc. 10.7a -c przedstawiają wartości średnich obliczone dla każdej z klas w poszczególnych głosach. Ogólna tendencja jest we wszystkich głosach taka sama. Najkrótsze samogłoski w każdym z głosów należą do klasy 1 (por. tab. 10.3). Są to samogłoski krótkie, nieakcentowane, występujące w sylabie niekońcowej i w nieprzedostatniej we frazie. W tej grupie (sylaba niekońcowa i nie przedostatnia) najdłuższy czas trwania mają zawsze samogłoski długie akcentowane (klasa 4).

Takie same relacje długości w badanych głosach dotyczą grupy samogłosek w sylabie przedostatniej akcentowanej (klasy 11 i 12) — średnia dla samogłoski



Ryc. 10.7. Wartości średnie iloczasu samogłosek: a) dla głosu LR, b) dla głosu JI, c) dla głosu MG, d) wartości standaryzowane dla 3 głosów łącznie

krótkiej jest niższa od średniej dla samogłoski długiej. W grupie dyftongów (klasy 13-16) najmniejszy czas trwania wykazują samogłoski nieakcentowane, występujące w sylabie niekońcowej i nieprzedostatniej.

Aby wyeliminować różnice wynikające z indywidualnego tempa mowy i porównać wyniki analiz dla kilku osób, przeprowadzono normalizację iloczasów głosek. Dane znormalizowano do wartości, których rozkłady charakteryzują się zerową średnią wartością oraz jednostkowym odchyleniem standardowym. Ryc. 10.7d przedstawia standaryzowane dla 3 głosów łącznie wartości średniego iloczasu w 20 klasach. Dane z wykresów pozwalają zaobserwować oddziaływanie poszczególnych zmiennych na iloczas samogłosek. I tak:

1. Średnie dla samogłosek w sylabie przedostatniej lub końcowej są wyższe od średnich w sylabach pozostałych: klasa 1 < klasa 8 i klasa 10, klasa 2 < klasa 6 i klasa 9, klasa 3 < klasa 7 i klasa 11, klasa 4 < klasa 5 i klasa 12, klasa 14 < klasa 13, klasa 16 < klasa 15, klasa 18 < klasa 17, klasa 20 < klasa 19.

2. Średnie dla samogłosek krótkich są niższe od średnich dla samogłosek długich, przy zachowaniu takich samych pozostałych warunków: klasa 1 < klasa 2, klasa 3 < klasa 4, klasa 7 < klasa 5, klasa 11 < klasa 12.

3. Średnie dla samogłosek nieakcentowanych są niższe od średnich dla samogłosek akcentowanych: klasa 1 < klasa 3, klasa 2 < klasa 4, klasa 9 < klasa 5, klasa 10 < klasa 7, klasa 16 < klasa 14, klasa 20 < klasa 18.

4. Średnie dla samogłosek w sylabie wygłosowej zamkniętej są niższe od średnich dla samogłosek w sylabie otwartej: klasa 9 < klasa 6, klasa 10 < klasa 8.

Uzyskane wyniki świadczą, iż w materiale doświadczalnym największy wpływ na iloczas samogłosek wywiera bliskość granicy frazowej. Wydłużanie samogłosek w sylabach ostatniej i przedostatniej we frazie pełni dla słuchaczy funkcję informacyjną o syntaktycznej i semantycznej strukturze wypowiedzi.

10.2.3. AKCENT RDZENNY

Percepcyjna segmentacja tekstu polegająca wyłącznie na lokalizacji granicy frazowej oraz akcentu jest niewystarczająca do klasyfikacji akustycznej akcentu rdzen-

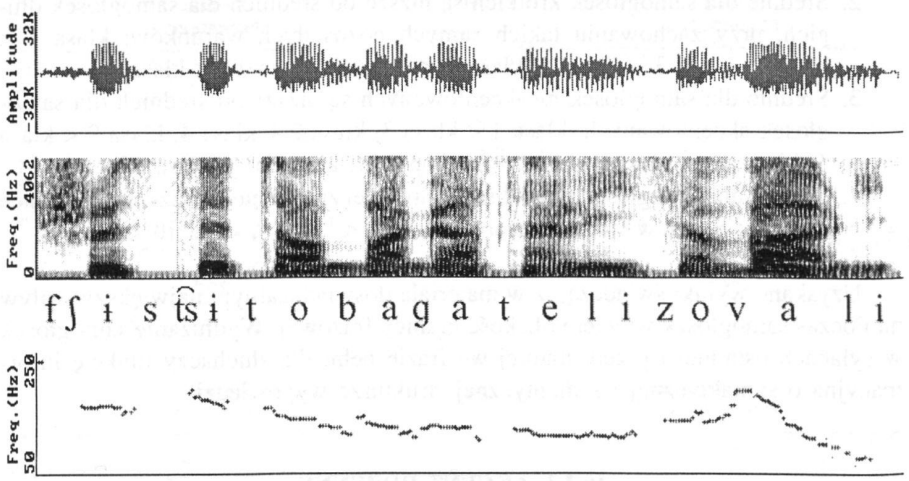
nego. Przeprowadzono dodatkowe doświadczenie odsłuchowe, w którym fonetycy identyfikowali intonację w końcowym fragmencie frazy, tj. intonację rdzenną

— od ostatniej sylaby akcentowanej do końca frazy. W celu ułatwienia odsłuchu wprowadzono w miejscu wystąpienia granicy frazowej 6-sekundowe przerwy. Wprowadzenie ciszy po granicy frazowej ułatwiło koncentrację uwagi słuchacza i zwiększyło jego wrażliwość percepcyjną. W większości przypadków klasyfikacja usłyszanych intonacji pokrywała się z wynikami analiz akustycznych, te przebiegi które oznaczono jako rosnące charakteryzowały się rosnącym przebiegiem parametru F0, przebiegi opadające identyfikowano jako intonacje opadające. W badanym materiale wyznaczono łącznie (w 3 głosach) 432 granice frazowe. Analiza percepcyjna oraz akustyczna pozwoliła wyznaczyć udział procentowy poszczególnych typów: (LM oraz LH) — 38% , (HL oraz ML) — 36%, MH — 10%, HM — 4%, MM — 6%, xL — 3%, LHL — 3 przypadki.

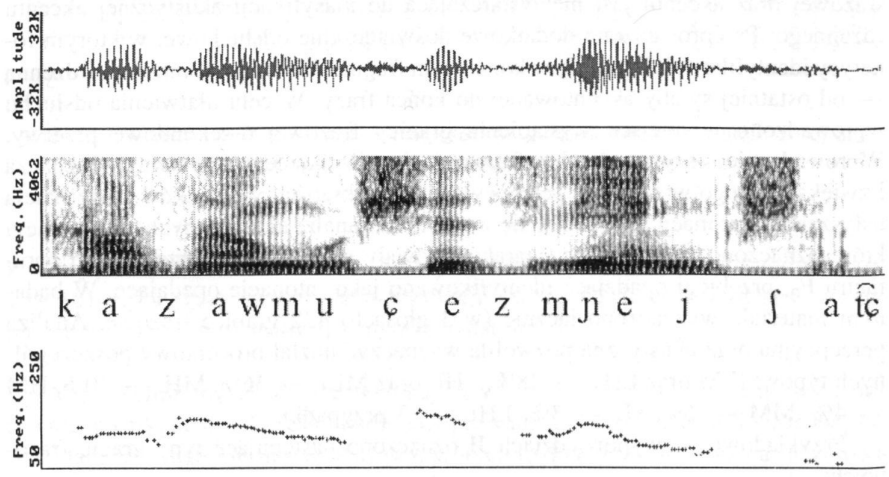
Przykładowo, w wypowiedziach JI oznaczono następujące typy granic frazowych:

Lecz mimo tych przygód **(MH)** *żył nadal* **(ML),** *choć wciąż się zmniejszał* **(xL),** *dzięki czemu* **(LM)** *jest obecnie jednym z atomów* **(LH),** *który tym się różni od pozostałych* **(HL),** *że ma wyraz twarzy* **(LH)** *ręce* **(ML)** *i nogi* **(ML).**

Na ryc. 10.8 przedstawiono przykład akcentu rdzennego HL we frazie *Wszyscy to bagatelizowali*. Na sylabach *wali* obserwuje się znaczny spadek częstotliwości podstawowej (ponad oktawę) od wartości maksymalnej we frazie wynoszącej 220 Hz w stosunku do wartości minimalnej wynoszącej 70 Hz.



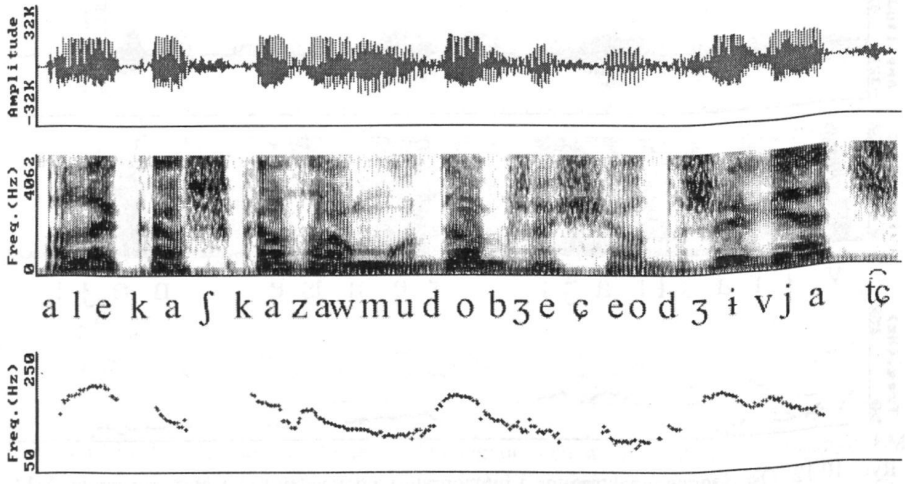
Ryc. 10.8. Oscylogram, spektrogram i intonogram wypowiedzi z akcentem rdzennym HL *wszyscy to bagatelizowali*



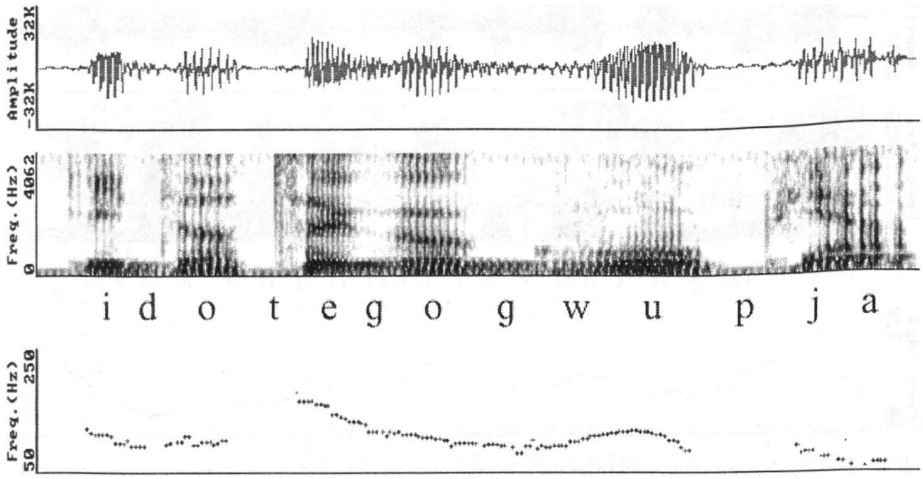
Ryc. 10.9. Oscylogram, spektrogram i intonogram wypowiedzi z akcentem rdzennym ML *kazał mu się zmniejszać*

Rycina 10.9 ilustruje akcent rdzenny typu ML zrealizowany we frazie *...kazał mu się zmniejszać.* Na sylabach *zmniejszać* występuje spadek częstotliwości podstawowej rzędu kilkudziesięciu Hz. Na ostatniej sylabie *szać* zauważa się bardzo niskie wartości częstotliwości podstawowej (zmiany długości okresu są nieregularne).

Na ryc. 10.10 przedstawiono przykład intonacji opadającej wysokiej HM we frazie *...a lekarz kazał mu dobrze się odżywiać...* Na sylabie *ży* częstotliwość pod-



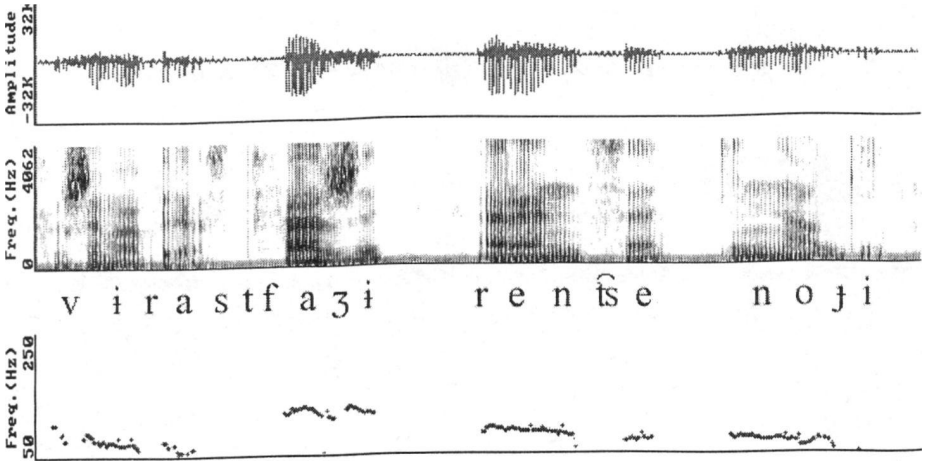
Ryc. 10.10. Oscylogram, spektrogram i intonogram wypowiedzi z akcentem rdzennym HM *a lekarz kazał mu dobrze się odżywiać*



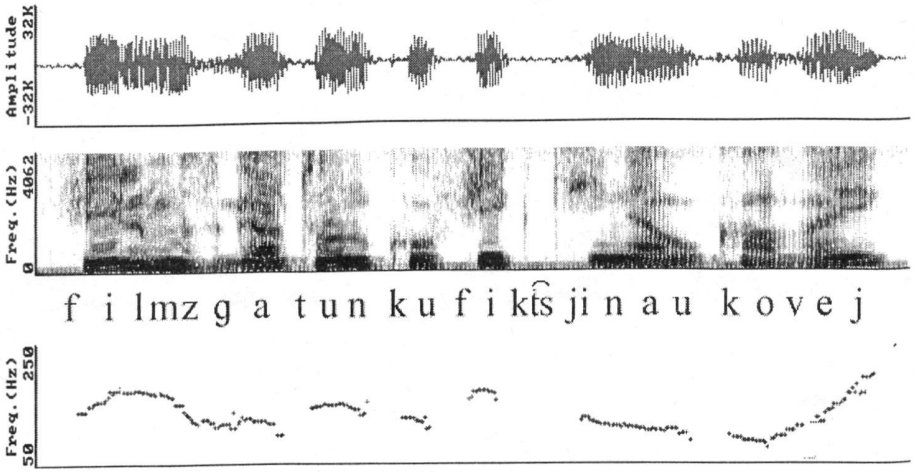
Ryc. 10.11. Oscylogram, spektrogram i intonogram wypowiedzi z akcentem rdzennym xL *i do tego głupia*

stawowa osiąga wartość maksymalną we frazie (225 Hz), sylaba *wiać* znajduje się na poziomie średnim (M), 170 Hz.

Ryc. 10.11 ilustruje akcent typu xL we frazie *i do tego głupia*. Na ostatniej sylabie zauważa się bardzo niską częstotliwość. Sylaba rdzenna *głu* leży w dolnym zakresie zmian częstotliwości podstawowej. Na sylabie następującej po sylabie rdzennej — *pia* zmiany parametru F0 są nieregularne, częstotliwość spada poniżej



Ryc. 10.12. Oscylogram, spektrogram i intonogram wypowiedzi z akcentem rdzennym MM *wyraz twarzy, ręce, nogi*

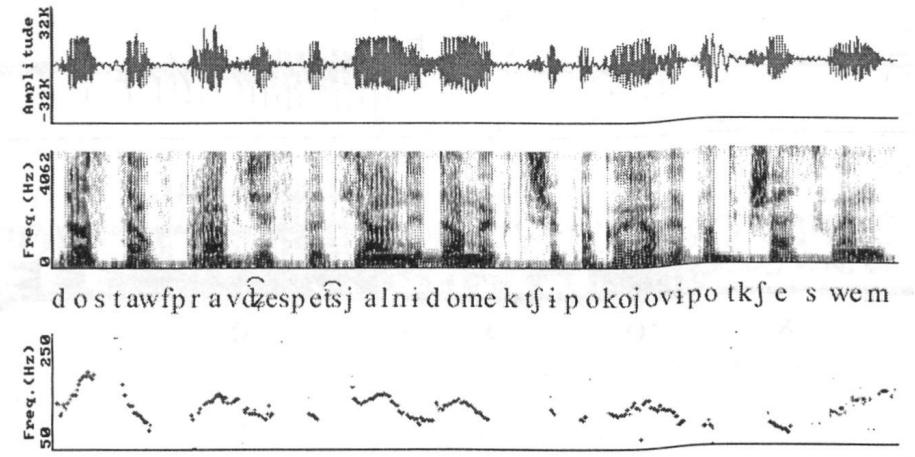


Ryc. 10.13. Oscylogram, spektrogram i intonogram wypowiedzi z akcentem rdzennym LH *film z gatunku fikcji naukowej*

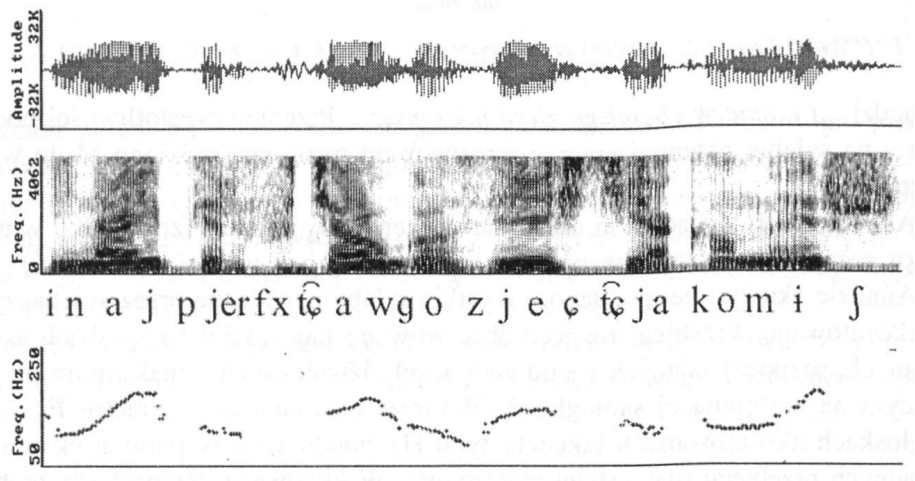
statystycznej dla tego głosu wartos'ci Fmin. Akcenty typu xL występują najczęściej na końcu zdania, przed pauzą akustyczną.

Przykładowo, na ryc. 10.12 przedstawiono intonację stałą MM na sylabach *twarzy* we frazie *...że ma wyraz twarzy..* Wahania parametru F0 w tej strukturze wynoszą 10 Hz. Na spółgłoskach obserwuje się głównie efekty mikroprozodii.

Na ryc. 10.13 przedstawiono przebieg parametru F0 w wypowiedzi z akcentem rdzennym LH *...film z gatunku fikcji naukowej...* Na sylabie rdzennej *ko* wartość



Ryc. 10.14. Oscylogram, spektrogram i intonogram wypowiedzi z akcentem rdzennym LM *dostał wprawdzie specjalny domek trzypokojowy pod krzesłem*



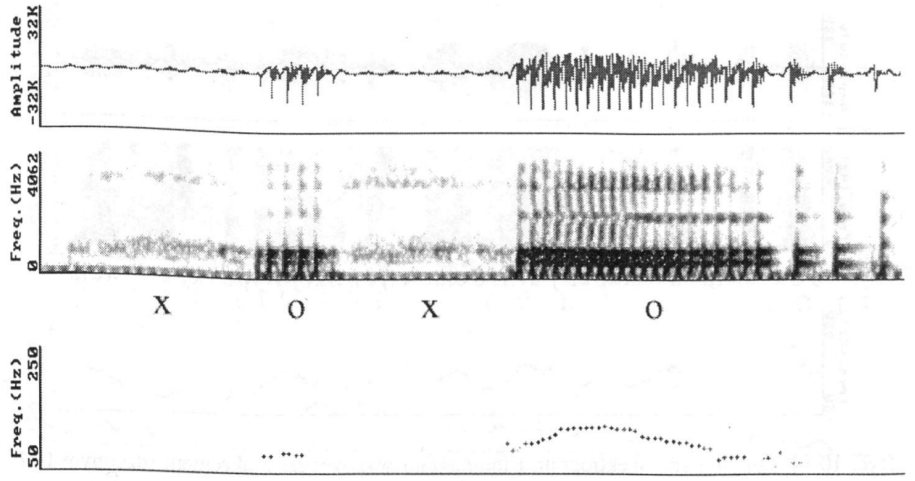
Ryc. 10.15. Oscylogram, spektrogram i intonogram wypowiedzi z akcentem rdzennym MH

i najpierw chciał go zjeść jako mysz

parametru F0 osiąga minimum (75 Hz). Na sylabie *wej* obserwuje się szybki wzrost częstotliwości do 190 Hz (do globalnego maksimum).

Na ryc. 10.14 zilustrowano intonację LM w wypowiedzi *…dostał wprawdzie specjalny domek trzypokojowy pod krzesłem…* Na sylabie rdzennej *krze* występuje minimum globalne w przebiegu częstotliwości podstawowej, na sylabie następnej *łem* obserwuje się typowy wzrost parametru F0 charakterystyczny dla wzorca LM (od poziomu niskiego L do średniego M).

Ryc. 10.15 ilustruje frazę z akcentem rdzennym MH na wyrazie *mysz* w wy-



Ryc. 10.16. Oscylogram, spektrogram i intonogram wypowiedzi z akcentem rdzennym LHL

ho, ho…

powiedzi *…i najpierw chciał go zjeść jako mysz…* Przebieg częstotliwości podstawowej na sylabie rdzennej *mysz* jest rosnący od poziomu średniego M do wysokiego H.

Akcenty LHL w badanym materiale występowały sporadycznie, tylko w przypadku specjalnej emfazy, jak np. *Ho, ho… ślicznie pan wyrósł* (ryc. 10.16).

Analizie akustycznej poddano wszystkie sylaby oznaczone przez słuchaczy jako akcentowane. Przebiegi rosnące obserwowano najczęściej na sylabach akcen-

towanych, rozpoczynających się od spółgłosek dźwięcznych z maksimum przypadającym na następującej samogłosce. Wartości ekstremalne parametru F0 na samogłoskach akcentowanych (akcenty typu H) znajdują się w pobliżu ekstremów globalnych przebiegu (najczęściej maksimów). W kilkunastu przypadkach zaobserwowano wartość Fmin stanowiącą minimum przebiegu (akcent typu L) — na samogłosce nie znajdującej się w pobliżu granicy frazy.

Ocena percepcyjnej i akustycznej segmentacji sygnału oraz sylab akcentowanych wymaga szczegółowej analizy statystycznej. Dla wszystkich występujących w tekście sylab oraz samogłosek wyznaczono oddzielne zbiory parametrów charakteryzujące zmienność częstotliwości podstawowej, iloczasu oraz energii. Dane te stanowią podstawę do automatycznej klasyfikacji/rozpoznawania typu akcentu w języku polskim (rozdział 13).

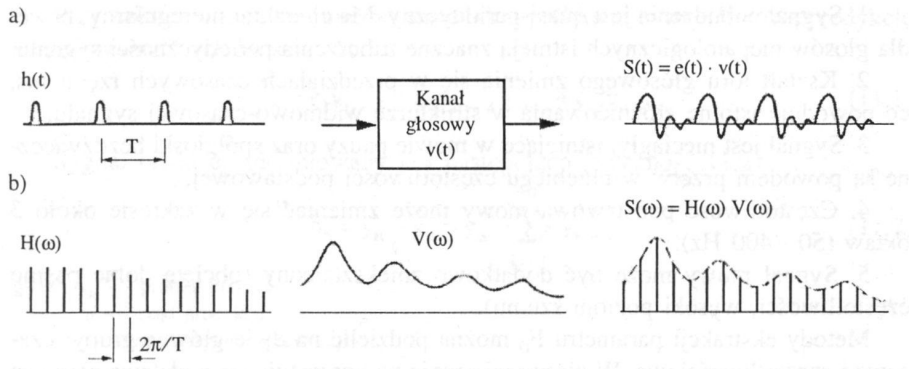
11 PODSTAWY MATEMATYCZNEGO OPISU SUPRASEGMENTALIÓW

11.1. POMIAR I PRZETWARZANIE CZĘSTOTLIWOŚCI PODSTAWOWEJ

11.1.1. EKSTRAKCJA SKŁADOWEJ PODSTAWOWEJ SYGNAŁU MOWY

Sygnał mowy można modelować jako odpowiedź kanału głosowego, przedstawionego w postaci odpowiednio pobudzanego układu liniowego o parametrach zmiennych w czasie. Jeśli kształt kanału głosowego zmienia się powoli, to odpowiedź tę można aproksymować w krótkich przedziałach czasu jako splot pobudzenia i odpowiedzi impulsowej kanału głosowego. Model taki (według Oppenheima 1982, s. 13) ilustruje ryc. 11.1.

Głoski bezdźwięczne powstają przy wymuszeniu przepływu powietrza poprzez przewężenie kanału głosowego. Turbulentny strumień powietrza wytwarza źródło



Ryc. 11.1. Model wytwarzania mowy w postaci odpowiedzi liniowego układu quasi-stacjonamego: a) opis w dziedzinie czasu, b) opis w dziedzinie częstotliwości

szumowe pobudzające kanał głosowy. W przypadku głosek dźwięcznych, jeśli pobudzenie w określonym przedziale czasu jest okresowe o stałej częstotliwości podstawowej, to odpowiedź układu jest także okresowa. Widmo pobudzenia okresowego jest widmem prążkowym o harmonicznych rozmieszczonych równomiernie co 2π/t i o obwiedni zależnej od kształtu impulsów krtaniowych.

Z punktu widzenia **wytwarzania** mowy okres podstawowy T0 może być zdefiniowany jako czas upływający między dwoma kolejnymi impulsami krtaniowymi. Punkt początkowy pomiaru przyjmuje się arbitralnie, najczęściej jako punkt zamknięcia głośni lub maksymalnego jej otwarcia. Sygnał pobudzenia ma w przybliżeniu asymetryczny, trójkątny kształt o bogatym widmie z tłumionymi (z szybkością 12 dB/oktawę) wyższymi harmonicznymi. Źródłem tonu krtaniowego są drgania fałdów głosowych, których napięcie i masa wpływają bezpośrednio na wytwarzanie drgań.

Z punktu widzenia **przetwarzania** sygnału T0 jest zdefiniowane w dziedzinie czasu jako średni czas kilku impulsów krtaniowych lub w dziedzinie częstotliwości jako częstotliwość podstawowa (w przybliżeniu) harmonicznej struktury w krótkookresowym widmie sygnału mowy.

Sposób uśredniania, w szczególności przedział czasowy, w którym się go przeprowadza, zależy od indywidualnej metody. Długoterminowa definicja okresu nie ma praktycznego zastosowania, ponieważ przedziały czasu między kolejnymi impulsami tonu krtaniowego nie są stałe, lecz zmieniają się w sposób ciągły. Częstotliwość próbkowania uzależniona jest najczęściej od dolnej mierzonej częstotliwości. Dla głosu o F0min = 100 Hz można przyjąć okres próbkowania 10 ms, dla F0min = 200 Hz okres próbkowania nie powinien być krótszy od 5 ms.

W ciągu ostatnich kilkudziesięciu lat powstały setki algorytmów pomiaru częstotliwości podstawowej mowy. Nie uzyskano jednak w pełni satysfakcjonującego rozwiązania problemu, o czym pisał Hess (1983) w swojej pracy przeglądowej poświęconej ekstraktorom parametru F0.

Złożoność problemu wynika ze specyficznych cech sygnału mowy:

1. Sygnał pobudzenia jest quasi-periodyczny. Ma charakter nieregularny, nawet dla głosów niepatologicznych istnieją znaczne zaburzenia periodyczności sygnału.

2. Kształt toru głosowego zmienia się w przedziałach czasowych rzędu ms, co powoduje istotne zróżnicowania w strukturze widmowo-czasowej sygnału.

3. Sygnał jest nieciągły, istniejące w mowie pauzy oraz spółgłoski bezdźwięczne są powodem przerw w przebiegu częstotliwości podstawowej.

4. Częstotliwość podstawowa mowy może zmieniać się w zakresie około 3 oktaw (50 - 400 Hz).

5. Sygnał mowy może być dodatkowo zniekształcony (obcięte dolne pasmo częstotliwości, wysoki poziom szumu).

Metody ekstrakcji parametru F0 można podzielić na dwie główne grupy: czasowe i częstotliwościowe. W pierwszej grupie wykorzystuje się podstawę czasową sygnału, a w drugiej krótkookresową transformację w dziedzinie częstotliwości. Główną zaletą ekstraktorów wykorzystujących czasową strukturę sygnału jest możliwość pomiaru długości każdego okresu, nawet w przypadku sygnałów nieregularnych (np. w mowie osób z patologiami narządu głosu). Krótkookresowa reprezentacja sygnału umożliwia ekstrakcję sekwencji średnich estymat okresu w krótkich przedziałach czasowych. Poniżej przedstawiono najczęstsze algorytmy ekstrakcji bazujące na metodach wykorzystujących cechy sygnału w dziedzinie czasu oraz krótkookresową reprezentację sygnału mowy (według Hessa 1983).

11.1.1.1. Metody wykorzystujące cechy sygnału w dziedzinie czasu

1. Ekstrakcja równoległego przetwarzania

Podstawą metody jest założenie, że najbardziej wiarygodny pomiar można osiągnąć przez właściwą kombinację wyników kilku ekstraktorów podstawowej składowej sygnału. Periodyczność szacuje się na podstawie analizy wybranych punktów ekstremalnych odpowiednio odfiltrowanego (do około 900 Elz) sygnału mowy.

2. Ekstraktor redukcji danych

Metoda wykorzystuje zasadę efektywnej redukcji informacji. Charakterystyczną cechę sygnału stanowi tzw. cykl przejścia, zdefiniowany jako suma wszystkich próbek między dwoma najbliższymi przejściami sygnału przez poziom zerowy. W znaczących cyklach przejścia (na początku każdego okresu) sygnał ma zwykle dużą amplitudę i długi czas trwania (a więc dużą energię).

3. Ekstraktory wykorzystujące technikę inwersyjnej filtracji

Założeniem metody jest rekonstrukcja funkcji pobudzenia — przebiegu tonu krtaniowego i na tej podstawie oszacowanie długości każdego okresu. Metoda ta

jest szczególnie przydatna w zastosowaniach medycznych, szczególnie w analizie wszelkiego rodzaju dysfonii.

4. Ekstrakcja bazująca na metodzie LPC (Linear Prediction Analysis)

Przy założeniu korelacji między sąsiednimi próbkami mowy, liniową predykcję można wyrazić zależnością (11.1)

$$\hat{y}_n = \varepsilon_1 y_{n-1} + \varepsilon_2 y_{n-2} + \ldots + \varepsilon_m y_{n-m} \tag{11.1}$$

Sygnał błędu δn zdefiniowany jest następująco (zależność 11.2):

$$\delta_n = y_n - \hat{y}_n = y_n - \sum_{i=1}^{m} \varepsilon_i y_{n-i} \tag{11.2}$$

gdzie: y_n — próbka bieżąca,
m — rząd predykcji,
ε — współczynniki predykcji.

Wyznaczanie parametrów pobudzenia z zastosowaniem predykcji liniowej polega na badaniu sygnału błędu otrzymanego w wyniku filtracji pierwotnego sygnału mowy w układzie o transmitancji będącej odwrotnością aproksymacji funkcji transmitancji kanału głosowego. Sygnał błędu jest aproksymacją sygnału pobudzenia.

11.1.1.2. Metody wykorzystujące krótkookresową reprezentację sygnału mowy

1. Metody autokorelacyjne.

W przypadku funkcji autokorelacji sygnał wejściowy skorelowany jest sam ze sobą. Jeżeli sygnał jest periodyczny lub prawie periodyczny to po upływie czasu T0 funkcja autokorelacji powinna przyjąć wysoką wartość. Analiza opóźnienia, po którym funkcja autokorelacji przyjmuje wartość maksymalną bezpośrednio wyznacza okresowość funkcji. Autokorelację często stosuje się w połączeniu z innymi metodami. Funkcja autokorelacji sygnału rdif zdefiniowana jest następująco (zależność 11.3):

$$r_{dif} = \frac{1}{N} \sum_{n=0}^{N-dif-1} y_n y_{(n+dif)} \tag{11.3}$$

gdzie: y_n — próbka bieżąca,
dif — opóźnienie,
N — interwał czasowy.

2. Metoda funkcji różnicowej AMDF (Average Magnitude Difference)

Funkcja różnicowa AMDFdif opisana jest zależnością (11.4):

$$AMDF_{dif} = \frac{1}{N} \sum_{n=q}^{q+N-1} | y_n - y_{(n+dif)} | \tag{11.4}$$

gdzie: y_n — próbka bieżąca,
dif — opóźnienie,
N — długość framy sygnału.

Jest to rodzaj funkcji autokorelacji. Przy różnych opóźnieniach *dif* bada się sygnał różnicowy yn - y(n+dif). Spodziewać się można, że funkcja AMDF osiągnie silne minimum przy opóźnieniu między sygnałami równym T0. Dla sygnału dokładnie periodycznego funkcja AMDFdif powinna być równa zeru.

3. Ekstraktory wykorzystujące homomorficzną analizę sygnału mowy.

Widmo segmentów dźwięcznych sygnału mowy jest iloczynem obwiedni widma reprezentującej kanał głosowy oraz mikrostruktury widma przedstawiającej pobudzenie. Logarytm widm jest sumą logarytmu obwiedni widma i logarytmu widma pobudzenia. Logarytm obwiedni widma zmienia się powoli w dziedzinie częstotliwości. Logarytm pobudzenia jest okresowy. Parametry pobudzenia wydzielane są z części cepstrum występującej dla dużych wartości czasu. W segmentach dźwięcznych w cepstrum pojawiają się maksima występujące dla wielokrotności okresu tonu krtaniowego. Ekstraktory bazujące na metodzie rozplotu homomorficznego należą do najbardziej wiarygodnych metod pomiaru F0.

11.1.1.3. Porównanie metod pomiarowych

Żadna z powyższych metod nie ma charakteru uniwersalnego. Poszczególne techniki mają swoje specyficzne wady wynikające z przyjętej metody przetwarzania sygnału (np. wrażliwość na strukturę formantową lub zmiany poziomu sygnału). Błędy ekstrakcji zalicza się najczęściej do jednej z trzech grup:

- a. błędy tzw. duże (często wynikające z niewłaściwego pomiaru drugiej lub wyższych harmonicznych), w wyniku których wyznaczona wartość parametru przekracza kilkanaście lub kilkadziesiąt procent wartości pomiarów sąsiednich,
- b. błędy drobne, które wynikają z niedokładności stosowanej metody,
- c. błędy w detekcji dźwięczności/bezdźwięczności sygnału.

Optymalnym, stosowanym obecnie rozwiązaniem jest wykorzystanie kilku równolegle pracujących ekstraktorów częstotliwości podstawowej i uwzględnienie statystycznie najbardziej wiarygodnej kombinacji wyników. Z uwagi na naturalne nieregularności sygnału nie jest możliwe otrzymanie na wyjściu układu technicznego regularnego ciągu pomiarów częstotliwości podstawowej.

Ogólne sterowanie zmiennością częstotliwości podstawowej jest przez mówcę kontrolowane (np. rozkład, rodzaj akcentów), natomiast mikrofluktuacje sygnału (wynikające z kontekstu fonematycznego, np. z kontekstu ze spółgłoskami zwartymi) uwarunkowane są zjawiskami aerodynamicznymi. Znaczna część zaburzeń periodyczności jest więc w sygnale mowy naturalna, możliwa do zaobserwowania np. na początku lub końcu fonacji albo też w sąsiedztwie spółgłosek zwartych.

W rzeczywistości system słuchowy człowieka wygładza te nieregularności i zmiany parametru F0 percypuje jako ciągłe melodyczne struktury. Mikroprozodia i pauzy spowodowane bezdźwięcznością spółgłosek nie mają wpływu na słuchowy odbiór akcentu, przyczyniają się natomiast do wrażenia naturalności sygnału. Fakt ten jest bardzo istotny w interpretacji zmian parametru częstotliwości podstawowej w wypowiedzi. Zmiana częstotliwości podstawowej rzędu 20 Hz występująca na spółgłosce jest dla modelowania oraz detekcji akcentu nieistotna. Tego samego rzędu zmiana parametru F0 na samogłosce może powodować percepcyjne wrażenie akcentu.

Na podstawie analizy 312 języków Whalen i Levitt (1995) określili mikroprozodię jako zjawisko uniwersalne. Zróżnicowanie wyników w estymacji wielkości zmian mikroprozodycznych (w zakresie 3% - 20%) dla poszczególnych języków, związane jest z niejednakową statystyczną istotnością badań, opartych na różnorodnym materiale (opracowanym na podstawie kilku do kilkudziesięciu głosów). Znaczne efekty koartykulacyjne określili dla języka francuskiego Di Cristo i Hirst (1986), a dla języka szwedzkiego Lyberg (1984).

Wartość parametru na początku samogłoski jest wyższa po spółgłoskach bezdźwięcznych zwartych niż po dźwięcznych zwartych (różnica ta może być nawet większa od 3 półtonów). Szeroko zaakceptowana jest opinia, że częstotliwość podstawowa opada po bezdźwięcznych spółgłoskach zwartych, a po ich dźwięcznych odpowiednikach wzrasta. Zjawisko to nie występuje jednak regularnie. Między innymi Silverman (1986) postawiła hipotezę, że kierunek zmian parametru F0 po spółgłoskach zwartych zależy nie tylko od cech segmentalnych, ale również od

prozodycznej struktury wypowiedzi. W większości języków różnice koartykulacyjne dochodzą do 1,5 półtonu.

W języku polskim zjawisko to badała Steffen-Batogowa (1970, 1973) oraz Matuszkina (1976). Względna wielkość spadków związanych z koartykulacją maleje w następującej kolejności: spółgłoski zwarte 8,5%, trące 6,4%, nosowe 4,6%, płynne r, 1 - 6,7% (dane dla tekstów czytanych, Matuszkina 1976).

Na ryc. 11.2a i 11.2b przedstawiono przykłady przebiegów parametru F0 w wypowiedzi: *mama myje rano lalę małej Joli* oraz w wypowiedzi *Grzegorz zjada rano w domu dużo dżemu* (z 6 różnymi pozycjami akcentu rdzennego). W przeprowadzonym teście percepcyjnym słuchacze ocenili przebiegi intonacyjne w obu przykładach jako takie same. Wizualna ocena oraz analiza akustyczna wykazują względnie gładki kontur intonacyjny w wypowiedziach *mama myje rano lalę małej Joli* (spółgłoski nosowe i płynne), natomiast w wypowiedzi *Grzegorz zjada rano w domu dużo dżemu* zauważalne są duże perturbacje parametru F0 występujące głównie na spółgłoskach zwartych i zwarto-trących.

Istotny problem stanowi więc interpretacja występujących i uzasadnionych artykulacyjnie nieregularności zmian — wygładzanie sygnału.

Najprostszą próbą rozwiązania tego zagadnienia jest przyjęcie jednej z możliwości wygładzania danych za pomocą bardziej lub mniej złożonego wielomianu i jako ocenę jakości wygładzania zastosowanie kryterium błędu np. średniokwadratowego. Metoda ta może funkcjonować poprawnie w połączeniu z odpowiednią interpretacją zmian częstotliwości podstawowej na spółgłoskach oraz samogłoskach.

Kontur intonacyjny zawiera fragmenty o zróżnicowanej ważności dla percepcji. Ważna jest więc interpretacja nieregulamości zmian z punktu widzenia syntezy i rozpoznawania struktur melodycznych. Wygładzanie przebiegu częstotliwości podstawowej musi być połączone z segmentacją sygnału mowy i analizą błędów pomiaru na szczególnie istotnych fragmentach, odpowiadających samogłoskom. Do wygładzania przebiegu często stosuje się 3-punktowe liniowe okno Hanna lub medianę (zależności 11.5 oraz 11.6)

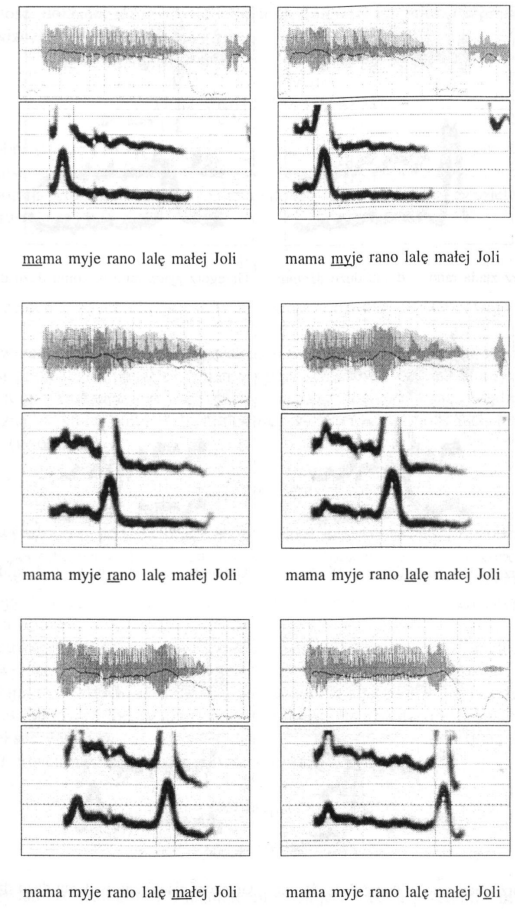
$$y_{(q)}' = 0{,}25y_{(q-1)} + 0{,}5y_{(q)} + 0{,}25y_{(q+1)} \qquad (11.5)$$

gdzie: y_q' — dane wygładzone,
y_q — dane pomiarowe,
q — próbka bieżąca

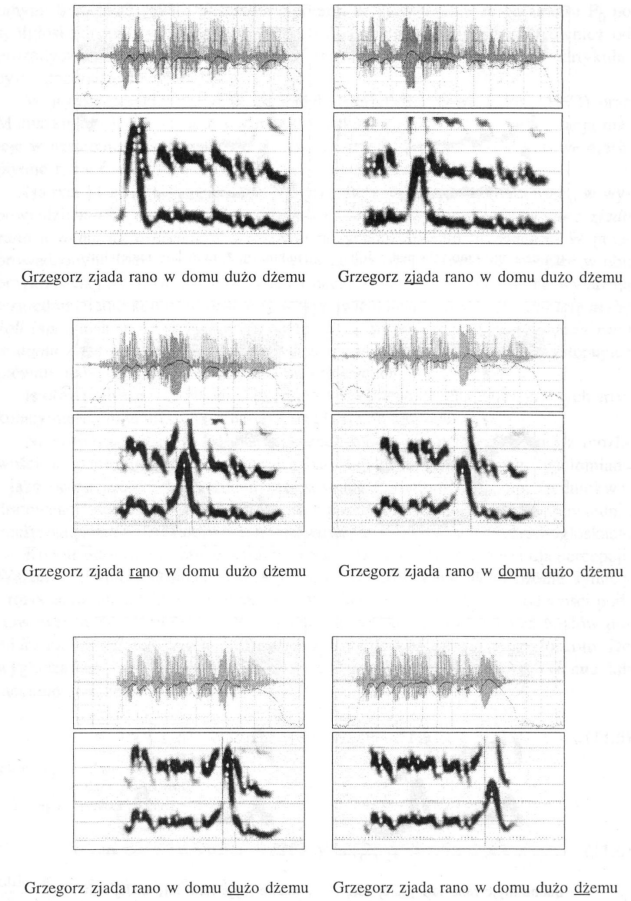
$$\mathrm{med}(y,K) = y\,[(K-1)/2], \quad y_i \le y_{(i+1)} \quad i = 0,1,\ldots,k-2 \qquad (11.6)$$

gdzie: K — mediana z próbek y_0 do $y_{(K-1)}$.

W zależności od ukierunkowania pomiaru częstotliwości podstawowej, ze względu na określone zastosowania w syntezie, rozpoznawaniu czy patologii mowy oraz nauce intonacji, jako priorytetowe wymagania dla ekstraktora przyjmuje się wia-



Ryc. 11.2a. Przebieg częstotliwości podstawowej w wypowiedzi: *mama myje rano lalę małej Joli* z 6 pozycjami akcentu rdzennego



Ryc. 11.2b. Przebieg częstotliwości podstawowej w wypowiedzi: *Grzegorz zjada rano w domu dużo dżemu* z 6 pozycjami akcentu rdzennego

rygodność, dokładność, szybkość pomiaru, odporność na zniekształcenia oraz jakość wizualizacji konturu intonacyjnego.

11.1.2. SKALE POMIAROWE

Dobór skali, na której można byłoby zgodnie z psychoakustycznym wrażeniem wysokości tonu odzwierciedlać zmiany częstotliwości podstawowej, jest najczęściej pomijanym problemem w analizie intonacji. Niezależnie od przyjętej metody pomiaru relacja 11.7

$$F_0 = \frac{1}{T_0} \tag{11.7}$$

uwzględnia wyrażenie okresu w ms, a częstotliwości podstawowej w Hz.

Powszechnie stosowana (zwłaszcza we wcześniejszych pracach) skala liniowa nie jest przydatna do analizy zmian względnych częstotliwości, istotnych w percepcji tonu. Alternatywnie więc stosuje się skalę logarytmiczną, między innymi muzyczną, w której zmiany częstotliwości wyrażone są w półtonach lub ćwierćtonach (wzór 11.8).

$$F_{\text{półtony}} = 12\log_2 \frac{f_2}{f_1} \quad (11.8)$$

F wyrażone w półtonach określa odległość między dwoma częstotliwościami (f| — początkową zmianą częstotliwości wyrażoną w Hz i f2 — końcową wyrażoną w Hz). Skale — melowa oraz barkowa — stosowane w psychoakustyce, nie mają do analizy intonacji zastosowania, ponieważ poniżej 1 KHz są w przybliżeniu liniowe.

Na uwagę zasługuje wprowadzana w ostatnich latach do analizy melodii mowy skala wyrażona w erbach (zależności 11.9 i 11.10). W skali tej pasma krytyczne (określające selektywność systemu słuchowego poniżej 500 Hz) są pośrednie między skalą liniową i logarytmiczną. Pasma te ustalono na drodze psychoakustycznych eksperymentów (np. Hermes et al. 1991). Różnice między skalą wyrażoną w Hz i erbach przedstawiono we wzorach 11.9 oraz 11.10.

$$\text{Erb} = 16{,}7 \log_{10}(1 + \frac{f}{165{,}4}) \quad (11.9)$$

$$f = 165{,}4\,(10^{0.06\text{Erb}} - 1) \quad (11.10)$$

Przydatność skali erbów w analizie intonacji jest obecnie przedmiotem dalszych badań (Hermes i Rump 1994, Hermes 1995). Problem wyboru skali ma różne konsekwencje w określonych zastosowaniach. Dla syntezy mowy nie jest obojętny wybór skali, w jakiej ma być jednakowo percepcyjnie wyróżniona sylaba akcentowana. Dotyczy to szczególnie uwzględnionego zakresu wysokości głosu — od niskiego męskiego do wysokiego kobiecego. Np. w męskim głosie zmiana od 120 Hz do 180 Hz oznaczać będzie porównywalną zmianę częstotliwości w głosie kobiecym od 240 Hz do 300 Hz (przy założeniu liniowej skali) lub zmianę od 240 Hz - 360 Hz (przy uwzględnieniu skali logarytmicznej) oraz zmianę od 240 Hz do 325 Hz (przy zastosowaniu skali wyrażonej w erbach). Efektywność skali zależy od celu analizy. Transformacja logarytmiczna a priori nie zawsze jest konieczna, ponieważ technika normalizacyjna zastosowana do logarytmicznych czy też liniowych danych może funkcjonować jednakowo dobrze. Praktycznym rozwiązaniem wydaje się w analizie intonacji stosowanie skali logarytmicznej pozwalającej na standardowy opis zmian tonu dla różnych zastosowań. W obecnej pracy w analizie zmian wysokości tonu stosowano skalę logarytmiczną.

11.1.3. NORMALIZACJA

Istniejące różnice w subiektywnej i obiektywnej ocenie zmian wysokości tonu są źródłem trudności w wyborze kryteriów klasyfikujących jednostki intonacyjne.

Zapisy obiektywne częstotliwości podstawowej mogą wykazywać zróżnicowania w zakresie:

- a. ciągłości/nieciągłości (przerwy w ciągłości przebiegu uwarunkowane są występowaniem spółgłosek bezdźwięcznych),
  b. długości frazy (uwarunkowanej głównie liczbą sylab),
  c. różnego rozkładu ekstremów (określonego lokalizacją akcentów).

Do wymienionych powyżej uwarunkowań językowych należy dodać zróżnicowania pozajęzykowe związane głównie z wysokością głosu i tempem wypowiedzi. Jedną z metod osiągnięcia inwariantności jest normalizacja danych. Aby osiągnąć

np. inwariantność ze względu na przesunięcia oraz modyfikacje skal, należy dane tak przeskalować, aby miały wartość średnią równą zeru i jednostkową dyspersję. Normalizacja jest ważnym aspektem przygotowania danych do analizy, ponieważ już proste przeskalowanie współrzędnych może prowadzić do odmiennego podziału na grupy. Wyniki dotychczasowych badań nie pozwalają w sposób jednoznaczny na wyjaśnienie relacji między percepcyjną i fizyczną normalizacją, wykazują jednak istotność zmian względnych w percepcji tonu. Najczęstsze metody normalizacji uwzględniają zmiany parametru F0 w odniesieniu do wybranych arbitralnie wartości. Przykładowo normalizacja według poniższej zależności (11.11) uwzględnia zmiany częstotliwości podstawowej względem zakresu i wybranego punktu odniesienia (por. np. Rose 1991).

$$F_{0i}' = \frac{(F_{0i} - F_{0ref})}{F_{0range}} \qquad (11.11)$$

gdzie: F0i'— częstotliwość znormalizowana,
F0i — częstotliwość normalizowana,
F0ref — punkt odniesienia,
F0range — zakres.

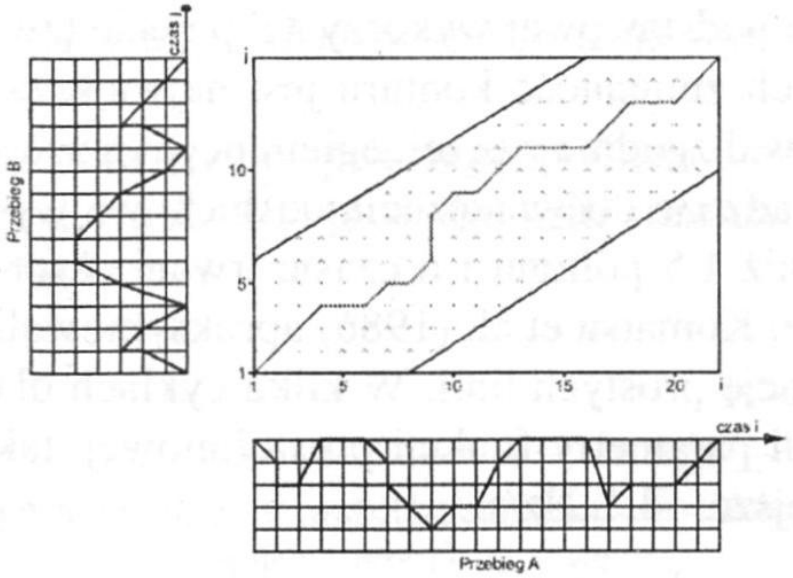
Często stosowana normalizacja wykorzystuje parametry średnie rozkładów częstotliwości podstawowej, wartość średnią F0 oraz odchylenie standardowe δ(11.12.)

$$F_{0i}' = \frac{(F_{0i} - \bar{F}_{0)}}{\delta} \qquad (11.12)$$

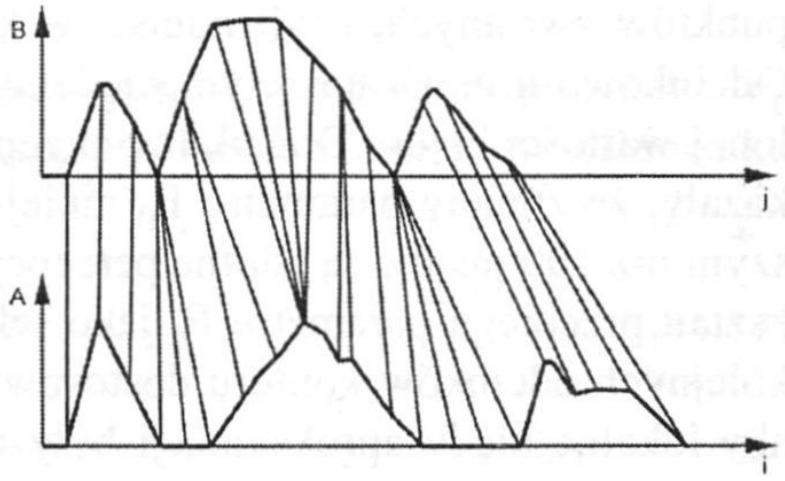
Jedną z przyczyn trudności w modelowaniu i analizie struktur intonacyjnych jest nierównomierność tempa wypowiedzi. Z licznych prób rozwiązania tego zagadnienia, dla problemu normalizacji wypowiedzi w rozpoznawaniu mowy na uwagę zasługuje metoda nieliniowego dopasowania (time warping) oparta na technice programowania dynamicznego.

Na ryc. 11.3a i b przedstawiono koncepcje metody nieliniowej normalizacji czasowej DTW (Dynamic Time Warping) w postaci klasycznej (według Sickerta 1983). Nad osiami i oraz j umieszczono dwa przebiegi sygnału A i B reprezentowane przez kolejne próbki mowy A = a(l),…a(i)….a(I) oraz B = b(l),…b(j)….b(J).

Liczba danych I oraz J określa rozmiary macierzy odległości między poszczególnymi punktami przebiegu A i B. Optymalne dopasowanie obu ciągów nastąpi wówczas, gdy znaleziona zostanie taka ścieżka łącząca lewy dolny i prawy górny róg macierzy, na której suma odległości między elementami a(i) i b(i) będzie minimalna. Celem metody jest znalezienie takiego odwzorowania, które optymalnie wyeliminuje różnice czasowe w przebiegach A i B. Od punktu początkowego do końcowego wylicza się dla wszystkich kolejnych punktów skumulowane odległości między tymi punktami a punktem końcowym według rekursywnej reguły (wzór 11.13)



Ryc. 11.3a. Ilustracja nieliniowego czasowego dopasowania przebiegu A oraz B



Ryc. 11,3b. Rezultat nieliniowego czasowego dopasowania przebiegu A oraz B

$$L(i, j) = l(i, j) + \min\{L(i-1, j), L(i-1, j-1), L(i, j-1)\} \qquad (11.13)$$

gdzie: l(i, j) — różnica między segmentem i-tym jednej próbki oraz j-tym segmentem drugiej,
L(i, j) — minimalna suma odległości między punktem (i, j) a końcowym.

Omawiana procedura ma wiele odmian. Dla przeprowadzenia normalizacji częstotliwości podstawowej metodą dopasowania nieliniowego konieczne są modyfikacje, głównie w zakresie określenia granic, w których przeprowadzane jest dopasowywanie. Próby normalizacji czasowej konturów intonacyjnych języka polskiego metodą DTW (por. np. Jassem, Demenko 1986, 1989) wykazały konieczność precyzyjnej, odrębnej normalizacji w obrębie samogłosek i spółgłosek. Problem

dopasowania czasowego przebiegów intonacyjnych jest szczególnie istotny dla wizualizacji przebiegu (np. w nauce intonacji). W systemach analizy i syntezy suprasegmentaliów zagadnienie normalizacji czasowej jest złożone, należy uwzględnić informację bezpośrednią, wynikającą z wpływu określonych źródeł zmienności na iloczas elementów fonetycznych.

11.2. PARAMETRYZACJA KONTURU INTONACYJNEGO

11.2.1. APROKSYMACJE PRZEBIEGÓW CZĘSTOTLIWOŚCI PODSTAWOWEJ

Często spotykaną metodą parametryzacji intonacji jest aproksymacja zmian tonu arbitralnie wybraną funkcją. W literaturze znaleźć można co najmniej kilkanaście sposobów aproksymacji przebiegów częstotliwości podstawowej opartych na bardziej lub mniej globalnym matematycznym dopasowaniu określonej funkcji do danych empirycznych w obrębie frazy bądź jej fragmentu.

1. Aproksymacje funkcjami liniowymi.

Scheffers (1981) w opisie częstotliwości podstawowej wykorzystał pojęcie tzw. punktów zwrotnych, czyli miejsc, w których zmienność konturu jest największa. Odcinkowo-liniową aproksymację kontrolował zgodnie z przebiegiem pewnej ustalonej wartości błędu. Dodatkowo przeprowadzone doświadczenia odsłuchowe wykazały, że zmiany parametru F0 mniejsze niż 1,5 półtonu i o czasie trwania krótszym niż 100 ms nie są istotne percepcyjnie. Komatsu et al. (1986) aproksymowali kształt przebiegu parametru F0 jako sekwencję prostych linii. W kilku cyklach dla kolejnych odcinków konturu dostosowywali parametry funkcji prostoliniowej, tak aby lokalne błędy aproksymacji były mniejsze od 1 Hz/s.

2. Aproksymacje wielomianami.

Levitt i Rabiner (1971) opisywali przebiegi częstotliwości podstawowej w krótkich (80 ms) oknach czasowych (przesuwanych co 40 ms) wielomianami ortogonalnymi (liniowymi oraz nieliniowymi 2. oraz 3. stopnia). Autorzy zastosowali średniokwadratowe kryterium błędu dopasowania funkcji (błąd nie może przekraczać od 5,5 Hz). Olive (1975) opisywał przebiegi częstotliwości podstawowej w prostych zdaniach wielomianami 4. stopnia. Z uwagi na trudności z interpretacją współczynników, w dalszym etapie swojej pracy aproksymował przebiegi parametru F0 oddzielnie dla każdego wyrazu, za pomocą krzywej opisanej 4 wartościami: dla punktu początkowego, środkowego, końcowego oraz parametrem określającym stromość krzywej w końcowych jej fragmentach. Hirst et al. (1991) aproksymowali przebiegi częstotliwości podstawowej języka francuskiego funkcjami sklejanymi (spline function) drugiego rzędu. Testy percepcyjne wykazały użyteczność tego rodzaju opisu dla języka francuskiego. ’t Hart (1991) testował percepcyjnie paraboliczną *(„sharp”)* oraz prostoliniową *(„fiat”)* aproksymację konturu intonacyjnego. Odsłuchy wykazały, że paraboliczny opis daje podobne wyniki jak liniowy (jeśli zastąpi się maksima lub minima globalne płaskim fragmentem przebiegu rzędu 30-40 ms). d’Alessandro i Mertens (1995) przeprowadzili podobną aproksymację jak ’t Hart odcinkami prostymi, ale uwzględnili dodatkowo percepcyjną stylizację (przyjęli priorytet aproksymacji końcowego fragmentu konturu). Katae et al. (1995) zastosowali na skali logarytmicznej aproksymację przebiegów częstotliwości podstawowej trapezami opisanymi siedmioma danymi.

3. Aproksymacja funkcjami trygonometrycznymi.

Reinecke i Lehning (1994) aproksymowali przebiegi częstotliwości podstawowej szeregami Fouriera. Poprawny opis krótkiego (do 2 sekund) zdania osiągnięto przy wykorzystaniu 25 współczynników szeregu.

4. Aproksymacja zbiorem funkcji.

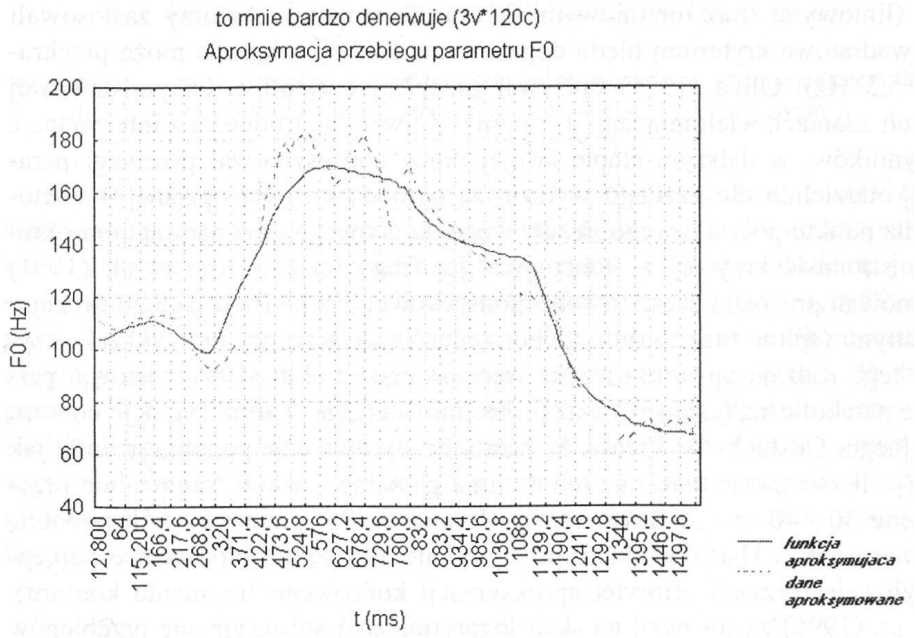
Moore et al. (1994) badali przydatność w modelowaniu przebiegu intonacyjnego 16 funkcji (1 liniowej i 15 nieliniowych: wykładniczych, trygonometrycznych, sigmoidalnych). Zbiór funkcji testowano na intonacjach w krótkich wypowiedziach. Najwyższy współczynnik korelacji między danymi doświadczalnymi i danymi wyznaczonymi funkcją wykładniczą wynosił 0,83.

Dla języka polskiego podjęto próbę aproksymacji przebiegów parametru F0 (zależność 11.14) w krótkich frazach funkcją wykładniczo-potęgową (Demenko 1983).

$$F_{0j} = at_j^b e^{ct_j} \qquad (11.14)$$

gdzie: a, b, c — współczynniki funkcji aproksymującej,
t_j — kolejne próbki czasowe.

Osiągnięto dobre wyniki aproksymacji krótkich fragmentów przebiegów częstotliwości podstawowej (z jednym akcentem). Opracowanie praktycznych reguł



Ryc. 11.4. Aproksymacja przebiegu parametru F0 w wypowiedzi *to mnie bardzo denerwuje*

aproksymacji przebiegów w wypowiedziach dłuższych (z kilkoma akcentami) stanowiło istotny problem. Próby aproksymacji przebiegów częstotliwości podstawowej w wypowiedziach języka polskiego według modelu Fujisaki zaprezentowanego w rozdziale 5 nie przyniosły zadowalających wyników. Na ryc. 11.4 zilustrowano aproksymację przebiegu częstotliwości podstawowej w krótkiej wypowiedzi *To mnie bardzo denerwuje.* Zastosowano 1 funkcję frazową (a = 0,031 oraz *Kp* = 0,41) oraz 3 funkcje aproksymujące składowe akcentowe odpowiednio o parametrach: β = 0,16, Ka = 0,150 (*tp* = 10 ms, *tk* = 140 ms), β = 0,16, *Ka*= 0,77 (*tp* = 280 ms, *tk* = 740 ms), β = 0,16, *Ka* = 0,65 (*tp* = 750 ms, *tk* = 1080 ms). Oznaczenia funkcji przyjęto według modelu Fujisaki (rozdział 5).

Wprawdzie funkcja aproksymująca znacznie wygładziła przebieg, jednak duże rozbieżności między danymi eksperymentalnymi a wartościami funkcji aproksymującej są nie do zaakceptowania zarówno w analizie, jak i syntezie intonacji.

11.2.2. OPIS STRUKTURALNY

Inną możliwością parametryzacji konturu intonacyjnego jest próbkowanie go w określonych momentach czasowych tl…tn. W ten sposób każdy przebieg określony jest przez wektor w przestrzeni n-wymiarowej, dogodnej do analizy numerycznej. Jeżeli jednak punktów pomiarowych jest dużo (a w przypadku częstotliwości podstawowej mogą być setki pomiarów) to problem parametryzacji komplikuje się. Stosuje się więc transformacje przestrzeni umożliwiające skoncentrowanie informacji na kilku pierwszych składowych i eliminację nieistotnych współrzędnych.

Przeprowadzono próbę wykorzystania metody Karhunena-Loevego (K-L) w opisie zmienności częstotliwości podstawowej kilkudziesięciu wypowiedzi języka polskiego (Demenko 1984). Wyniki pracy wykazały możliwość opisu zmian tonu w przestrzeni kilkuwymiarowej. Istotną trudność stanowiło ustalenie korelacji wektorów bazowych z cechami fizycznymi transformowanych przebiegów.

Do efektywnego modelowania zmian wysokości tonu ma służyć etykietyzacja „tilt" (Taylor 1993, 1995). Tilt reprezentuje zmienność częstotliwości podstawowej (sumę wzrostu oraz spadku) w obrębie przebiegu parametru F0 zawierającego lokalne ekstremum (zależność 11.15).

$$\text{tilt} = \frac{|\,\text{Arise}\,| - |\,\text{Afall}\,|}{|\,\text{Arise}\,| + |\,\text{Afall}|} \qquad (11.15)$$

gdzie: Arise — amplituda wzrostu,
Afall — amplituda spadku częstotliwości podstawowej na samogłosce.

Tilt = -1 oznacza wyłącznie spadek, 1 — wzrost, 0 — równy wzrost i równy spadek częstotliwości podstawowej. Etykietyzację „tilt" autorzy stosowali zarówno do syntezy, jak i analizy oraz rozpoznawania intonacji. Wyniki prac wykazały użyteczność tego rodzaju opisu w automatycznym przetwarzaniu zmian częstotliwości podstawowej, zarówno w tekstach czytanych, jak i w dialogach.

W najnowszych opracowaniach poświęconych automatycznej analizie oraz klasyfikacji suprasegmentaliów bierze się pod uwagę tzw. cechy strukturalne częstotliwości podstawowej (powstałe na bazie pomiarów pierwotnych), iloczasu oraz intensywności. Najczęściej wykorzystywane parametry opisujące zmienność parametru F0, to np. wartość średnia, minimalna oraz maksymalna, stosunek wartości końcowej przebiegu częstotliwości do wartości średniej, współczynniki regresji, umiejscowienie maksimum/minium przebiegu F0 względem początku/końca samogłoski, parametry opisujące relacje między zmianami tonu na sąsiednich samogłoskach. W zakresie opisu intensywności (energii) sygnału najczęściej uwzględnia się: maksymalną lub średnią energię w obrębie sylaby, współczynniki regresji, błąd kwadratowy między linią regresji i konturem intensywności sylaby lub ciągu sylab. W opisie iloczasu na potrzeby automatycznej analizy wykorzystuje się: średni iloczas fonemu, znormalizowany iloczas samogłoski (sylaby), tempo mowy oraz czas trwania przerwy akustycznej.

W automatycznych analizach suprasegmentaliów, w około 80% wykorzystuje się opisy zmienności częstotliwości podstawowej, a w kilku bądź w kilkunastu procentach zmiany iloczasu oraz energii sygnału (por. np. McAllister 1991, Kuijk i Boves 1997). Jak dotychczas, podjęto niewiele prób parametryzacji strukturalnej. Szczegóły dotyczące tego zagadnienia przedstawia np. praca Streefkerk (1997).

Parametryzacja strukturalna wydaje się ekonomiczna i stosunkowo mało skomplikowana w implementacjach syntezy oraz rozpoznawania mowy.

11.3. STATYSTYCZNE METODY ANALIZY SUPRASEGMENTALIÓW

11.3.1. OGÓLNA CHARAKTERYSTYKA METOD KLASYFIKACJI

Wykorzystanie wiedzy doświadczalnej przy przejściu od jakościowego opisu zjawiska (w przypadku suprasegmentaliów mogą to być obserwacje dotyczące zmienności np. częstotliwości podstawowej) do badań ilościowych wymaga dokładnej oceny wyników pomiarów. Należy stwierdzić, czy są one w zgodzie z przewidywaniami teoretycznymi oraz czy umożliwiają przyjęcie bądź odrzucenie założonej hipotezy. Wyniki pomiarów nie zawsze są jednoznacznie określone poprzez procedurę doświadczalną i podlegają fluktuacjom. Częściowo losowy charakter danych opisujących suprasegmentalia zawarty jest w naturze eksperymentów (badaniom podlegają różni parlatorzy oraz słuchacze) oraz w niedokładnościach ekstrakcji parametru F0, iloczasu czy energii sygnału. W praktyce wyznacza się kształt rozkładu badając próbę — zbiór złożony ze skończonej liczby pomiarów. Oszacowanie statystycznych parametrów rozkładu wymaga obszernego materiału doświadczalnego.

W przypadku suprasegmentaliów najczęściej wykorzystywane są metody matematyczne (w tym statystyczne). Metody matematyczne można podzielić na deterministyczne, które przyjmują pewien aparat matematyczny, ale nie wymagają żadnych założeń co do statystycznych własności klas, jak to ma miejsce w przypadku uczących iteracyjnych algorytmów. Klasycznym przykładem jest metoda perceptronowa (reward-punishment algorithm) opracowana przez Rosenblatta w r. 1957 (za Tadeusiewiczem 1993). Metoda została przetestowana praktycznie na sieci neuropodobnej *Perceptron,* stanowiącej pierwszą realnie funkcjonującą imitację sieci neuronowej (por. np. Tadeusiewicz 1993).

W podejściu statystycznym do analizy lub klasyfikacji danych wykorzystuje się statystyczne własności rozkładów cech opisujących badaną populację. Najbardziej podstawowymi narzędziami są najczęściej: analiza wariancji, analiza dyskryminacyjna oraz metoda K-L (Karhunen-Loève). Szczegółowe opisy większości tych metod podano np. w pracach Sobczak i Malina (1985), Lachenbruch (1975), Morrison (1990), Brandt (1998).

Popularną, statystyczną techniką stosowaną w analizie/klasyfikacji/rozpoznawaniu, szczególnie przydatną w przypadku zmian czasowych (np. w mowie), są modele Markowa (HMMs — Hidden Markov Models). Modele wykorzystują łańcuchy Markowa o ograniczonej liczbie słów i ograniczonych zbiorach rozkładów prawdopodobieństw wyjściowych. Wadą tej metody są duże obciążenia obliczeniowe oraz konieczność wyznaczania statystycznych parametrów rozkładów.

W praktycznych zastosowaniach często spotyka się również inne grupy metod klasyfikacji/rozpoznawania, np. heurystyczne i syntaktyczne. Heurystyczne podejście oparte jest na doświadczeniu i wiedzy eksperymentatora. Klasyfikacja składa się ze sformułowanych ad hoc procedur przeznaczonych do określonego zadania. Metody te często występują łącznie np. z metodami matematycznymi.

Metody lingwistyczne (syntaktyczne) stosowane są w odniesieniu do sygnału mowy i równie często w projektach klasyfikacji/rozpoznawania obrazów. Klasa może być opisana przez hierarchiczną strukturę podklas (analogicznie jak w języku). Szczególnie ważne jest sformułowanie gramatyki pozwalającej na sterowanie regułami w analizie syntaktycznej.

Dla zbiorów danych, opisanych nie tylko ilościowo, ale również jakościowo najczęściej wykorzystywaną techniką są tzw. zbiory rozmyte (np. Węglarz, Czogała i Łęcki 1997).

Również w rozpoznawaniu mowy podejmowano próby stosowania tej metodologii (np. Gubrynowicz et al. 1981).

Poniżej zostaną omówione pokrótce tylko dwie z wymienionych procedur statystycznych: analiza dyskryminacyjna i HMMs. Pierwsza z nich stosowana była także w polskich badaniach, a druga jest coraz bardziej popularna w wielu pracach dotyczących analizy sygnału mowy.

11.3.2. ANALIZA DYSKRYMINACYJNA

Zagadnienie dyskryminacji obiektów w wielowymiarowej przestrzeni analizował Fisher w 1936 r. (według np. Lachenbruch 1975, Gatnar 1998). Zaproponował on funkcje liniowe opisujące hiperpłaszczyzny rozdzielające zbiory obiektów w ten sposób, aby otrzymać jak najlepsze odseparowanie poszczególnych klas obiektów.

W celu znalezienia współczynników *ai* określających położenie hiperpłaszczyzny, w równaniu (11.16):

$$g(\mathbf{x}) = \sum_{i=1}^{m} a_i x_i + a_0 \qquad (11.16)$$

gdzie: a_i — współczynniki funkcji dyskryminacyjnej,

Fisher obliczył odległości między środkami skupień dla standaryzowanych cech obiektów. Określił kierunek r, dla którego wyrażenie 11.17 przyjmuje największą wartość

$$\frac{\mathbf{r}^T \bar{\mathbf{x}}_1}{\sqrt{\mathbf{r}^T \mathbf{S}\mathbf{r}}} - \frac{\mathbf{r}^T \bar{\mathbf{x}}_2}{\sqrt{\mathbf{r}^T \mathbf{S}\mathbf{r}}} \qquad (11.17)$$

gdzie: $\bar{\mathbf{x}}_1$ oraz $\bar{\mathbf{x}}_2$ — wektory średnich arytmetycznych w obu klasach,
$\mathbf{S}$ — wspólna dla nich macierz wariancji.

Po przekształceniach równania (11.17) można otrzymać równanie 11.18

$$\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 = \frac{\mathbf{r}^T (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)\mathbf{S}\mathbf{r}}{\sqrt{\mathbf{r}^T \mathbf{S}\mathbf{r}}} \qquad (11.18)$$

gdzie: $\mathbf{r}$ — szukany kierunek.

Postać funkcji dyskryminacyjnej jest w zasadzie dowolna, jednak w praktyce najczęściej stosuje się funkcje liniowe lub kwadratowe (np. Lachenbruch 1975).

Zagadnienie klasyfikacji w ramach analizy dyskryminacyjnej dotyczy głównie rozkładów normalnych lub takich, które mogą być opisane wystarczająco dokładnie za pomocą momentów 1. i 2. stopnia. W odniesieniu do badań suprasegmentaliów metodę tę wykorzystuje się dość często (np. Hunt 1994, Sagisaka et al. 1997). Szczególnie istotne jest w tym przypadku staranne testowanie własności statystycznych rozkładów prawdopodobieństw cech opisujących poszczególne klasy, ponieważ wykazano (np. Jassem et al. 1968 oraz Steffen-Batóg et al. 1970), że rozkłady wartości częstotliwości podstawowej w mowie zbliżone są do lognormalnych.

Dla własnego materiału językowego wykorzystano metodę analizy dyskryminacyjnej do klasyfikacji tonów nuklearnych w dwusylabowej wypowiedzi *dobrze*

(Demenko 1986, Demenko et al. 1988). Wypowiedzi sparametryzowano 8-elementowym wektorem opisującym zmienność parametru F0 w 8 wybranych punktach czasowych. Zastosowano liniowe i kwadratowe funkcje dyskryminacyjne, które wykazały zgodnie (w 80%) rozłączność badanych klas. Najlepiej rozpoznano najwyraźniejsze akustycznie i percepcyjnie akcenty (LHL, LH, HL).

Wyniki obecnie przeprowadzonych badań na złożonych melodycznie wypowiedziach oraz mowie ciągłej, przedstawione w rozdziale 13, a także wyniki klasyfikacji akcentów przeprowadzonej w roku 1988 na materiale językowym (zawierającym tylko izolowaną wypowiedź), wskazują, że w języku polskim istnieje co najmniej 9 tonów rdzennych (por. rozdział 6). Różnice akustyczne między nimi są na tyle wyraźne, że pozwalają na ich automatyczną klasyfikację/rozpoznawanie.

11.3.3. MODELE MARKOWA

Ukryte modele Markowa wykorzystuje się do modelowania procesów stochastycznych. Każdy z modeli jest zbiorem połączonych ze sobą stanów S = {S1…sn}, gdzie n jest liczbą stanów. W kolejnych momentach czasowych t = (l, 2,. .T) modelowany proces przechodzi z jednego stanu do drugiego, generując sygnały Stany są ukryte przed obserwatorem. Dla każdej pary stanów Si oraz Sj określone jest prawdopodobieństwo przejścia z jednego stanu do drugiego. Zachowanie modelu zależy w chwili t + 1 wyłącznie od stanu, w którym znajdował się proces w poprzedzającej chwili t. Istnieją dwa procesy stochastyczne: jeden, dotyczy nieznanej sekwencji stanów (symboli), którą trzeba „odkryć” (rozpoznać), i drugi opisuje przyporządkowywanie poszczególnym stanom (symbolom) niejednoznacznych sygnałów, np. akustycznych.

Technice modelowania sygnału mowy z zastosowaniem łańcuchów Markowa poświęcono od lat 80. bardzo obszerne prace (np. Rabiner 1989). Technika ta jest bardzo złożona obliczeniowo. Współczesne implementacje algorytmów rozpoznawania mowy bazujące na HMM (np. Naturally Speaking — firmy Dragon, Via Voice — firmy IBM, FreeSpeech 98 — firmy Philips), są efektem prac wielu zespołów badawczych na całym świecie i dużych nakładów finansowych.

Modelowanie łańcuchami Markowa zmienności suprasegmentalnej mowy, jak dotąd nie przyniosło zadowalających rezultatów. O ile wyniki rozpoznania akcentu w dwusylabowych wypowiedziach są względnie zachęcające, to w zakresie mowy ciągłej nie udało się uzyskać tak dobrych rezultatów, jak w przypadku cech segmentalnych mowy. Freij i Falliside (1988) uzyskali przy zastosowaniu 5-stanowego HMMs dokładność klasyfikacji akcentu w dwusylabowych wyrazach równą 94%. Butzberger (1990) w 3-stanowym HMMs uzyskał poprawność rozróżniania wzorców intonacyjnych (na wyrazach izolowanych) przypisanych 5 kategoriom: pytaniu, stwierdzeniu, wątpliwości, komendzie, kontynuacji, wynoszącą 89%. Taylor et al. (1997) analizowali struktury intonacyjne dialogu w mowie spontanicznej i uzyskali na podstawie modelowania w 8-stanowym HMMs średni procent rozpoznawania 67%. Wstępne eksperymenty modelowania zmienności intonacji łańcuchami Markowa w systemie text-to-speech dla języka japońskiego przeprowadził Fukada et al. (1994). Osiągnięto średni błąd kwadratowy między danymi eksperymentalnymi a danymi pochodzącymi z modelowania równy 9,2% (dla przeciętnej wartości 120 Hz).

Jednym z powodów trudności implementacji HMMs do analizy cech melodycznych mowy mogą być specyficzne własności suprasegmentaliów, związane np. z koniecznością przetwarzania informacji w obrębie różnych iloczasowo fragmentów wypowiedzi. Należy uwzględnić klasyfikację intonacji na preiktyczną oraz rdzenną i nie wystarczy samo stwierdzenie obecności bądź braku akcentu w obrębie sylaby.

Ostatnio metodą częściej wykorzystywaną w analizie suprasegmentaliów są sieci neuronowe, w których problem reprezentacji czasowej struktury sygnału

można rozwiązać pośrednio, np. poprzez podanie informacji o iloczasach sylab docelowych (jak i sylab sąsiednich) wchodzących w skład danego wzorca intonacyjnego.

12 SIECI NEURONOWE W ANALIZIE SUPRASEGMENTALIÓW

12.1. SFORMUŁOWANIE PROBLEMU

Klasycznym, często spotykanym obszarem technicznych zastosowań sieci neuronowych jest klasyfikacja/rozpoznawanie sygnałów wizualnych oraz dźwiękowych, w tym także mowy. Problematyce tej poświęcono szereg praktycznych i teoretycznych opracowań (por. np. Morgan i Scofield 1992, Reichl et al. 1995, Ramachandran i Mammone 1995). Zagadnienie wykorzystania sieci neuronowych zarówno w projektach badawczych, jak i konkretnych aplikacjach przedstawiono również w pracach polskich autorów, np. Tadeusiewicza (1993), Mikruta (1993), Tadeusiewicza (1994), Tadeusiewicza i Flasińskiego (1991), Tadeusiewicza i Mikruta (1994), Rutkowskiej et al. (1997), Tadeusiewicza et al. (1998) oraz Izworskiego i Wszołka (1999).

W analizie suprasegmentaliów mowy podejmowano dotychczas tylko nieliczne, pilotażowe próby wykorzystania sieci neuronowych do klasyfikacji akcentu (np. Lee et al. 1995, Taylor 1995, Ying et. al. 1995, Chen et al. 1995, Kiessling et al. 1996), rozpoznawania granic frazowych (np. Batliner et al. 1997, Hess et al. 1997) bądź syntezy intonacji (np. Traber 1997). Pagel et al. (1995) poddali rozpoznawaniu 5 kategorii suprasegmentalnych określonych na podstawie 9-minutowego tekstu radiowego: 1) intonacja rosnąca tzw. „continuation rise" (50 elementów), 2) maksimum (128 elementów), 3) początkowy wzrost (129 elementów), 4) deklinacja (105 elementów), 5) elementy pozostałe (1287 elementów). Wektory cech opisano poprzez wartości średnie parametru F0, współczynniki regresji oraz iloczas sylaby. Uzyskano około 70% poprawności klasyfikacji przy wykorzystaniu jednowarstwowej sieci MLP (posiadającej 7 neuronów wejściowych, 10 ukrytych i 5 wyjściowych).

Do klasyfikacji fraz przyjmuje się przeważnie następujące typy granic: B3 pełna intonacyjna granica z silnie zaznaczoną zmianą parametru F0, (oraz dodatkowym wydłużeniem), B2 granica pośrednia ze słabym zaznaczeniem zmian w cechach suprasegmentalnych, B0 granica wyrazowa nie dokładnie wyznaczona, B9 niegramatyczna granica wskazująca błędy językowe — wahania lub powtórki (według Batliner 1997). Dla każdej sylaby oraz dla trzech poprzedzających i 3 następujących po niej sylab wyznaczono: iloczas (bezwzględny i znormalizowany), wartości parametru F0 — minimalne, maksymalne, początkowe, końcowe, średnie oraz energię w obrębie sylaby. Wykorzystanie sieci neuronowej MLP (40/20 neuronów) pozwoliło (na podstawie analizy około 1000 granic w 21 dialogach) na dokładność rozpoznawania granic frazowych w 80%. Mast et al. (1996) wyróżnili 18 suprasegmentalnych jednostek opisu DAU (dialog acts unit) oraz DAC (dialog acts category). Jako jednostkę DAU oraz DAC przyjęto podstawowe jednostki dialogu (np. pytanie, akceptacja, sugestia). Dla każdej końcowej sylaby wyrazu wyznaczono wektor cech określający kontur intonacyjny, pauzy, energię oraz iloczas. Najlepsze wyniki klasyfikacji uzyskano w sieci typu MŁP (zawierającej 60/30 neuronów) wykorzystującej 117 prozodycznych cech dla każdej sylaby w zdaniu.

W projekcie rozpoznawania mowy Verbmobil wykorzystanie prozodii przewidziano głównie dla ułatwienia segmentacji mowy i określenia znaczenia wypowiedzi: np. Hess et al. (1997), Batliner (1997) oraz Kiessling, Kompe et al. (1996), Niemann et al. (1998). Dla poszczególnych sylab wyznaczono 242-elementowy wektor

cech (cechy określano dla każdej sylaby, 6 ją poprzedzających i 6 następujących po niej). Zastosowanie sieci MŁP z 2 ukrytymi warstwami i 6 neuronami na wyjściu pozwoliło na uzyskanie poprawności klasyfikacji: dla B3 — 91%, B2 — 89 %, B0 — 90%; dla 3 klas sylab akcentowanych (A3 — akcent główny, A2 — drugorzędny, A1 — brak akcentu) średnio 95%. Udane próby parametryzacji i rozpoznawania 9 tonów chińskich w zakresie sylab izolowanych przeprowadził Lee et al. (1995). Informację wejściową opisano 5-elementowym wektorem: 3 cechami ilustrującymi zmienność tonu oraz 2 cechami określającymi iloczas oraz energię. W wyniku uczenia trzywarstwowej klasycznej sieci (z 20 neuronami ukrytymi i 9 wyjściowymi) algorytmem propagacji wstecznej uzyskano procent klasyfikacji tonów równy 89%. Dla języka chińskiego podjęto próbę klasyfikacji tonów leksykalnych w mowie ciągłej (Ying et al. 1995). Przy wykorzystaniu sieci neuronowej, jednokierunkowej z 10 neuronami wejściowymi, 20 ukrytymi i 6 wyjściowymi osiągnięto poprawną klasyfikację równą 76%.

Podstawową trudnością, na jaką napotyka projektant sieci przeznaczonej do przetwarzania informacji suprasegmentalnej jest brak zweryfikowanego percepcyjnie i akustycznie modelu zmian częstotliwości podstawowej mowy. Na obecnym etapie badań nie wiadomo, jakie informacje należy podawać na wejścia sieci i w jaki sposób projektować sygnały wyjściowe. Wybór cech opisujących jednostki intonacyjne (akcenty rdzenne, poboczne bądź też wzorce intonacyjne ustalone np. według kryterium percepcyjnego) ciągle jest problemem otwartym. W literaturze przedmiotu nie ma zgodności, jak adekwatnie opisać kontur intonacyjny. Jednym z możliwych rozwiązań tego problemu wydaje się oparcie modelu intonacji na naturalnym systemie percepcji słuchowej i opisie struktur melodycznych poprzez te cechy, które są istotne dla percepcyjnej klasyfikacji: kierunek przebiegu tonu, położenie na skali częstotliwości oraz lokalizacja zmiany tonu względem samogłoski.

Klasyfikacja struktur prozodycznych języka polskiego opisana została w niniejszym opracowaniu kilku- lub kilkunasto-elementowym wektorem cech (zależnie od liczby i rodzaju klasyfikowanych jednostek). Cechy te wybrano na podstawie eksperymentów odsłuchowych. Przyjęto tezę, że cechy akustyczne wyodrębnionych percepcyjnie wzorców pozwalają na automatyczną analizę struktury frazy intonacyjnej: klasyfikację/rozpoznawanie akcentu rdzennego oraz preiktycznego według określonego wcześniej sekwencyjnego modelu frazy intonacyjnej.

Dla automatycznej klasyfikacji słuchowo wyodrębnionych struktur suprasegmentalnych konieczne jest przyjęcie szeregu założeń uwzględniających percepcyjno-akustyczne cechy sygnału.

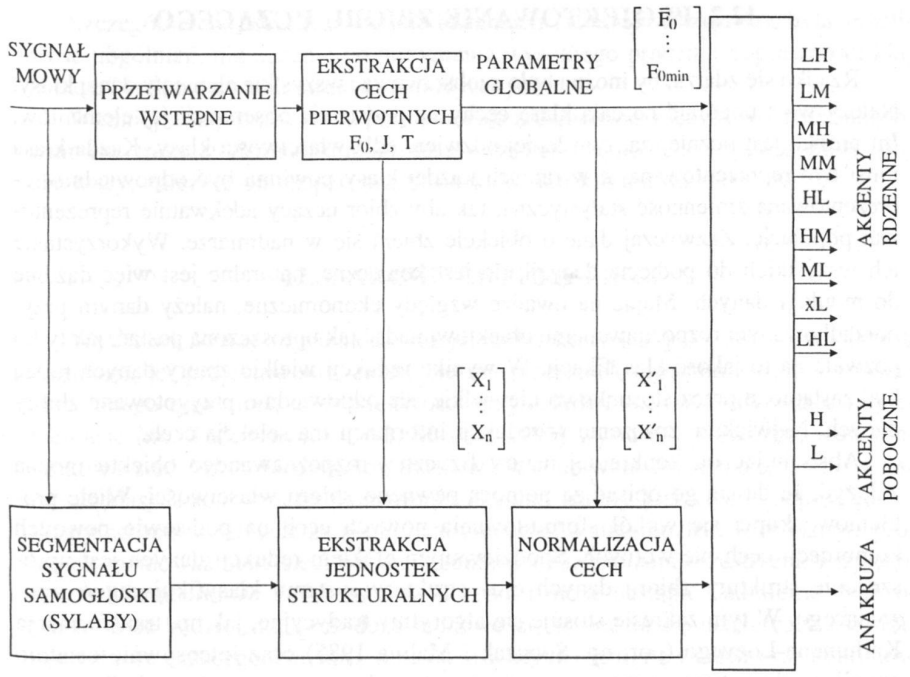
1. Jako podstawę klasyfikacji/rozpoznawania jednostek melodycznych należy uznać całościowe wzorce intonacyjne pojawiające się w sygnale mowy. Wyodrębnianie zmian wysokości tonu rozpatrywanych z sylaby na sylabę nie umożliwia jednoznacznej klasyfikacji (ta sama zmiana wysokości tonu z sylaby na sylabę może realizować różne wzorce intonacyjne). Ponadto rozkład sylab nieakcentowanych i akcentowanych w mowie jest nierównomierny. Sylab nieakcentowanych jest w mowie ciągłej kilkakrotnie więcej, co oznacza zaliczenie kilkakrotnie większej liczby przypadków bądź do jednej, bądź do drugiej klasy i z punktu widzenia statystycznego stanowi poważne utrudnienie w estymacji parametrów rozkładów prawdopodobieństw.

2. Cechy akustyczne wzorców intonacyjnych określone są głównie poprzez względne zmiany częstotliwości podstawowej zachodzące na poszczególnych samogłoskach/sylabach (przebieg tonu rosnący, równy i opadający) oraz zmiany wysokości tonu między sylabami (sylaba następująca po bieżącej może występować na równym poziomie, powyżej lub poniżej sylaby bieżącej). Zmiany te określają łączny kontur intonacyjny wzorca. Dla opisu wzorca konieczne jest więc nie tylko scharakteryzowanie przebiegu częstotliwości podstawowej, ale również określenie umiejscowienia samogłosek w poszczególnych fragmentach przebiegu. Podobne przebiegi częstotliwości podstawowej mogą oznaczać zupełnie różne wzorce (lub ich fragmenty),

zależnie od umiejscowienia w nim samogłosek (np. przebieg rosnąco-opadający może oznaczać typ wzorca LHL lub fragment wzorca rosnącego LH lub opadającego HL).

3. W mowie ciągłej, istotną rolę dla segmentacji sygnału mowy na frazy odgrywają cechy akustyczne ułatwiające rozdzielenie fraz, takie jak iloczas (wydłużenie sylab końcowych frazy) oraz intensywność (zmniejszenie energii końcowego fragmentu sygnału).

Na ryc. 12.1 przedstawiono schemat blokowy klasyfikacji jednostek intonacyjnych wybranych struktur, takich jak: akcenty rdzenne (9 klas) i opcjonalne akcenty poboczne (2 klasy) oraz anakruza. Z sygnału mowy ekstrahowana jest co 10 ms chwilowa wartość częstotliwości podstawowej oraz chwilowa wartość obwiedni energii sygnału. Automatyczny pomiar parametrów umożliwia oprogramowanie spektrografu cyfrowego Kay 5500. Segmentacja sygnału przeprowadzana jest w zakresie



Ryc. 12.1. Schemat klasyfikacji akcentów w języku polskim

określenia granic samogłosek oraz sylab. Złożoność zagadnienia akustycznej segmentacji mowy, problem automatyzacji segmentacji przedstawiono obszernie w pracy Kvale (1993). Zgodnie z aktualnymi tendencjami przyjęto, że sylaba może być zdefiniowana akustycznie poprzez maksima poziomu sygnału w obrębie samogłosek.

W przypadku analizy suprasegmentaliów precyzyjne określenie granic samogłoski lub sylaby nie jest wymagane. Przeciętna dokładność pomiaru rzędu 10- - 15 ms zapewnia wystarczająco poprawną segmentację. Parametry pierwotne: częstotliwość podstawowa, poziom intensywności, iloczas samogłoskowy, ekstrahowane z mowy, podlegają weryfikacji. Szczególnie istotna jest analiza błędów ekstrakcji częstotliwości podstawowej i w przypadku niepewnego pomiaru lepszym rozwiązaniem jest pominięcie wartości niż przyjęcie błędu (np. pomiaru drugiej harmonicznej). Normalizacja dotyczy głównie eliminacji różnic średnich wysokości głosów mówców oraz tempa mowy. Parametry pierwotne bez segmentacji sygnału nie mogą stanowić bezpośrednio podstawy klasyfikacji.

Wektor cech strukturalnych zawierający informacje o strukturze suprasegmentalnej w dziedzinie samogłoski/sylaby bądź ciągu sylab podawany jest na wejścia klasycznej jednokierunkowej sieci neuronowej MLP.

12.2. PROJEKTOWANIE ZBIORU UCZĄCEGO

Rzadko się zdarza, by można było zaobserwować wszystkie elementy danej klasy. Należy wiec uogólnić na całą klasę cechy na podstawie obserwacji jej elementów. Im próbka jest liczniejsza, tym lepiej odzwierciedla właściwości klasy. Każda klasa musi być reprezentowana, a w ramach każdej klasy powinna być odpowiednio reprezentowana zmienność statystyczna, tak aby zbiór uczący adekwatnie reprezentował populację. Zazwyczaj dane o obiekcie zbiera się w nadmiarze. Wykorzystanie ich wszystkich do podjęcia decyzji nie jest konieczne, naturalne jest więc dążenie do redukcji danych. Mając na uwadze względy ekonomiczne, należy danym przyporządkowanym rozpoznawanemu obiektowi nadać tak uproszczoną postać, jak tylko pozwala na to jakość klasyfikacji. W wyniku redukcji wielkie zbiory danych mogą być zastąpione przez stosunkowo niewielkie, ale odpowiednio przygotowane zbiory danych. Największe znaczenie w redukcji informacji ma selekcja cech.

Abstrahując od konkretnej natury fizycznej rozpoznawanego obiektu można założyć, że da się go opisać za pomocą pewnego zbioru właściwości. Wiele problemów skupia się wokół sformułowania nowych cech na podstawie pewnych kombinacji cech pierwotnych. Spodziewanym efektem redukcji danych jest uproszczenie struktury zbioru danych oraz struktury systemu klasyfikującego/rozpo- znającego. W tym zakresie stosuje się algorytmy tradycyjne, jak np. transformacja Karhunena-Loevego (por. np. Sobczak i Malina 1985) oraz intensywnie ostatnio rozwijane algorytmy genetyczne, często zresztą wykorzystywane jako technika wspomagająca w sieciach neuronowych (np. Goldberg 1998, Rutkowska et. al 1997).

Reprezentatywność danych oraz określenie wektorów cech opisujących obiekty w odniesieniu do suprasegmentaliów mowy są istotnym problemem wymagającym wyraźnego roztrzygnięcia. Z uwagi na specyfikę analiz struktur melodycznych danego języka badanie intonacji jest w obecnym stanie wiedzy badaniem częściowym. Istotne jest przygotowanie zbioru danych uwzględniających przynajmniej podstawowe źródła zmienności intonacji. Opracowanie zbioru uczącego np. tylko dla jednego głosu może zapewnić całkowitą lub prawie 100-procentową klasyfikację, lecz być może nie umożliwi przetestowania sieci w równie wysokim stopniu dla innej grupy osób. Sieć taka nie będzie miała własności uogólniania. Konieczne jest przygotowanie zbioru uczącego na zróżnicowanym materiale językowym, jakim jest mowa ciągła oraz uwzględnienie przynajmniej kilku głosów. Ekstrakcja cech suprasegmentalnych mowy, nawet przy dzisiejszym stanie wiedzy technicznej jest zadaniem bardzo czasochłonnym i wymagającym od eksperymentatora dużego doświadczenia w zakresie analizy instrumentalnej. Pomiary częstotliwości podstawowej oraz segmentacja sygnału uzyskana w wyniku automatycznych procedur (np. programami spektrografu cyfrowego Kay) są poddawane weryfikacji, najczęściej manualnej, która zapewnia względnie poprawną korektę. Ponieważ przeprowadzenie szczegółowych analiz suprasegmentaliów na obszernym materiale językowym (rzędu kilkunastu lub kilkudziesięciu godzin) wymagałoby oprócz dużej ilości czasu znacznych nakładów finansowych, problem starannej selekcji materiału badawczego jest bardzo istotny.

Szczególnie w dotychczasowych badaniach intonacji ważniejsza była możliwość uogólnień, niż dążenie do otrzymania wysokiego procentu poprawności klasyfikacji/rozpoznawania struktur melodycznych.

W niniejszym opracowaniu przyjęto więc zróżnicowany językowo materiał (powtarzane frazy izolowane, dialogi, teksty czytane, wypowiedzi syntezowane) pochodzący, zależnie od eksperymentu od kilku lub kilkunastu mówców.

12.3. ARCHITEKTURA SIECI

Wybór właściwego typu sieci oraz jej architektury jest jednym z ważniejszych zagadnień jej projektowania. Jeżeli nie ma przesłanek uzasadniających przyjęcie określonego typu sieci, to w większości przypadków, jak wykazały liczne implementacje (np. Tadeusiewicz 1993, Masters 1993) wystarczy klasyczna, jednokierunkowa sieć MLP. Sieć wielowarstwowa jest uniwersalnym narzędziem w zakresie klasyfikacji oraz rozpoznawania. Bardzo istotny jest wybór odpowiedniej liczby neuronów. Użycie za małej liczby neuronów uniemożliwi rozwiązanie problemu, ponieważ małe sieci nie mają zdolności rozdzielania przestrzeni. Przyjęcie zbyt wielu neuronów zwiększy czas uczenia i może nastąpić nadmierne dopasowanie. Sieć będzie miała tak duże możliwości, że będzie się uczyć cech nieistotnych zbioru uczącego. Nie majak na razie ścisłych reguł, ile warstw oraz neuronów należy wykorzystać i w jaki sposób zaprojektować warstwy. Wstępnego oszacowania struktury sieci można dokonać na podstawie cech geometrycznych analizowanego zbioru danych. Ryc. 12.2 ilustruje możliwości rozdzielania przestrzeni za pomocą nieliniowej sieci neuronowej (przykład Rumelharta zaczerpnięty z pracy Tadeusiewicza 1993, s. 54). Sieć dwuwarstwowa pozwala na rozpoznawanie wypukłych oraz jednospójnych obszarów, tj. simpleksów. Za pomocą trzeciej warstwy możliwe jest tworzenie dowolnych obszarów, niewypukłych i niejednospójnych.

Podstawową cechą sieci wielowarstwowych jest ich zdolność do realizacji dowolnie złożonych odwzorowań wejściowo-wyjściowych lub powierzchni decyzyjnych rozdzielających klasy obrazów.

Liczbę neuronów w warstwach ustala się eksperymentalnie. Dla większości problemów praktycznych nie ma teoretycznych przesłanek stosowania więcej niż dwóch warstw ukrytych. Wykorzystanie większej liczby warstw wydłuża proces uczenia i nie zawsze w sposób znaczący poprawia wyniki. Na wyjściu sieci liczba neuronów zależy od sposobu reprezentacji klasy. Jednym ze sposobów sygnalizowania kategorii obrazu przez klasyfikator jest reprezentacja lokalna. Układ sygnalizuje klasę i0 poprzez wartość 1 na wyjściu i-tego neuronu i wartości -1 na wyjściach pozostałych neuronów (dla neuronów unipolarnych wartości 0), czyli

$$y_i = 1 \quad \text{dla} \quad i = i_0$$

$$y_i = -1 \quad \text{dla} \quad i = 1,2,...R$$

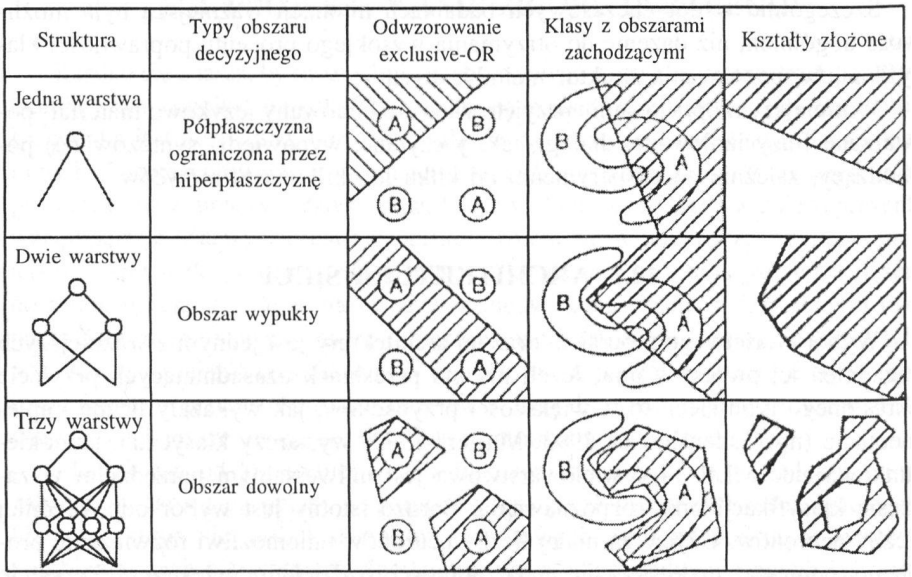
gdzie: R — liczba klas.

Ryc. 12.2. Własności dyskryminacyjne sieci neuronowej jednokierunkowej

12.4. PROCES UCZENIA

Podstawą konstrukcji większości algorytmów automatycznego uczenia jest reguła delta. Reguła ta zakłada, że wraz z każdym wektorem wejściowym podawany jest sygnał opisany jako zadana (wymagana) odpowiedź neuronu na sygnał wej-

ściowy x. Wektory: wejściowy i wyjściowy (zwane również obrazem wejściowym i wyjściowym) mają postać (12.1):



$$\mathbf{x} = [\, x_1 x_2 \ldots x_n \,]^t, \quad \mathbf{z} = [\, z_1 z_2 \ldots z_m \,]^t \tag{12.1}$$

Wagi *wij* łączą neuron *i* z wejściem *j*. Neuron ma zdolność adaptacji. Jego wagi podlegają modyfikacji w trakcie nauki. Według ogólnej zasady nauki przyjętej dla sieci neuronowych wektor wag Wj = [wi1,wi2,...win]t rośnie proporcjonalnie do iloczynu sygnałów wejściowego i uczącego.

Klasyczną regułą obowiązującą dla neuronów z ciągłymi funkcjami aktywacji i nadzorowanym trybem uczenia jest reguła delta. Uczący sygnał delta zdefiniowany jest następująco:

$$r = [\, z_i - f(\mathbf{w}_i^t \mathbf{x}) \,]\, f'(\mathbf{w}_i^t \mathbf{x}) \tag{12.2}$$

gdzie: z_i — pożądany sygnał wyjściowy,
$f'(\mathbf{w}_i^t \mathbf{x})$ — pochodna funkcji aktywacji f(net)

Reguła delta może być otrzymana z warunku najmniejszego kwadratu błędu Q między *oi* oraz *zi* (zależność 12.3)

$$Q = \frac{1}{2}(z_i - o_i)^2 \tag{12.3}$$

gdzie: o_i — rzeczywisty wyjściowy sygnał neuronu.

Zależność 12.3. jest ekwiwalentna zależności (12.4)

$$Q = \frac{1}{2}[(z_i - f(\mathbf{w}_i^t \mathbf{x})]^2 \tag{12.4}$$

Gradient błędu określony jest więc zależnością (12.5)

$$\nabla Q = -(z_i - o_i)\, f'(\mathbf{w}_i^t \mathbf{x})\, \mathbf{x} \tag{12.5}$$

Składowe gradientu (dla j = l,2,...,n) określa wzór (12.6)

$$\frac{\delta Q}{\delta w_{ij}} = -(z_i - o_i)\, f'(\mathbf{w}_i^t \mathbf{x})\, x_j \tag{12.6}$$

Ponieważ minimalizacja błędu wymaga zmian wag w kierunku ujemnym (spadku wartości funkcji) przyjmujemy zależność (12.7)

$$\Delta \mathbf{w_i} = \eta \nabla Q \tag{12.7}$$

gdzie: η — dodatnia stała.

Z równań 12.5 i 12.7 otrzymujemy zależność 12.8

$$\Delta \mathbf{w_i} = \eta (z_i - o_i)\, f'(\mathrm{net}_i)\, \mathbf{x} \tag{12.8}$$

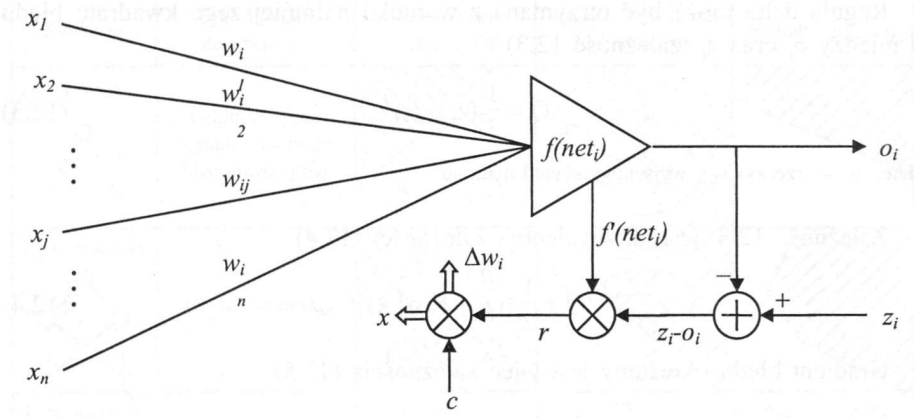
Zmianę pojedynczej wagi określa wzór 12.9

$$\Delta w_{ij} = \eta (z_i - o_i)\, f'(\mathrm{net}_i)\, x_j \tag{12.9}$$

Na ryc. 12.3 zilustrowano regułę delta (por. Żurada 1992, s. 66).

Metoda propagacji wstecznej wykorzystująca regułę delta jest uniwersalnym algorytmem uczenia sieci wielowarstwowych, którego szczegóły przedstawiono między innymi w kilku pracach (Tadeusiewicz 1993, 1998, Żurada 1992, Żurada et al. 1996). Metodę propagacji wstecznej błędu można uważać za metodę optymalizacji minimalizującą zadane kryterium jakości za pomocą metody gradientowej. Kryterium jakości jest funkcją błędu wyjściowego zdefiniowanego w przestrzeni wag. Algorytm uczenia modyfikuje wagi w kierunku zmniejszania się błędu.

Mając wyznaczony błąd podczas realizacji *j*-tego kroku procesu uczenia w neu-



Ryc. 12.3. Uczenie według reguły delta (za Żuradą 1992, s. 66)

ronie o numerze *m* można rzutować ten błąd wstecz do wszystkich tych neuronów, których sygnały stanowiły wejścia dla m-tego neuronu. W literaturze istnieje bardzo dużo przykładów poprawnego działania sieci po zastosowaniu metody propagacji wstecznej. W załączniku 8 podano schemat tego ważnego algorytmu.

Przy praktycznym uczeniu sieci brak jest ogólnych optymalnych metod postępowania i wciąż jest to proces, w którym eksperyment i intuicja eksperymentatora odgrywają zasadniczą rolę. W złożonych sieciach istotną rolę odgrywa sprawność uczenia. Metodom przyśpieszania procesu uczenia sieci poświęcono wiele uwagi. Jednym ze sposobów optymalizacji uczenia jest odpowiedni dobór jego współczynników. Efektywność i zbieżność metody propagacji wstecznej błędu zależy w istotny sposób od wartości współczynnika korekcji. Współczynnik ten powinien być dobierany eksperymentalnie dla każdego problemu w zakresie wartości od 10-3 do 10. Zmniejszanie się błędu może być powolne, gdy jednocześnie gradient i współczynnik uczenia są niewielkie. Jednym ze sposobów przyspieszenia zbieżności jest metoda momentu, która dokonuje korekty wag według reguły 12.10 (Żurada et al. 1996 s. 144):

$$\Delta \mathbf{w}(k) = -\eta \nabla Q(k) + \gamma \Delta \mathbf{w}(k-1) \tag{12.10}$$

gdzie: k — aktualny krok uczenia,
k – 1 — poprzedni krok uczenia,
γ — dodatnia stała,
η — składnik momentu określający drugi składnik sumy.

Podczas uczenia sieci podstawową miarą jej jakości jest zwykle błąd średniokwadratowy na wyjściu. Obliczany w metodzie propagacji wstecznej błąd łączny jest sumą błędów po wszystkich *p* obrazach uczących.

$$Q = \frac{1}{2} \sum_{l=1}^{p} \sum_{k=1}^{K} (z_{lk} - o_{lk})^2 \tag{12.11}$$

Znormalizowany błąd średni kwadratowy, uwzględniający liczbę obrazów uczących określony jest zależnością 12.12.

$$Q_{rms} = \frac{1}{pK} \sum_{l=1}^{p} \sum_{k=1}^{K} (z_{lk} - o_{lk})^2 \tag{12.12}$$

gdzie: p — liczba obrazów uczących,
o — sygnał wyjściowy sieci,
z — pożądany sygnał wyjściowy,
K — liczba neuronów.

Jako praktyczny wskaźnik ilustrujący perspektywy uczenia sieci proponuje się wskaźnik DELTA (Mikrut 1993), uwzględniający błąd obliczony dla elementu rozpoznającego zadany obraz oraz maksymalny błąd dla pozostałych elementów. Jak wykazały eksperymenty (por. np. Izworski, Wszołek 1999) wskaźnik DELTA pozwala na efektywną ocenę pracy sieci.

W ocenie jakości funkcjonowania sieci należy uwzględnić właściwe jej przetestowanie. Po procesie uczenia sieć powinna zostać zweryfikowana na podstawie danych nie wchodzących w skład zbioru uczącego. Jeżeli sieć dobrze interpoluje nowe dla niej obrazy, można założyć, że zarówno sieć, jak i zbiór uczący zostały właściwie zaprojektowane. Technika projektowania sieci neuronowych wyłania się jako oddzielna dyscyplina naukowa. Pomimo intensywnego rozwoju w ostatnich latach tej dziedziny w Polsce, aktualne informacje na temat sieci nie są jeszcze

łatwo dostępne (szczegółowy przegląd programów oraz informacji dotyczących sieci neuronowych zawiera praca Tadeusiewicza 1998).

13 AUTOMATYCZNA KLASYFIKACJA INTONACYJNEJ STRUKTURY FRAZY

13.1. WYPOWIEDZI IZOLOWANE

Podstawowym zadaniem modelowania struktur intonacyjnych jest klasyfikacja wyodrębnionych na drodze percepcyjnej jednostek melodycznych wyłącznie w zakresie typu akcentu rdzennego (w najprostszych intonacyjnie wypowiedziach zawierających tylko sylabę rdzenną lub sylabę rdzenną i sylaby następujące po niej).

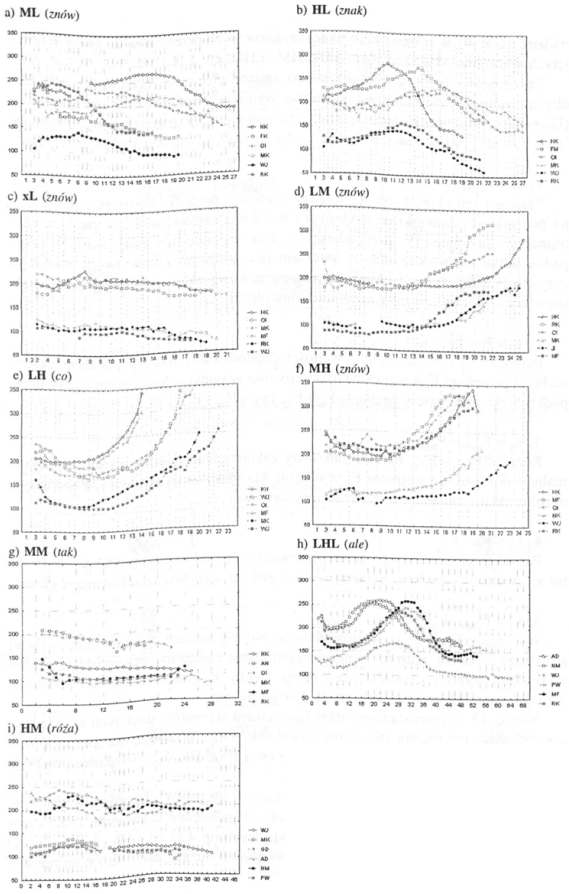
Wstępnie przyjęto do analiz wypowiedzi jedno oraz dwusylabowe z materiału zróżnicowanego fonetycznie, np. *znak, znów, tak, proszę…* itd. wymówione z 9 najczęstszymi typami akcentu rdzennego języka polskiego, traktowanymi jako wzorcowe. Wyniki analiz akustycznych tych wypowiedzi przedstawiono w rozdziale 8 niniejszej pracy. Do klasyfikacji przyjęto 1535 struktur jedno i dwusylabowych. Liczba replikacji poszczególnych wzorców mieściła się w zakresie 129- 150.

Klasyfikowanie przebiegu jako rosnącego, opadającego czy równego, w przypadku konturu intonacyjnego nie jest trywialne. Przebiegi percypowane jako opadające są najczęściej przebiegami rosnąco-opadającymi (ze wzrostem na spółgłosce i spadkiem na samogłosce), przebiegi odbierane przez słuchaczy jako rosnące składają się z fragmentu zawierającego spadek częstotliwości i fragmentu zawierającego mniejszy lub większy wzrost. Przebiegi percypowane jako równe — są w rzeczywistości przebiegami rosnącymi lub opadającymi.

Ryc. 13.1a — i ilustruje przykłady typów przebiegów intonacyjnych zaliczonych percepcyjnie do tych samych klas o nieidentycznym konturze i zakresie częstotliwości. Dla ilustracji wybrano przebiegi parametru F0 uzyskane dla wypowiedzi 6 osób, dysponujących różnymi skalami wysokości głosu oraz różnym tempem wypowiedzi. Przebiegi częstotliwości podstawowej przedstawiono w postaci danych odczytywanych co 20, 15 lub 10 ms (w zależności od wartości dolnej mierzonej częstotliwości, nie znormalizowane czasowo ani częstotliwościowo).

Dla wizualnego uwypuklenia różnic w poszczególnych realizacjach wzorców zastosowano liniową skalę częstotliwości.

Nawet pobieżna wizualna analiza zamieszczonych przykładów pozwala prze-



Rys 13.1. Przebiegi częstotliwości podstawowej w realizacji 9 tonów rdzennych

a) replikacja frazy *znów,* b) replikacja frazy *znak*, c) replikacja frazy *znów,* d) replikacja frazy *znów,* e) replikacja frazy *co*, f) replikacja frazy *znów,* g) replikacja frazy *tak*, h) replikacja frazy *ale*, i) replikacja frazy *róża*

widzieć trudności w zakresie klasyfikacji replikowanych przebiegów do grupy 9 przyjętych wzorców: MM, LM, ML, MH, HM, LHL, xL, LH i HL.

Poszczególne typy akcentu rdzennego opisano pięcioma cechami (x1…x5). Cechy te wybrano arbitralnie na podstawie wyników poprzednich eksperymentów (Demenko 1998) oraz badań z zakresu percepcji intonacji przedstawionych w pracy ’t Hart et al. (1990). Dwie pierwsze cechy x1 oraz x2 opisują kształt przebiegu.

1. x1= Fvp - Fe

Parametr ten określa różnicę między wartością początkową parametru F0 (Fvp na początkowej samogłosce struktury) i wartością częstotliwości w punkcie ekstremalnym przebiegu (Fe przypadającym na samogłosce lub spółgłosce). W przypadku braku ekstremum między punktem początkowym i końcowym przebiegu przyjęto Fe = Fvp (x1 =0). Jako wartość początkową założono częstotliwość podstawową w początkowym fragmencie samogłoski.

2. x2 = Fe - Fk

Parametr x2 opisuje różnicę wartości częstotliwości między punktem ekstremalnym przebiegu Fe a wartością częstotliwości w punkcie końcowym Fk. W przypadku braku ekstremum przyjęto Fe = Fvp (x2 = Fvp - Fk).

3. x3 = Fmax - Fmin

Różnica Fmax - Fmin określa całkowity zakres zmian między wartością maksymalną i minimalną parametru F0 w obrębie klasyfikowanej struktury. Parametr x3 opisuje zakres zmian częstotliwości podstawowej.

4. x4 = Fsr - Fsrg

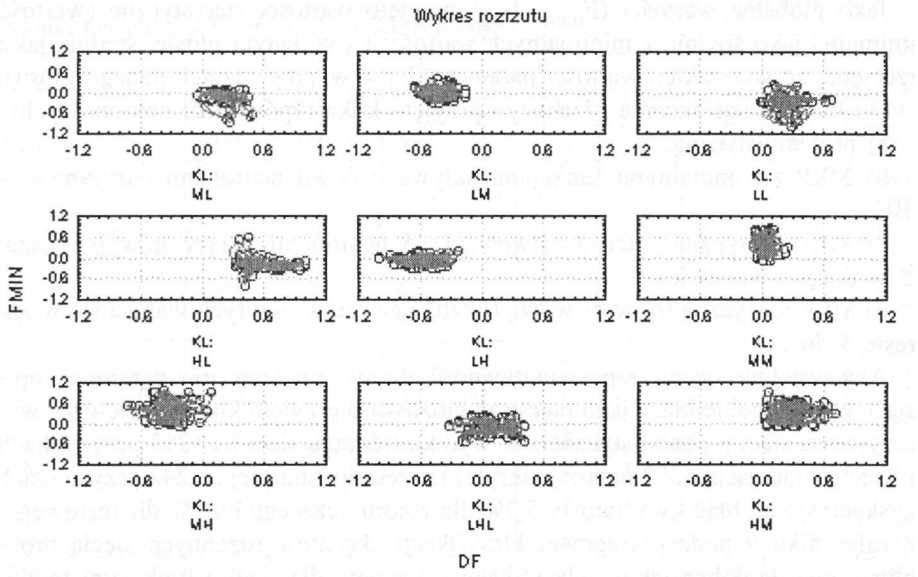
Parametr x4 określa różnicę między wartością średnią częstotliwości w strukturze i wartością średnią globalną we wszystkich wypowiedziach danego głosu.

5. x5 = Fmin - Fming

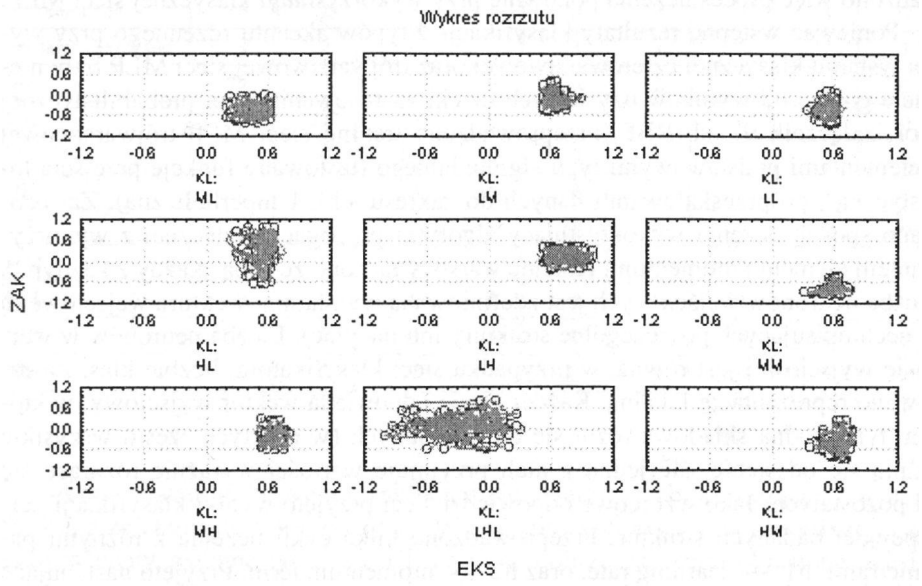
Wyrażenie Fmin - Fming opisuje różnicę między wartością minimalną w danej strukturze i średnią wartością minimalną globalną częstotliwości podstawowej we wszystkich wypowiedziach danego głosu. Cechy x4 oraz x5 określają położenie wzorca intonacyjnego na skali głosu danego mówcy.

Na ryc. 13.2a przedstawiono dane opisujące 9 akcentów rdzennych w układzie: częstotliwość minimalna (Fmin) oraz kierunek (DF). Współrzędna kierunek rozgranicza przebiegi rosnące od opadających np. LH od HL, Fmin rozdziela przebiegi niskie od wysokich, np. xL od MH.

Na ryc. 13.2b przedstawiono dane opisujące 9 akcentów rdzennych w układzie: zakres zmian (ZAK) oraz ekstremum (EKS). Współrzędna ekstremum (EKS) rozgranicza przebiegi rosnące od rosnąco-opadających np. LH od LHL, zakres (ZAK) przebiegi o małym zakresie zmian od przebiegów o dużym zakresie zmian np. LM od LH.



Ryc. 13.2a. Dane opisujące 9 akcentów nuklearnych we współrzędnych: kierunek (DF) oraz częstotliwość minimalna (Fmin)



Ryc. 13.2b. Dane opisujące 9 akcentów nuklearnych we współrzędnych: ekstremum (EKS) oraz zakres (ZAK)

Jako globalne wartości (Fming, Fsrg) przyjęto wartości statystyczne (wartość minimalną jako średnią z minimalnych wartości F0 w danym głosie, średnią jako przeciętną ze wszystkich wartości parametru F0 w wypowiedziach danego głosu).

Do klasyfikacji akcentu rdzennego przyjęto kilka typów sieci neuronowych:

- a. probabilistyczną,
  b. MLP z z radialnymi funkcjami aktywacji (z 90 neuronami ukrytymi) — RBF,
  c. MŁP klasyczną czterowarstwową (z 5 neuronami ukrytymi w pierwszej i 5 w drugiej warstwie),
  d. MŁP klasyczną trójwarstwową (liczba neuronów ukrytych wahała się w zakresie 5 do 20).

Aby wstępnie ocenić reprezentatywność zbioru uczącego oraz parametry opisujące poszczególne klasy jako pierwszą możliwość przyjęto klasyfikację przy wykorzystaniu sieci probabilistycznej. W wyniku treningu sieci na 1293 przypadkach (a więc posiadającej 1293 neurony ukryte) i przetestowaniu jej na 242 przypadkach uzyskano średni błąd kwadratowy 5,3% dla zbioru uczącego i 9,1% dla testowego. W załączniku 9 podano fragment klasyfikacji akcentów rdzennych siecią probabilistyczną. Podobną jakość klasyfikacji osiągnięto dla sieci z funkcjami radialnymi (60 neuronów ukrytych).

Klasyfikacja przeprowadzona przy wykorzystaniu sieci probabilistycznej wykazała, że zbiór uczący został właściwie przygotowany i opisany, jednak wady sieci probabilistycznej (duży rozmiar) wykluczają ją z praktycznych aplikacji. Przeprowadzono więc proces uczenia ponownie przy wykorzystaniu klasycznej sieci MŁP.

Ponieważ wstępne rezultaty klasyfikacji 9 typów akcentu rdzennego przy wykorzystaniu klasycznej czterowarstwowej oraz trójwarstwowej sieci MLP były niewiele gorsze od wyników uzyskanych dzięki zastosowaniu sieci probabilistycznej (por. załącznik 9) lub RBF, przeprowadzono trening sieci MLP trójwarstwowej z elementami podstawowymi typu sigmoidalnego (testowano funkcje przejścia logistyczną i po przeskalowaniu danych do zakresu +1-1 hiperboliczną). Zastosowano sposób uczenia wykorzystujący algorytm propagacji wstecznej z wykorzystaniem elementu momentum. Kolejne warstwy łączono ze sobą „każdy z każdym”. Liczba neuronów wejściowych jest zdefiniowana wymiarami wektora wejściowego

(5 cech opisujących poszczególne struktury intonacyjne). Liczba neuronów w warstwie wyjściowej jest równa, w przypadku sieci klasyfikatora, liczbie klas. Zastosowano reprezentację lokalną. Każdej klasie odpowiada wektor wyjściowy, w którym tylko jedna składowa różni się od pozostałych (w praktyce często wszystkie różnią się od siebie, ale jedna z nich przyjmuje wartości wyraźnie różniące się od pozostałych. Jako wzorcowe odpowiedzi sieci przyjęto wyniki klasyfikacji percepcyjnej badanych struktur. Przeprowadzono kilka cykli uczenia z różnymi parametrami: hl — learning rate, oraz h2 — momentum term. Przyjęto następujące współczynniki: hl =0,9; 0,6; 0,3; 0,09; 0,06; 0,03 oraz h2 = 0,6; 0,4; 0,2; 0,06; 0,03; 0,01 i w kierunku rosnących wartości. W każdej sesji treningowej podawano po 1000 prezentacji zbioru uczącego.

Tabela 13.1 Wyniki klasyfikacji 9 typów akcentów rdzennych

Zbiór uczący									
	ML	**LM**	**xL**	**HL**	**LH**	**MM**	**MH**	**LHL**	**HM**
Ogółem	**193**	**124**	**129**	**125**	**126**	**197**	**141**	**198**	**60**
Poprawne	**161**	**98**	**106**	**111**	**114**	**180**	**109**	**178**	**48**
Niepoprawne	**12**	**14**	**10**	**7**	**7**	**7**	**19**	**6**	**3**
Niesklasyfikowane	**20**	**12**	**3**	**7**	**5**	**10**	**13**	**14**	**9**
ML	**161**	**0**	**8**	**3**	**0**	**7**	**0**	**0**	**0**
LM	**0**	**98**	**0**	**0**	**6**	**0**	**6**	**0**	**0**
xL	**10**	**0**	**106**	**0**	**0**	**0**	**0**	**0**	**0**
HL	**0**	**0**	**0**	**111**	**0**	**0**	**0**	**6**	**3**
LH	**0**	**8**	**0**	**0**	**114**	**0**	**9**	**0**	**0**
MM	**2**	**0**	**2**	**0**	**0**	**180**	**4**	**0**	**0**
MH	**0**	**6**	**0**	**0**	**1**	**0**	**109**	**0**	**0**
LHL	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**178**	**0**
HM	**0**	**0**	**0**	**4**	**0**	**0**	**0**	**0**	**48**
Zbiór testowy									
	ML	**LM**	**xL**	**HL**	**LH**	**MM**	**MH**	**LHL**	**HM**
Ogółem	**38**	**26**	**24**	**29**	**28**	**33**	**14**	**33**	**17**
Poprawne	**29**	**22**	**17**	**26**	**25**	**29**	**11**	**28**	**13**
Niepoprawne	**6**	**2**	**3**	**1**	**0**	**3**	**1**	**3**	**3**
Niesklasyfikowane	**3**	**2**	**4**	**2**	**3**	**1**	**2**	**2**	**1**
ML	**29**	**0**	**3**	**0**	**0**	**2**	**0**	**0**	**0**
LM	**0**	**22**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
xL	**4**	**0**	**17**	**0**	**0**	**1**	**0**	**0**	**0**
HL	**0**	**0**	**0**	**26**	**0**	**0**	**0**	**3**	**0**
LH	**0**	**2**	**0**	**0**	**25**	**0**	**1**	**0**	**0**
MM	**0**	**0**	**0**	**0**	**0**	**29**	**0**	**0**	**3**
MH	**0**	**0**	**0**	**0**	**0**	**0**	**11**	**0**	**0**
LHL	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**28**	**0**
HM	**2**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**13**

Średni procent poprawnej klasyfikacji elementów ze zbioru uczącego wyniósł 85,5%, a ze zbioru testowego 82,6%. Wyższy procent poprawnej klasyfikacji uzyskały akcenty rdzenne typu LH, HL (wzorce dobrze wyraźnie określone akustycznie i percepcyjnie), niższy procent akcenty typu xL oraz ML. W tabeli 13.1 podano przykład klasyfikacji przy zastosowaniu progu akceptacji = 0,95 i odrzucenia = 0,05, które oznaczają w przypadku funkcji aktywacji większej niż 0,95 poprawne zaklasyfikowanie obiektu, mniejszej niż 0,05 — błędną klasyfikację a w przypadku funkcji aktywacji mieszczącej się między 0,05 -0,95 brak decyzji.

Klasyfikacja struktur melodycznych w wypowiedziach wielosylabowych jest problemem bardziej złożonym niż klasyfikacja akcentu rdzennego w izolowanych jedno- lub dwusylabowych frazach.

W szczególności trudność może stanowić:

- a. odróżnienie akcentu preiktycznego typu L od akcentu rdzennego typu LH i LM,
  b. odróżnienie akcentu preiktycznego typu H od akcentu rdzennego typu HL i ML,
  c. określenie danej sylaby (znajdującej się w pobliżu ekstremum lokalnego/globalnego przebiegu parametru F0 jako sylaby nieakcentowanej, akcentowanej preiktycznej lub rdzennej.

Klasyfikację przeprowadzono na podstawie następujących typów wypowiedzi:

— z akcentem rdzennym na początku lub końcu wielosylabowej frazy np.: ”Znowu ten owariat (intonacja pełna opadająca, HL).

— wypowiedzi zawierające jeden akcent preiktyczny (L lub H) np.:

To jest ˈjakiś ‘‘znak (intonacja preiktyczna rosnąca + intonacja rdzenna HL)

— złożone melodycznie wypowiedzi z kilkoma akcentami preiktycznymi np.:

To był ˈcałkiem ˈniezły i uczˈciwy człowiek (anakruza + intonacja preiktyczna rosnąca + intonacja rdzenna ML).

W rozdziale 8 opracowania przedstawiono przeanalizowany akustycznie powyższy materiał.

Na jego bazie przeprowadzono klasyfikację 12 różnych struktur intonacyjnych:

- a. struktury z akcentami rdzennymi typu: HL, ML, xL, HM,
  b. struktury z akcentami rdzennymi typu: LH, LM, MH ,
  c. struktury z akcentem rdzennym typu: LHL,
  d. struktury z akcentem rdzennym typu M,
  e. struktury z akcentem preiktycznym typu H,
  f. struktury z akcentem preiktycznego typu L,
  g. anakruza (oznaczona dalej jako P).

Jako początek każdej struktury przyjęto sylabę akcentowaną, jako koniec ostatnią sylabę występującą przed następną akcentowaną (lub ostatnią sylabę we frazie). Jako sylaby akcentowane uznano te sylaby, które zostały uwydatnione przez mówcę w wypowiedziach wzorcowych, a w imitacjach ocenione (na drodze eksperymentu percepcyjnego) jako podobne do wzorcowych. W przypadku anakruzy jako pierwszą sylabę struktury przyjęto pierwszą sylabę frazy, jako ostatnią sylabę struktury — ostatnią sylabę przed pierwszą sylabą akcentowaną we frazie.

Analizowane struktury mogły się więc składać z jednej sylaby lub ciągu sylab i występować w różnych kontekstach (z akcentem rdzennym na początku lub końcu frazy). Do klasyfikacji przyjęto 1930 struktur wielosylabowych (wybranych losowo 1630 do uczenia sieci i 300 do testowania).

Przyjęte do klasyfikacji struktury opisano 11-elementowym wektorem (x1′ … x11′) określonym cechami wymienionymi poniżej.

1. x1′ = Fvp-Fe

Parametr x1 ‘określa różnicę między wartością początkową parametru F0 (Fvp — na początkowej samogłosce struktury) i wartością częstotliwości w punkcie ekstremalnym przebiegu Fe. W przypadku braku ekstremum między punktem początkowym i końcowym przebiegu przyjęto Fe = Fvp.

2. x2′ = Fe-Fk

Parametr x2 określa różnicę między wartością ekstremalną przebiegu (Fe) a wartością częstotliwości w punkcie końcowym przebiegu (Fk).

Parametry X1′ oraz x2′ określają kształt struktury (przebieg rosnący, opadający, opadająco-rosnący).

3. x3′ = Fmax — Fmin

Różnica: Fmax - Fmin określa całkowity zakres zmian między wartością maksymalną i minimalną parametru F0 w obrębie danej struktury.

Cecha x3′ powinna odróżnić struktury o dużym zakresie zmian (np. LH) od struktur o małych zmianach (np. LM, MM).

4. x4′ = Fsr — Fsrg

Parametr x4 wyraża różnicę między wartością średnią częstotliwości w klasyfikowanej strukturze i wartością średnią globalną we frazie. Cecha ta powinna zróżnicować akcenty typu H i typu L.

5. x5′ = Fmin — Fming

Wyrażenie Fmin - Fming określa różnicę między wartością minimalną w danej strukturze i wartością minimalną globalną, wyznaczoną na podstawie wszystkich wypowiedzi danego głosu.

Cecha ta powinna ułatwić klasyfikację akcentu np. LM (z niskim Fmin) i MH (z wysokim Fmin).

Cechy x4′ i x5′ określają umiejscowienie danej struktury na skali głosu mówcy.

6. x6′ — Fve — Fke

Cecha x6′ określa zmianę parametru F0 na sylabie, na której występuje wartość ekstremalna częstotliwości w przebiegu (Fve — oznacza wartość na początku samogłoski, Fke — oznacza wartość częstotliwości na końcu sylaby). W intonacjach rdzennych typu rosnącego największa zmienność parametru występuje na końcu (w pobliżu maksimum globalnego). Najczęściej obserwuje się duży wzrost parametru F0 na ostatniej samogłosce we frazie. Parametr ten powinien ułatwić oddzielenie akcentu rdzennego rosnącego od akcentu preiktycznego typu L.

7. x7′ = |Fvp- Fk| - |Fva- Fka|

Parametr x7′ określa różnicę między bezwzględną zmianą częstotliwości w całej strukturze (Fvp-Fk, gdzie Fvp i Fk oznaczają odpowiednio początkową i końcową wartość parametru F0 w strukturze) i bezwzględną zmianą częstotliwości na sylabie akcentowanej (Fva - Fka).

8. x8′= |Fva - Fka| - |Fka - Fkr|

Parametr x8′ określa różnicę między zmianą częstotliwości podstawowej na sylabie akcentowanej (Fva - Fka), i zmianą (Fka - Fkr) określającą bezwzględną różnicę częstotliwości mierzoną od końcowej sylaby akcentowanej do końca frazy. Na sylabie rdzennej w intonacjach opadających występuje zwykle duża zmienność parametru F0 na początku struktury, wartość parametru x8′ powinna być więc większa dla akcentów rdzennych opadających, niż preiktycznych typu H.

Cechy x6′, x7′, oraz x8′ związane są z charakterystycznymi zmianami tonu w akcentach rdzennych.

9. x9′ = Dvi

Cecha x9′ określa znormalizowany iloczas ostatniej samogłoski w klasyfikowanej strukturze.

Na końcowej samogłosce frazy obserwuje się efekt wydłużenia. Spodziewać się więc można, że samogłoski w strukturach zawierających akcenty rdzenne czyli w strukturach końcowych frazy, będą dłuższe niż samogłoski w strukturach zawierających akcenty preiktyczne (w niekońcowych strukturach frazy). Na ryc. 13.3 zilustrowano znormalizowany iloczas samogłoski ostatniej w poszczególnych klasach struktur.

10. x10′ = AF/ADvi

Cecha X10′ określa stromość wzrostu/spadku częstotliwości podstawowej na ostatniej samogłosce w strukturze. Jako ΔF przyjęto Fvp - Fk, jako ΔDvi — iloczas samogłoski.

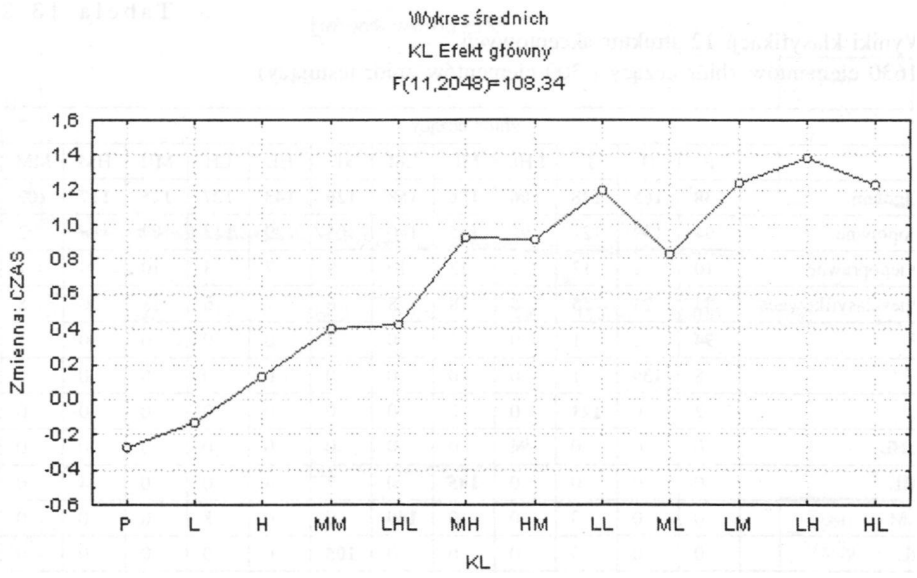
11. x11′ = Ei

Cecha X11′ określa znormalizowaną energię (względem średniej i odchylenia standardowego dla iloczasów samogłoskowych w wypowiedzi) ostatniej samogłoski w klasyfikowanej strukturze.

Cechy x9′, x10′ oraz X11′ związane są z iloczasem oraz energią wyznaczoną dla samogłosek.

Do klasyfikacji akcentu rdzennego przyjęto kilka typów sieci neuronowych:

- a. probabilistyczną,
  b. MLP z z radialnymi funkcjami aktywacji (z 90 neuronami ukrytymi) — RBF,
  c. MLP klasyczną czterowarstwową (z 6 neuronami ukrytymi w pierwszej i 6 w drugiej warstwie,



Ryc. 13.3. Znormalizowany iloczas ostatniej samogłoski w klasyfikowanych strukturach

- a. MLP klasyczna trójwarstwową (przyjęto liczbę neuronów w środkowej warstwie w zakresie 5 - 25).

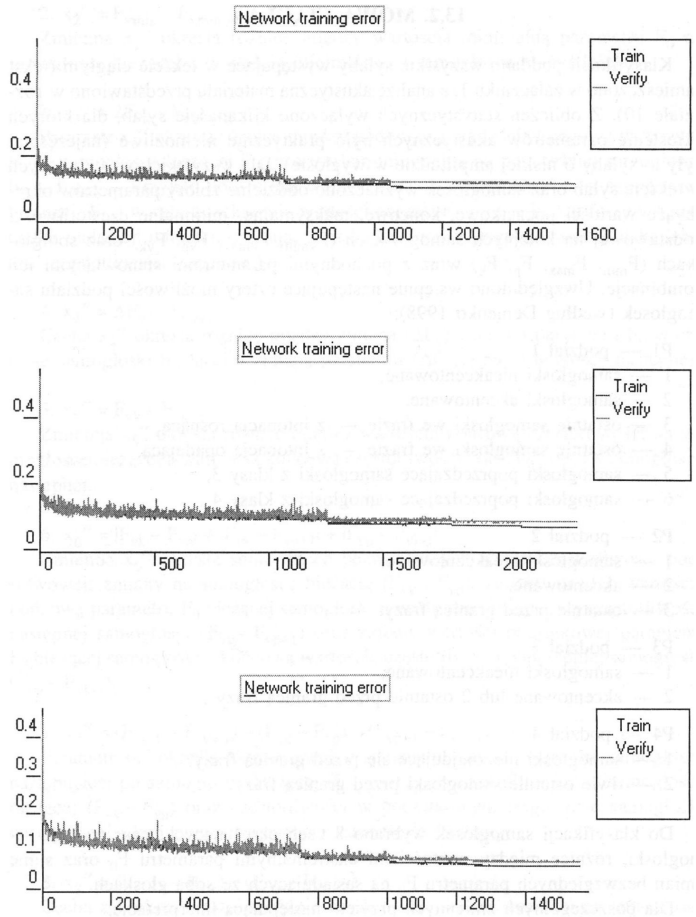
Uzyskano podobne wyniki klasyfikacji 12 typów akcentu rdzennego przy wykorzystaniu klasycznej, trój warstwowej sieci MLP, sieci probablistycznej oraz RBF.

Przeprowadzono szczegółowo trening sieci MLP trójwarstwowej (fragment wyników klasyfikacji przedstawiono w załączniku 10). Liczbę neuronów ukrytych zmieniano w zakresie 5 - 25. Przeprowadzono kilka cykli uczenia z różnymi parametrami hl — (learning rate) oraz h2 — (momentum term). Do uczenia zastosowano 1630 przypadków, a do testowania 300 przypadków. Wszystkie podawano w kolejności losowej. Na ryc. 13.4 podano wykresy błędów.

Przy 11 neuronach osiągnięto błąd globalny na zbiorze uczącym równy 0,08 a na zbiorze testowym 0,09 (Eu = 0,08, Et = 0,09, ryc. 13.4). Przy 16 neuronach błąd na zbiorze uczącym wyniósł 0,069, na testowym 0,08. Dodawanie większej liczby neuronów nie poprawiło wyników uczenia sieci. W załączniku 10 przedstawiono fragment klasyfikacji kilkudziesięciu struktur akcentowych. Dla zbioru uczącego otrzymano poprawną klasyfikację w 83%, dla testowego w 80%. Najgorzej zostały sklasyfikowane sylaby nieakcentowane na początku wypowiedzi (klasa P): tylko 67% dla zbioru uczącego i 60% dla zbioru testowego. Dobrze sklasyfikowano akcenty LH, HL oraz LHL (por. szczegółowe dane w tabeli 13.2).

Tabela 13.2 Wyniki klasyfikacji 12 struktur akcentowych (1630 elementów zbiór uczący i 300 elementów zbiór testujący)

zbiór uczący												
	P	**H**	**L**	**LHL**	**ML**	**LM**	**xL**	**HL**	**LH**	**MH**	**HM**	**MM**
Ogółem	**138**	**165**	**148**	**106**	**176**	**158**	**120**	**143**	**127**	**125**	**117**	**107**
Poprawne	**94**	**139**	**121**	**96**	**145**	**140**	**105**	**129**	**113**	**102**	**88**	**92**
Niepoprawne	**10**	**2**	**12**	**2**	**13**	**10**	**9**	**7**	**8**	**10**	**9**	**7**
Niesklasyfikowane	**34**	**24**	**15**	**8**	**18**	**8**	**6**	**7**	**6**	**13**	**20**	**8**
P	**94**	**2**	**1**	**0**	**1**	**0**	**2**	**0**	**0**	**0**	**0**	**7**
H	**8**	**139**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
L	**2**	**0**	**121**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
LHL	**0**	**0**	**0**	**96**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
ML	**0**	**0**	**0**	**0**	**145**	**0**	**7**	**4**	**0**	**0**	**4**	**0**
LM	**0**	**0**	**0**	**0**	**0**	**140**	**0**	**0**	**8**	**0**	**0**	**0**
xL	**0**	**0**	**0**	**0**	**0**	**0**	**105**	**0**	**0**	**0**	**0**	**0**
HL	**0**	**0**	**0**	**2**	**10**	**0**	**0**	**129**	**0**	**0**	**0**	**0**
LH	**0**	**0**	**0**	**0**	**0**	**5**	**0**	**0**	**113**	**8**	**0**	**0**
MH	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**102**	**0**	**0**
HM	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**3**	**0**	**0**	**88**	**0**
MM	**0**	**0**	**10**	**0**	**2**	**5**	**0**	**0**	**0**	**2**	**5**	**92**
zbiór testowy												
	P	**H**	**L**	**LHL**	**ML**	**LM**	**xL**	**HL**	**LH**	**MH**	**HM**	**MM**
Ogółem	**21**	**29**	**25**	**30**	**35**	**37**	**18**	**27**	**20**	**22**	**15**	**20**
Poprawne	**13**	**21**	**18**	**26**	**30**	**31**	**15**	**23**	**17**	**17**	**12**	**16**
Niepoprawne	**3**	**5**	**6**	**2**	**2**	**5**	**2**	**2**	**2**	**4**	**1**	**3**
Niesklasyfikowane	**5**	**3**	**1**	**2**	**3**	**1**	**1**	**2**	**1**	**1**	**2**	**1**
P	**13**	**0**	**0**	**0**	**0**	**2**	**0**	**0**	**0**	**0**	**0**	**3**
H	**2**	**21**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
L	**1**	**0**	**18**	**0**	**0**	**3**	**0**	**0**	**0**	**0**	**0**	**0**
LHL	**0**	**0**	**0**	**26**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
ML	**0**	**0**	**3**	**0**	**30**	**0**	**2**	**2**	**0**	**0**	**0**	**0**
LM	**0**	**0**	**0**	**0**	**0**	**31**	**0**	**0**	**1**	**0**	**0**	**0**
xL	**0**	**0**	**3**	**0**	**0**	**0**	**15**	**0**	**0**	**0**	**0**	**0**
HL	**0**	**0**	**0**	**2**	**1**	**0**	**0**	**23**	**0**	**0**	**0**	**0**
LH	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**17**	**0**	**0**	**0**
MH	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**17**	**0**	**0**
HM	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**12**	**0**
MM	**0**	**5**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**4**	**1**	**16**



Ryc. 13.4. Wykres błędu podczas treningu sieci MLP a) 7 neuronów (Eu = 0,11, Et = 0,13), b) 11 neuronów (Eu = 0,08, Et = 0,09), c) 16 neuronów (Eu = 0,069, Et = 0,08)

13.2. MOWA CIĄGŁA

Klasyfikacji poddano wszystkie sylaby występujące w tekście ciągłym (tekst zamieszczono w załączniku 1, a analizę akustyczną materiału przedstawiono w rozdziale 10). Z obliczeń statystycznych wyłączono kilkanaście sylab, dla których określenie parametrów akustycznych było praktycznie niemożliwe (najczęściej były to sylaby o niskiej amplitudzie w wygłosie). Dla wszystkich występujących w tekście sylab oraz samogłosek wyznaczono oddzielne zbiory parametrów określające wartości początkowe, końcowe, maksymalne, minimalne częstotliwości podstawowej na kolejnych samogłoskach (Fvmin, Fvmax, Fvp, Fvk) oraz spółgłoskach (Fmin, Fmax, Fp, Fk) wraz z pochodnymi parametrami stanowiącymi ich kombinacje. Uwzględniono wstępnie następujące cztery możliwości podziału samogłosek (według Demenko 1998):

P1 — podział 1
1 — samogłoski nieakcentowane,
2 — samogłoski akcentowane,
3 — ostatnie samogłoski we frazie — z intonacją rosnącą,
4 — ostatnie samogłoski we frazie — z intonacją opadającą,
5 — samogłoski poprzedzające samogłoski z klasy 3,
6 — samogłoski poprzedzające samogłoski z klasy 4 .

P2 — podział 2

1 — samogłoski nieakcentowane,

2 — akcentowane,

3 — ostatnie przed granicą frazy.

P3 — podział 3

1 — samogłoski nieakcentowane,

2 — akcentowane lub 2 ostatnie przed granicą frazy .

P4 — podział 4

1 — samogłoski nie znajdujące się przed granicą frazy,

2 — dwie ostatnie samogłoski przed granicą frazy.

Do klasyfikacji samogłosek wybrano 8 cech określających: czas trwania samogłoski, różnicę między wartościami ekstremalnymi parametru F0 oraz sumę zmian bezwzględnych parametru F0 na sąsiadujących ze sobą głoskach.

Dla poszczególnych zmiennych przyjęto następującą interpretację:

1. x1□ = DV

Cecha ta grupuje iloczasy samogłosek. Największe różnice w parametrach statystycznych zauważono między iloczasami samogłosek nieakcentowanych i niekońcowych we frazie oraz iloczasami samogłosek znajdujących się bezpośrednio przed granicą frazy.

2. x2□ = Fvmin — Fvmin -1

Zmienna x2□ określa różnicę między wartością minimalną parametru F0 na samogłosce bieżącej a wartością minimalną na samogłosce poprzedniej.

3. x3□ = 2Fvk-Fvmin -1

Parametr x3□ określa różnicę między podwojoną wartością końcową parametru F0 na samogłosce bieżącej Fvk a wartością minimalną na poprzedniej samogłosce Fvk - Fvmin -1. Przyjęto podwojoną wartość końcową parametru F0 na samogłosce bieżącej (Fvk) w celu zwiększenia wartości parametru x3□ (na końcu frazy w przypadku realizacji intonacji opadającej na końcu frazy występuje zawsze niska wartość częstotliwości końcowej).

4. x4□ = ΔFv - ΔFv+1

Cecha x4□ określa różnicę między zmianą (ΔFv) wartości parametru F0 w obrębie samogłoski bieżącej a zmianą parametru (ΔFv+1) na samogłosce następnej.

5. x5□ = Fvk - Fvk+1

Zmienna x5□ określa różnicę między wartością końcową częstotliwości na samogłosce bieżącej a wartością końcową częstotliwości podstawowej na samogłosce następnej.

6. x6□ = |Fvk - Fvp| + |Fvk - Fvp+1| + |Fvp - Fvk-1|

Zmienna x6□ określa sumę trzech bezwzględnych zmian częstotliwości podstawowej: zmiany na samogłosce bieżącej (Fvk-Fvp), zmiany między wartością końcową parametru F0 bieżącej samogłoski a wartością początkową częstotliwości następnej samogłoski (Fvk-Fvp+1) oraz zmiany wartości początkowej parametru F0 bieżącej samogłoski a końcową wartością częstotliwości poprzedniej samogłoski (Fvp-Fvk-1)·

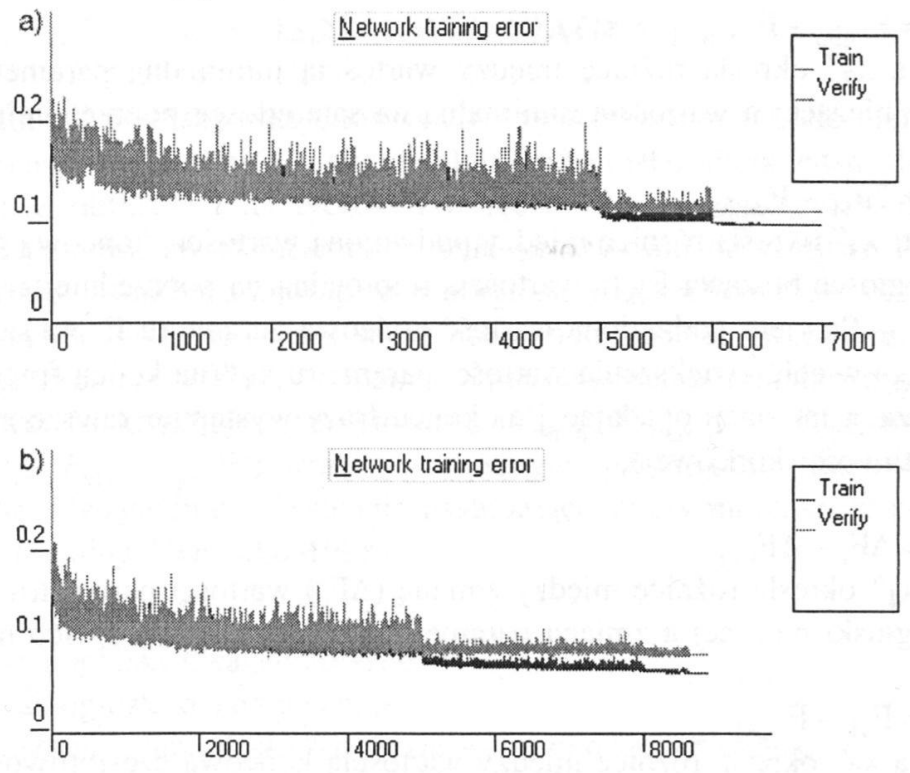
7. x7□ = (Fvk-1 - Fvp-1) - (Fvk - Fvp) - (Fvp+1 - Fvk+1)

Parametr x7□ określa różnicę między zmianami: częstotliwości na samogłosce następującej po samogłosce bieżącej (Fvk-1 - Fvp-1), częstotliwości na samogłosce bieżącej (Fvk-Fvp) oraz częstotliwości w początkowym fragmencie samogłoski następującej i końcowym fragmencie samogłoski bieżącej.

8. x8□ = (Fvsr-1 - Fvsr) - (Fvsr - Fvsr+1)

Cecha x8□ określa różnicę różnic między wartościami średnimi na samogłosce bieżącej i poprzedniej oraz bieżącej i następnej.

Do klasyfikacji badanych struktur zastosowano klasyczną trójwarstwową sieć typu MLP. Liczba neuronów wejściowych jest zdefiniowana wymiarami wektora wejściowego (8 cech). Przyjęto podział P2 (samogłoska akcentowana, nieakcentowana, ostatnia lub przedostatnia przed granicą frazy). Przeprowadzono uczenie sieci z 2, 4, 8 i 20 elementami w warstwie ukrytej. Najlepsze rezultaty otrzymano dla



Ryc. 13.5. Błąd uczenia sieci: a) z 9 neuronami ukrytymi, b) z 15 neuronami ukrytymi

20 neuronów. Elementami zbioru uczącego są pary złożone z wektorów sygnałów wejściowych (informacji podawanej na wejście sieci — opisanej 8 cechami) i wymaganych sygnałów wyjściowych (wzorcowych odpowiedzi sieci). Jako wzorcowe odpowiedzi sieci przyjęto wyniki klasyfikacji percepcyjnej badanych struktur.

Przeprowadzono kilka cykli uczenia z różnymi parametrami hl oraz h2. Uzyskano wysoki procent poprawnej klasyfikacji elementów ze zbioru uczącego: samogłoski nieakcentowane — 91%, akcentowane 86% oraz końcowe frazy 84% (por. załącznik 11 — dane dla pierwszych 106 sylab). Znacznie gorsze wyniki uzyskano dla rozpoznawania. Samogłoski nie pochodzące ze zbioru uczącego zostały średnio prawidłowo rozpoznane w 75%.

Dla klasyfikacji określonych w rozdziale 13.1 struktur akcentowych przeprowadzono dodatkowe doświadczenie odsłuchowe, w którym słuchacze (3 fonetyków) określali typy akcentu w mowie ciągłej. Z uwagi na niską liczebność niektórych typów akcentów rdzennych (np. HM, MH), słuchacze nie przeprowadzili szczegółowej klasyfikacji (nie wyróżniali np. akcentu typu HL, ML, xL, HM), oznaczali jedynie akcent jako opadający (F), rosnący (R) lub równy (MM). Jedna z odsłuchujących osób przeprowadziła szczegółową percepcyjną ocenę akcentów preiktycznych. W ogólnym wyniku doświadczenia otrzymano następujące typy akcentu: H (496 przypadki), L (35 przypadków), R (209 przypadków), F (175 przypadków), MM (34 przypadki) oraz anakruzę P (78 przypadków). Najczęściej występował akcent poboczny typu H oraz akcenty rdzenne rosnące i opadające. Do przetestowania możliwości automatycznego rozpoznania akcentów mowy ciągłej, wykorzystano sieć wyuczoną na frazach izolowanych wielosylabowych (rozdz. 13). Sieć tę poddano ponownemu procesowi uczenia na przykładach pochodzących z fraz izolowanych. Sygnały wejściowe nie uległy zmianie, tylko na wyjściu sieci określono te kategorie, które otrzymano na podstawie tekstów czytanych: H, L, R, F, MM oraz dodatkowo przyjęto klasę P określającą anakruzę. Modyfikacja polegała więc na połączeniu klas MH, LM, LH w klasę R oraz klas HL, HM, xL, ML w klasę F.

Jako zbiór testowy przyjęto klasyfikacje percepcyjne z mowy ciągłej. Otrzymano średni procent poprawnej klasyfikacji w granicach 79 — 83% zależnie od typu akcentu. Najlepsze wyniki uczenia sieci otrzymano przy 9 neuronach ukrytych. Ryc. 13.5 ilustruje wykresy błędów dla 9 i 15 neuronów ukrytych). Wyniki klasyfikacji dla zbioru uczącego i testowego przedstawiono w tabeli 13.3.

Tabela 1 3 .3 Wyniki klasyfikacji 6 struktur akcentowych

Zbiór uczący — wypowiedzi izolowane						
	P	**H**	**L**	**MM**	**R**	**F**
Ogółem	**138**	**163**	**148**	**107**	**556**	**410**
Poprawne	**79**	**134**	**116**	**90**	**495**	**380**
Niepoprawne	**8**	**4**	**10**	**6**	**21**	**9**
Niesklasyfikowane	**51**	**25**	**22**	**11**	**40**	**21**
P	**79**	**2**	**0**	**0**	**0**	**0**
H	**6**	**134**	**0**	**6**	**0**	**9**
L	**1**	**0**	**116**	**0**	**20**	**0**
MM	**0**	**1**	**10**	**90**	**1**	**0**
R	**1**	**1**	**0**	**0**	**495**	**0**
F	**0**	**0**	**0**	**0**	**0**	**380**
Zbiór testowy — mowa ciągła						
	P	**H**	**L**	**MM**	**R**	**F**
Ogółem	**85**	**496**	**56**	**34**	**175**	**209**
Poprawne	**60**	**412**	**36**	**28**	**148**	**181**
Niepoprawne	**3**	**44**	**5**	**3**	**14**	**13**
Niesklasyfikowane	**22**	**40**	**15**	**3**	**13**	**15**
P	**60**	**20**	**5**	**3**	**7**	**0**
H	**0**	**412**	**0**	**0**	**7**	**0**
L	**3**	**0**	**36**	**0**	**0**	**9**
MM	**0**	**24**	**0**	**28**	**0**	**4**
R	**0**	**0**	**0**	**0**	**148**	**0**
F	**0**	**0**	**0**	**0**	**0**	**181**

14 SYNTEZA PRZEBIEGÓW INTONACYJNYCH W MOWIE CIĄGŁEJ

14.1. ZAGADNIENIA PODSTAWOWE

Istnieje obecnie wiele technicznych możliwości syntezy sygnału mowy. Do najczęściej wykorzystywanych w praktyce metod należą: artykulacyjna — modelująca wytwarzanie sygnału mowy, formantowa — wykorzystująca bezpośrednio akustyczne cechy sygnału oraz konkatenacyjna — polegająca na łączeniu krótkich segmentów sygnału w dłuższe jednostki (np. demisylab w sylaby, sylab w wyrazy itp.). Bez względu na stosowany typ syntezy elementów segmentalnych mowy modelowanie intonacji ważne jest z kilku zasadniczych powodów.

1. Intonacja wpływa na zrozumiałość mowy. Spełnia funkcję segmentacyjną wypowiedzi i ułatwia słuchaczowi wyodrębnianie z ciągłego sygnału mowy przekazywanych przez mówcę poszczególnych informacji.

2. Błędy w budowie segmentalnej są przez słuchacza w większym stopniu tolerowane niż błędy w strukturze suprasegmentalnej wypowiedzi. Niewłaściwe miejsce wystąpienia akcentu, bądź nieprawidłowy typ akcentu może całkowicie zmienić sens wypowiedzi lub wywołać wrażenie nienaturalności. Lepszym rozwiązaniem w syntezie jest modelowanie monotonnej intonacji niż nieodpowiednie odwzorowywanie cech melodycznych wypowiedzi.

3. Dla uzyskania mowy wysokiej jakości niezbędne jest poprawne kształtowanie cech prozodycznych. Słuchacze z trudem akceptują mowę monotonną, ponieważ wymaga ona od nich dużo większej koncentracji uwagi niż odbiór wypowiedzi naturalnych.

Problematyka związana z modelowaniem intonacji dla syntezy mowy obejmuje trzy następujące podstawowe zagadnienia:

1. Wybór sterowania sekwencją tonów (kolejność akcentów, typ akcentu oraz synchronizacja czasowa zmian tonu względem własności segmentalnych). Problem ten jest stosunkowo dobrze rozwiązany (zwłaszcza dla języka angielskiego, niemieckiego, francuskiego, holenderskiego i japońskiego). Tradycyjnie najlepiej rozwinięta została synteza z reguł, zwykle stosowana do sterowania zmianami wysokości tonu w układach typu „text to speech”, w których dokonuje się automatycznie konwersji tekstu ortograficznego na odpowiedni sygnał akustyczny. Istnieje co najmniej kilkadziesiąt algorytmów teoretycznych i implementacji praktycznych sterowania intonacją w mowie czytanej opracowanych dla różnych języków. Do najciekawszych rozwiązań należą systemy: INVOVOX — system syntezy text- to-speech opracowany dla języków: angielskiego, niemieckiego, francuskiego, hiszpańskiego, szwedzkiego i włoskiego, DECTALK — system przetwarzania znaków ASCII w naturalnie brzmiącą mowę (posiada możliwość wytworzenia 4 typów głosu kobiecego, 4 głosów męskich i 1 dziecięcego), HADIFIX — synteza konkatenacyjna dla języka niemieckiego, MBROLA jest systemem syntezy wysokiej jakości (porównywalnej z jakością syntezy PSOLI) opartej na difonach z przeznaczeniem dla wielu języków (np. angielskiego, hiszpańskiego, włoskiego i holenderskiego).

2. Uwydatnianie intonacyjne. Dotyczy ono podkreślania intonacyjnego szczególnie istotnych dla mówcy fragmentów zdania, może być także związane z mo-

delowaniem informacji paralingwistycznych (np. Bolinger 1989). Zagadnienie uwzględniania w syntezie mowy informacji paralingwistycznych oraz pozajęzykowych stanowi aktualnie na świecie ważny problem (por. np. Sagisaka et al. 1997). Jego rozwiązanie jest niezbędne dla uzyskania syntezy wysokiej jakości.

3. Globalne cechy intonacji. Nowoczesne układy syntezy wymagają również opracowania modelowania różnych zakresów zmian częstotliwości podstawowej, rejestrów oraz normalizacji percepcyjnej konturu intonacyjnego w obrębie frazy.

14.2. STEROWANIE CZĘSTOTLIWOŚCIĄ PODSTAWOWĄ W SYNTEZIE MOWY POLSKIEJ

Problem sterowania częstotliwością podstawową w syntezie mowy polskiej nie jest w sposób zadowalający rozwiązany. Nieliczne opracowania z tej dziedziny obejmują swym zakresem głównie wypowiedzi izolowane i dostarczają tylko fragmentarycznych wskazówek, które mogą być zaimplementowane w syntezie (np. Kacprowski 1965, Jassem et al. 1968, Myślecki 1979, Jassem et al. 1990). W tej sytuacji dla sformułowania zasad sterowania parametrem F0 w mowie ciągłej konieczne stało się wykorzystanie opracowań dla innych języków (por. np. de Pijper 1983, 't Hart et al. 1990, Möbius 1993 oraz Portele 1997), jak i własnych badań. Wstępne reguły modelowania zmian wysokości tonu w mowie ciągłej sformułowano w pracy Demenko et al. (1993). Na ich podstawie opracowano algorytm, który został wdrożony do praktycznego układu syntezy o nazwie „Kubuś"[7].

Założono, że program realizujący kształtowanie konturów intonacyjnych powinien uwzględniać następujące rodzaje informacji:

1. Dane opisujące zdanie.

a) Liczba fraz

Zdania mogą składać się z jednej lub kilku fraz. Liczba fraz wchodzących w skład zdania określa jego stopień złożenia i ma wpływ na sterowanie dynamiką zmian parametru F0.

b) Struktura frazy

Frazy mogą posiadać odmienne struktury, wynikające z liczby oraz rozkładu sylab akcentowanych. Struktura frazy ma bezpośredni wpływ na sposób sterowania wysokością tonu.

c) Pozycja frazy

Pierwsze frazy i końcowe zdania są szczególnie istotne w modelowaniu intonacji, określają dynamikę przebiegu i typ wypowiedzi.

d) Zakończenie frazy

Frazy mogą kończyć się następującymi znakami interpunkcyjnymi: [,.?!—].

2. Dane opisujące frazę.

a) Liczba akcentów

Frazy mogą posiadać odmienne struktury wynikające z liczby, pozycji oraz z rodzaju sylab akcentowanych preiktycznych.

b) Pozycja akcentu

Pierwszy akcent preiktyczny oraz akcent rdzenny odgrywają szczególnie istotną rolę, określają dynamikę zmian oraz typ wypowiedzi.

c) Długość frazy

Wyróżnia się 7 kategorii długości frazy wyrażonej w sekundach (0-1,5 s, 1,5 — 2,5 s, 2,5-3,5 s, 3,5-4,5 s, 4,5-5,5 s, 5,5-6,5 s oraz powyżej 6,5 s).

3. Dane opisujące sylabę

[7] System opracowano w Zakładzie Fonetyki Akustycznej IPPT PAN w Poznaniu (J. Imiołczyk, I. Nowak, G. Demenko 1993).

Samogłoski mogą być poprzedzone zbitkami spółgłoskowymi o różnej długości i różnej strukturze.

Przyjęto wstępnie możliwość sterowania częstotliwością podstawową według modelu Fujisaki (1981, 1983, 1988). Model ten zakłada superpozycję składowej frazowej (określającej deklinację) i składowych akcentowych, wyznaczonych dla poszczególnych sylab akcentowanych (por. rozdz. 5).

Funkcję Gpi, sterującą frazą opisano zależnością (14.1)

$$G_{pi}(t) = K_{pi}\,\alpha_i \exp(-\alpha_i t) \qquad (14.1)$$

Funkcję Gaj sterującą składową akcentową opisano zależnością (14.2)

Dla ustalenia współczynników funkcji sterujących składową akcentową i frazową przeprowadzono analizę akustyczną i statystyczną częstotliwości podstawowej w kilkuminutowych gazetowych tekstach czytanych przez 6 osób oraz dwóch zestawach zdań (por. załącznik 12 oraz Demenko 1995a). Celem badań była statystyczna ocena podobieństw w przebiegach parametru F0 w replikacjach tego samego zdania przez różnych mówców. Długość fraz zmieniała się w zakresie od 3 do 56 sylab. Współczynnik korelacji między przebiegami częstotliwości podstawowej w replikacjach tej samej wypowiedzi okazał się dość wysoki (w granicach 0,67-0,96), co pozwoliło na statystyczne uśrednianie zmian częstotliwości podstawowej. Największy zakres zmian parametru wystąpił na pierwszej akcentowanej sylabie — 56 Hz (średnio dla 31-sylabowego zdania) i 62 Hz (dla zdania 56-sylabowego). Wewnątrz frazy zakresy zmian tonu na kolejnych sylabach stopniowo malały, od 28 Hz do 19 Hz w krótkich zdaniach i od 39 Hz do 15 Hz w dłuższych. We wszystkich analizowanych zdaniach znaczna zmiana częstotliwości podstawowej wystąpiła również na sylabie rdzennej (przeważnie powyżej 50 Hz).

Analiza statystyczna wykazała, że wartość początkowa częstotliwości podstawowej zależy od struktury początku frazy i długości zdania. Zależnie od długości frazy przyjęto współczynniki wzmocnienia Kpi w zakresie 0,018 -0,633 oraz tłumienia w przedziale 1,14-8,00. Ustalono zbiór maksymalnych wartości funkcji frazowych aproksymujących zmiany tonu w zakresie 100- 124 Hz.

Wyznaczono 3 typy linii deklinacyjnej (niski, średni oraz wysoki) i w każdym z nich rozróżniono 7 konfiguracji parametrów Kpi zależnie od długości frazy. Przykładowo, pierwszej frazie długiego zdania przypisano maksymalną wartość współczynnika wzmocnienia Kpi (0,633).

Przyjęto 14 współczynników Kaj (pokrywających zakres 6-84 Hz) i 3 kategorie wartości współczynnika tłumienia βj (w zakresie 6,97 - 122) co pozwoliło modelować wolne, szybkie i bardzo szybkie zmiany parametru F0. W tabeli 14.1 oraz tabeli 14.2 przedstawiono współczynniki funkcji frazowych i akcentowych.

Typowe przebiegi aproksymujące zmienność częstotliwości podstawowej przedstawiono w załączniku 13.

Frazę podzielono na 3 części: wstępną — zawierającą pierwszy akcent preiktyczny, środkową — obejmującą następne akcenty poboczne oraz końcową — zawierającą ostatni akcent preiktyczny i akcent rdzenny. Ogólny schemat modelu dla jednofrazowego zdania oznajmującego, przedstawiono na ryc. 14.1.

$$G_{aj}(t) = K_{aj}\,(1 - (1 + \beta_j \exp(-\beta_j t))) \qquad (14.2)$$

gdzie: K_{aj}, K_{pi} — oznaczają współczynniki wzmocnienia,
α_i, β_j — współczynniki tłumienia,
i, j — numer kolejnego akcentu,
t — czas.

Tabela 14.1 Współczynniki sterujące frazą dla poszczególnych typów linii deklinacyjnej

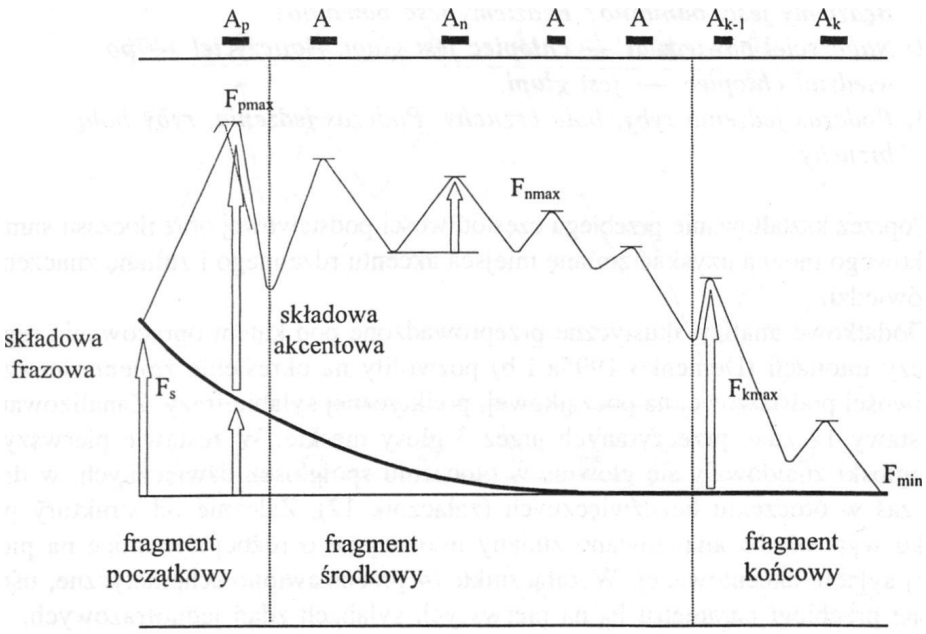
L.p.	NISKI			ŚREDNI			WYSOKI		
	ai	Kp	[Hz]	ai	Kp	[Hz]	αi	Kp	[Hz]
1	8,00	0,018	100	8,00	0,037	106	8,00	0,056	112
2	4,00	0,051	102	4,00	0,088	108	4,00	0,124	114
3	2,64	0,096	104	2,67	0,150	110	2,67	0,203	116
4	2,00	0,155	106	2,00	0,222	112	2,00	0,295	118
5	1,60	0,222	108	1,60	0,307	114	1,60	0,395	120
6	1,32	0,300	110	1,32	0,410	116	1,32	0,516	122
7	1,14	0,385	112	1,14	0,513	118	1,14	0,633	124

Tabela 14.2 Współczynniki sterujące akcentem w przypadku bardzo szybkich zmian częstotliwości podstawowej, szybkich oraz wolnych

L.p.	Współczynniki sterujące akcentem	Zmiana		
		bardzo szybka	szybka	wolna
	Ka	βes	βs	βw
1	0,0661	122,0	61,00	30,50
2	0,1236	95,92	47,96	23,98
3	0,1779	77,86	38,93	19,96
4	0,2294	65,78	32,93	16,95
5	0,2784	57,16	28,58	14,27
6	0,3251	50,72	25,36	12,68
7	0,3697	45,72	22,86	11,43
8	0,4124	41,70	20,85	10,43
9	0,4534	38,36	19,18	9,59
10	0,4927	35,62	17,81	8,91
11	0,5306	33,22	16,62	8,31
12	0,5671	31,18	15,59	7,80
13	0,6023	29,40	14,70	7,35
14	0,6362	27,88	13,94	6,97

Oznaczenia Ap, An oraz Ak określają odpowiednio początkową, n-tą oraz końcową sylabę akcentowaną. Wartości częstotliwości na sylabach nieakcentowanych leżą poniżej linii łączącej sąsiednie sylaby akcentowane.

Praktyczna implementacja reguł sterowania częstotliwością podstawową według założonych funkcji wykazała małą elastyczność w formowaniu konturu (poprawnie udało się tylko modelowanie dwóch akcentów rdzennych HL i ML). Istotne trudności sprawiała również synchronizacja czasowa maksimum funkcji względem początku/środka/końca samogłoski.



Ryc 14.1. Modelowanie intonacji w jednofrazowym zdaniu oznajmującym

Dla modyfikacji zastosowanego modelu przeprowadzono szereg testów odsłuchowych (Demenko 1995b), mających na celu ustalenie wpływu różnych realizacji akustycznych określonej wypowiedzi na percepcję syntetycznego akcentu. W eksperymentach wykorzystano metodę resyntezy liniowej predykcji LPC. Standardowa konfiguracja analizy spektrografu cyfrowego Kay 5500 pozwala na resyntezę wypowiedzi (pojedynczych fram sygnału lub całej wypowiedzi) metodą kowariancji lub korelacji, ustalenie liczby współczynników predykcji, określenia długości framy LPC, wyznaczenia współczynników emfazy lub preemfazy sygnału. Dla modelowania przebiegów parametru F0 wykorzystano opcję wpisywania wartości tego parametru do kolejnych 10 milisekundowych fram sygnału. Oprogramowanie spektrografu umożliwia natychmiastowy odsłuch uzyskanej wypowiedzi syntetycznej i naturalnej, analizę widmową oraz korektę danych za pomocą numerycznego edytora. Jakość wszystkich wykorzystanych w pracy syntetycznych wypowiedzi oceniano słuchowo i na bieżąco optymalizowano. Testy audytywne dotyczyły percepcyjnej oceny wpływu miejsca ekstremum w przebiegu częstotliwości podstawowej oraz dynamiki i szybkości zmian parametru F0 na akcent. Dodatkowo analizowano iloczas samogłosek akcentowanych.

Resyntezie poddano kilka par wypowiedzi, w których umiejscowienie oraz dynamika i szybkość zmian wysokości tonu na określonej sylabie decydowały o znaczeniu zdania.

1. *Będziemy jeść, **bambino** ? Będziemy **jeść** bambino?*

2. *Nauczyciel powiedział — chłopiec jest głupi. Nauczyciel — po wiedział chłopiec — jest głupi.*

3. *Podczas jedzenia ryby, bolą brzuchy. Podczas jedzenia, ryby bolą brzuchy.*

Poprzez kształtowanie przebiegu częstotliwości podstawowej oraz iloczasu samogłoskowego można uzyskać zmianę miejsca akcentu rdzennego i zmianę znaczenia wypowiedzi.

Dodatkowe analizy akustyczne przeprowadzone pod kątem opracowania reguł syntezy intonacji (Demenko 1995a i b) pozwoliły na określenie zmienności częstotliwości podstawowej na początkowej, preiktycznej sylabie frazy. Zanalizowano 2 zestawy 12 zdań przeczytanych przez 3 głosy męskie. W zestawie pierwszym samogłoski znajdowały się głównie w otoczeniu spółgłosek dźwięcznych, w drugim zaś w otoczeniu bezdźwięcznych (załącznik 12). Zależnie od struktury początku wypowiedzi analizowano zmiany intonacyjne o różnej dynamice na pierwszej sy-

labie akcentowanej. W załączniku 14 przedstawiono schematyczne, uśrednione przebiegi parametru F0 na pierwszych sylabach zdań jednofrazowych.

Opracowane, wyłącznie na drodze eksperymentalnej liczne modyfikacje algorytmu sterowania częstotliwością podstawową, doprowadziły w rezultacie do zmiany przyjętego schematu (według Fujisaki 1981) i kształtowania konturu intonacyjnego według modelu zaproponowanego w niniejszej pracy (por. rozdział 6).

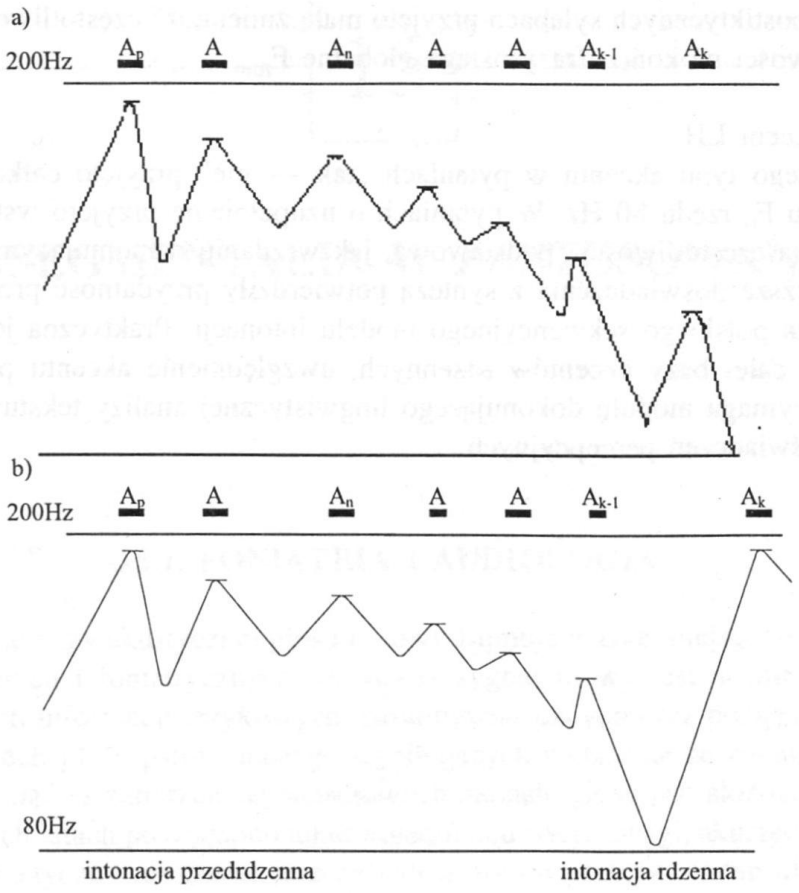
Na ryc. 14.2a zilustrowano przykładowy schemat zmian parametru F0 w jednofrazowym zdaniu oznajmującym, a na ryc. 14.2b w zdaniu pytającym o rozstrzygnięcie.

Ustalono następujące, podstawowe reguły sterowania częstotliwością podstawową.

1. Pierwszy akcent preiktyczny.

Na sylabach nieakcentowanych poprzedzających sylabę akcentowaną przyjęto małą prędkość zmian parametru F0: 5-6 półtonów/s. Prędkość zmian na akcentowanej sylabie zmienia się w zakresie 7-51 półtonów/s. Na sylabie akcentowanej poprzedzonej bezdźwięczną spółgłoską (spółgłoskami) tempo zmian częstotliwości wynosi około 7 półtonów/s, wartość początkowa częstotliwości jest wysoka np. 150 - 160 Hz. Na samogłosce występuje późne maksimum przebiegu. Jeżeli akcentowana sylaba nie była poprzedzona sylabami nieakcentowanymi, przyjęto prędkość zmian częstotliwości podstawowej około 50 półtonów/s oraz bardzo późne maksimum na samogłosce akcentowanej.

W sylabie akcentowanej poprzedzonej pojedynczą sylabą nieakcentowaną założono prędkość zmiany około 40 półtonów/s. Jeżeli sylaba akcentowana jest poprzedzona więcej niż jedną nieakcentowaną sylabą, przyjęto prędkość zmian parametru F0 w zakresie 25-40 półtonów/s. Towarzyszący spadek (12-45 półtonów/s) na następujących sylabach nieakcentowanych zależy od ich struktury oraz wartości maksimów występujących na samogłoskach akcentowanych. Wartość po-



Ryc. 14.2. Modelowanie intonacji a) w zdaniu oznajmującym b) w zdaniu pytającym o rozstrzygnięcie

cząstkowa parametru F0 zależy od pozycji frazy w zdaniu, długości frazy oraz struktury początku frazy (obecności/braku bezdźwięcznych spółgłosek i akcentowanych/nieakcentowanych sylab).

2. Pozostałe akcenty preiktyczne

Kolejne wartości maksimów na sylabach akcentowanych maleją wzdłuż frazy, spadki parametru F0 są spadkami całkowitymi lub częściowymi zależnie od struktury sylab nieakcentowanych pojawiających się pomiędzy sylabami akcentowanymi oraz od iloczasu sylaby akcentowanej. Granice między akcentami nie są zdefiniowane dokładnie i zależą od struktury akcentów. Maksima w przebiegu parametru F0 pojawiają się na początku, w środku lub na końcu samogłosek akcentowanych, minima na sylabach nieakcentowanych.

3. Akcent rdzenny

Zależnie od typu akcentu rdzennego przyjęto dla akcentu typu HL spadek 40 półtonów/s (50 - 80 Hz) a dla ML 25 - 30 półtonów/s (30 - 40 Hz). Na końcowych postiktycznych sylabach przyjęto małą zmienność częstotliwości. Wartość częstotliwości na końcu frazy osiąga globalne Fmin.

4. Akcent LH

Dla tego typu akcentu w pytaniach „tak — nie” przyjęto całkowity wzrost parametru F0 rzędu 80 Hz. W pytaniach o uzupełnienie przyjęto wstępnie model sterowania częstotliwością podstawową, jak w zdaniu oznajmującym.

Powyższe doświadczenia z syntezą potwierdziły przydatność proponowanego dla języka polskiego sekwencyjnego modelu intonacji. Praktyczna jednak implementacja całej bazy akcentów rdzennych, uwzględnienie akcentu preiktycznego typu L wymaga modułu dokonującego lingwistycznej analizy tekstu i szczegółowych doświadczeń percepcyjnych.

15 SUPRASEGMENTALIA W ZASTOSOWANIACH

15.1. FONIATRIA I AUDIOLOGIA

Metody analizy akustycznej głosu i mowy zajmują ważne miejsce w badaniach audiologicznych i foniatrycznych. Ponieważ sygnał mowy jest nośnikiem wielu różnorodnych informacji językowych, paralingwistycznych oraz pozajęzykowych, ekstrakcja cech głosu patologicznego wynikających wyłącznie ze zmian chorobowych krtani bądź z zaburzeń czynnościowych narządu głosu jest złożona. Pomimo że w ostatnich latach poświęcono temu zagadnieniu wiele uwagi, akustyczny „skrynning" profilaktyczny nie został jeszcze wdrożony do praktyki badań klinicznych. Intensywnie wzrasta w ostatnich latach wykorzystanie analiz (zwłaszcza cech suprasegmentalnych) mowy w audiologii i foniatrii zarówno dla potrzeb diagnozy, jak i rehabilitacji.

W Polsce problematyka ta stanowi przedmiot nielicznych opracowań poświęconych akustycznej analizie wybranych zaburzeń głosu/mowy (np. Gubrynowicz et al. 1980, 1981, Obrębowski 1982, Demenko et al. 1989, Pruszewicz et al. 1991, 1993, Pruszewicz 1992, Tarnowska et al. 1997, Świdziński 1999, Pruszewicz et al. 1999). Powstają nowatorskie metody automatycznej analizy akustycznej niektórych zaburzeń głosu przy wykorzystaniu sieci neuronowych (Tadeusiewicz et al. 1998, Izworski i Wszołek 1999).

Z licznych doniesień oraz badań własnych poświęconych analizie zaburzeń mowy o różnych etiologiach (np. Demenko et al. 1989, Pruszewicz 1992, Pruszewicz et al. 1991, 1993) wynika, że parametry suprasegmentalne (głównie zmienność „jittera" — odnosząca się do częstotliwości podstawowej, „shimmera" — odnosząca się odpowiednio do poziomu sygnału, iloczasu, przebiegu obwiedni intensywności w obrębie samogłosek i spółgłosek) istotnie ułatwiają obiektywną ocenę głosu/mowy patologicznej[8]8.

Zaburzenia mowy objawiają się często gwałtownymi zmianami częstotliwości podstawowej (nawet w obrębie samogłosek), szumem występującym w różnych nietypowych dla danego segmentu fonetycznego zakresach, niestabilnością głosu, niewłaściwym rytmem i nietypowymi zmianami poziomu sygnału. Gwałtowne zmiany tonu z okresu na okres (dochodzące nawet do kilkudziesięciu Hz) mogą być

[8] **Pod pojęciem jittera rozumie się zjawisko nieregularności zmian długości kolejnych okresów. Parametr ten jest określony liczbowo w procentach według następującego wzoru:**

$$PPQ = \frac{\frac{1}{n-2}\sum_{i=1}^{n-2}[(t(i)+t(i+1)+t(i+2))/3 - t(i+1)]}{\frac{1}{n}\sum_{i=1}^{n} t(i)} \times 100\%$$

gdzie: t(i) — długości kolejnych okresów,
n — numer kolejnego okresu.

symptomem zwiększonej masy fałdu głosowego. Niestabilność głosu oraz „jitter" może być objawem nieregularnej wibracji fałdów głosowych. W ostatnich latach analizę akustyczną mowy wykorzystywano w ocenie następujących stanów patologicznych narządu mowy i głosu.

1. Różnicowanie zaburzeń czynnościowych i organicznych narządu głosu

Stały wzrost zachorowań na nowotwory krtani zmusza do wielostronnej, ukierunkowanej profilaktyki, której najważniejszym zadaniem jest wczesne wykrywanie zmian organicznych w obrębie narządu głosu.

Różnicowanie patologii będącej następstwem zmian przede wszystkim w masie albo napięciu fałdów głosowych, jest zasadniczym kryterium oddzielenia zmian organicznych związanych najczęściej z przyrostem masy drgającego fałdu od zmian czynnościowych uwarunkowanych głównie zmianami napięcia i koordynacją drgań fałdów głosowych. O ile zmiany organiczne w obrębie głośni powodują zazwyczaj asymetryczny wzrost masy fałdu głosowego to zmiany czynnościowe wpływają na stopień jego napięcia. Czynnościowe zaburzenia głosu prowadzą do dyskoordynacji przede wszystkim fonacji i oddychania. Wśród przyczyn dysfonii organicznych szczególnie ważne miejsce zajmują wczesne nacieki nowotworowe w obrębie głośni.

W pracy Pruszewicz et al. (1991) podjęto próbę utworzenia podstaw akustycznego opisu dla różnicowania zaburzeń funkcjonalnych i organicznych. Na podstawie analizy (głównie suprasegmentaliów) w głosach 30 pacjentów (15 z zaburzeniami organicznymi i 15 z zaburzeniami czynnościowymi) postawiono i przetestowano hipotezę, że asymetria mas fałdów głosowych może sygnalizować zaburzenia organiczne, natomiast nieregularne sterowanie napięciem fałdów może sygnalizować zmiany funkcjonalne. Określenie na podstawie próbki głosu, czy zaburzenie jest wynikiem przyrostu masy czy dysfonii funkcjonalnej, byłoby niezwykle przydatne w diagnostyce wczesnych stanów nowotworowych.

Przeprowadzono szczegółowe badania foniatryczne i akustyczne mowy pacjentów z obu grup. Analizowano głównie parametry statystyczne rozkładów częstotliwości podstawowej oraz wartości maksymalne i średnie jittera.

Fragment wyników analiz akustycznych przedstawiono w tabeli 15.1. Jako zmienność intonacyjną w normie określono zakres zmian parametru F0, który wynosił co najmniej pół oktawy.

Tabela 15.1 Parametry akustyczne różnicujące zaburzenia mowy czynnościowe i organiczne

Cechy	**Grupa I zaburzenia organiczne**	**Grupa II** **zaburzenia czynnościowe**
Średnia wartość jittera	**8%**	**4%**
Maksymalna wartość jittera	**24%**	**28%**
Zakres harmonicznej struktury widma	**do 1600 Hz**	**do 2600 Hz**
Intonacja	**ograniczony zakres**	**w normie**

W obu grupach zaobserwowano podwyższoną wartość jittera. Zjawisko pewnej nieregularności zmian długości kolejno następujących po sobie okresów zaobserwowano również w mowie niepatologicznej. Szczegółowe badania poświęcone temu zagadnieniu przeprowadzał np. Horri (za Hess 1983). Przeciętne wartości jittera zależą od wartości częstotliwości: 51 µs dla F0 = 98 Hz i 24 µs dla F0 = 298 Hz.

Średnie wartości tego parametru dla mowy niepatologicznej mieszczą się w zakresie 0,5 - 1%. Badanie nieregularności zmian częstotliwości podstawowej należy przeprowadzać łącznie z analizą widmową sygnału.

O ile na początku lub końcu fonacji zaburzenia zmienności wysokości tonu są typowe (występuje nieregularność 2 lub 3 okresów), to często kilkuprocentowy jitter w środku samogłoski jest zjawiskiem wskazującym na zaburzenie mowy.

Nietypową zmienność jittera wytłumaczono brakiem koordynacji napięcia fałdów głosowych. W grupie I jest on dwukrotnie wyższy niż w drugiej, przy ogólnie niewielkim zakresie zmienności częstotliwości podstawowej. W obu grupach wystąpiła podobna maksymalna wartość jittera.

U pacjentów z zaburzeniami czynnościowymi zauważono trudności z utrzymaniem stałego tonu na samogłosce, pewne fragmenty samogłoski wymawiano wysoko, inne nisko. Zjawisko to świadczy o braku koordynacji w napięciu fałdów głosowych.

Przyrost masy fałdów głosowych jest zawsze zjawiskiem patologicznym i wymaga kontroli laryngologicznej. Jeżeli taką próbkę nagranego głosu udałoby się zanalizować i określić czy chrypka lub zaburzenie głosu jest wynikiem przyrostu masy, a nie dysfonii czynnościowej hiper czy hypofunkcjonalnej, to wówczas automatyczna analiza głosu stanowiłaby niezwykle przydatne narzędzie do wychwytywania w badaniach masowych zaburzeń organicznych, a zwłaszcza wczesnych stanów nowotworowych.

2. Ocena zaburzeń funkcjonowania narządu głosu spowodowanych niedosłuchem

Głos oraz mowa osób z niedosłuchem lub głuchych stanowi przedmiot badań zarówno audiologicznych, foniatrycznych, jak i akustycznych na całym świecie, również w Polsce[9]. Prowadzone są badania zmierzające do wykorzystania metod akustycznych w diagnozie i rehabilitacji niektórych typów niedosłuchów (np. Demenko et al. 1989, Pruszewicz et al. 1993). W pracach tych szczegółowej analizie poddano głos oraz mowę 18 pacjentów z głębokim niedosłuchem (powyżej 70 dB) lub całkowitą głuchotą. Pacjentów podzielono na 3 grupy: (1) osoby, u których uszkodzenie słuchu stwierdzono już w chwili urodzenia, (2) głuchota nabyta we wczesnym dzieciństwie (w okresie prelingwalnym czyli do 3 roku życia) i (3) głuchota nabyta po całkowitym zakończeniu przyswajania systemu językowego (powyżej 15 roku życia). Przeprowadzono szczegółowe badania audiometryczne oraz foniatryczne. Opracowany do analiz akustycznych materiał językowy składał się z izolowanych samogłosek, wyrazów i fraz oraz fragmentów mowy ciągłej. Mowa wszystkich pacjentów była zrozumiała (wyjątek stanowiły 3 osoby).

Jako cel badania przyjęto ustalenie związku między wiekiem pacjenta, w którym utracił on słuch i rozmiarem zaburzeń mowy. Założono, że również w mowie osób, które utraciły słuch w okresie postlingwalnym występują zniekształcenia, ale głównie cech suprasegmentalnych. Analizowano takie parametry, jak: średnia wartość i zakres zmian parametru F0, jitter, niestabilność głosu (określona maksymalną zmianą z okresu na okres) periodyczność sygnału na samogłoskach oraz spółgłoskach artykułowanych normatywnie z fonacją. Ponieważ najczęściej wypowiedzi osób z niedosłuchami charakteryzują się intonacją opadającą, analizowano zmiany wysokości tonu w wypowiedziach pytających. Fragment uzyskanych w analizach akustycznych wyników ilustruje tabela 15.2, w której umieszczono po dwa przypadki przykładowo z każdej grupy.

W kolejnych kolumnach oznaczono: płeć oraz wiek pacjenta, średnią wartość częstotliwości podstawowej, niestabilność głosu określoną maksymalną zmianą długości sąsiadujących ze sobą okresów tonu podstawowego, zakres zmian częstotliwości podstawowej, obecność szumu na samogłoskach i spółgłoskach, jitter oraz typ przebiegu intonacyjnego w pytaniu.

Największe zaburzenia mowy (wartości wszystkich badanych parametrów odbiegające od normy) zauważono w grupie I, a więc u osób, które urodziły się z uszkodzeniem narządu słuchu. W II grupie zauważono również znaczne zaburzenia mowy, zarówno w cechach segmentalnych (np. szum na samogłoskach),

[9] W nini ejszej pracy omawia się prace wykonane wyłącznie z udziałem autorki w Klinice Foniatrii i Audiologii Akademii Medycznej w Poznaniu, kierowanej przez prof. zw. dr hab. dr h.c. A. Pruszewicza.

jak i suprasegmentalnych (np. zmniejszony zakres zmian częstotliwości podstawowej).

Tabela 15.2 Parametry akustyczne mowy pacjentów z niedosłuchem lub całkowitą głuchotą

Grupa	Inicjały Płeć wiek	[Hz]	Max [Fi+1/Fi]	Zakres Fmin - Fmax [Hz]	Obecność szumu spółgłoski/ samogłoski	Jitter PPQ [%]	Typ intonacji
I	AW F-59	282	400 276	93-350	+/+	12	opadająca
	DB F-34	240	300 218	120-308	+/+	8	opadająca
II	SK F-32	225	237 210	186-265	+/-	4	opadająca
	HB F-25	321	382 318	289-440	+/-	6	równa
III	SB M-35	165	162 147	126-315	+/-	2	równa
	KF M-11	245	286 256	219-285	-/-	1,5	opadająca

Interesująca jest analiza zaburzeń mowy pacjentów z grupy III, czyli osób, które utraciły słuch po 15 roku życia. W cechach segmentalnych mowy tych pacjentów nie zauważono istotniejszych zmian. Znaczne zaburzenia zaobserwowano w cechach suprasegmentalnych, zmienność częstotliwości podstawowej była niewielka, występowały przeskoki w zmianach wysokości tonu, pytania realizowano z intonacją opadającą lub równą.

Podwyższenie wartości średniej częstotliwości podstawowej zinterpretowano jako brak skutecznego sprzężenia zwrotnego wynikającego z faktu, że mówca jest jednocześnie własnym słuchaczem. Hiperfunkcja krtani oraz jej asymetria wpływały na zwiększenie wartości jittera i shimmera.

3. Laryngektomie całkowite i częściowe.

Analiza akustyczna mowy (w tym również cech suprasegmentalnych) jest szczególnie przydatna u pacjentów po częściowych lub całkowitych laryngektomiach (Tarnowska et al. 1997). Mowa pacjentów po laryngektomiach posługujących się tzw. mową przełykową charakteryzuje się głównie bardzo niską częstotliwością podstawową, dłuższym czasem trwania wypowiedzi (co najmniej 1,5 razy dłuższym) oraz ubogą intonacją. Metody analizy akustycznej pozwalają ocenić obiektywnie postępy w rehabilitacji foniatrycznej pacjenta. Zagadnienie sterowania zmianami wysokości tonu wymaga rozwiązania w technicznych układach sztucznej krtani.

4. Wirylizacja głosu.

W ocenie klinicznej objawem wirylizacji narządu głosu u kobiet jest znaczne obniżenie zakresu głosu (odbiegające od średniej statystycznej wysokości głosu dla płci żeńskiej). Szczegółowe badania z tego zakresu przeprowadził między innymi Obrębowski (1982). Badał zjawisko wirylizacji po preparatach sterydowych i zjawisko mutacji perwersyjnej. W analizie akustycznej szczególnej uwadze poddano cechy suprasegmentalne mowy. Zaobserwowano znaczne obniżenie średniej wartości częstotliwości podstawowej głosu (w niektórych przypadkach nawet o oktawę), ubogość intonacji i charakterystyczne wielomodalne rozkłady częstotliwości podstawowej ilustrujące brak stabilnej średniej częstotliwości podstawowej. Zaraz po rozpoczęciu fonacji obserwowano gwałtowne obniżenie głosu.

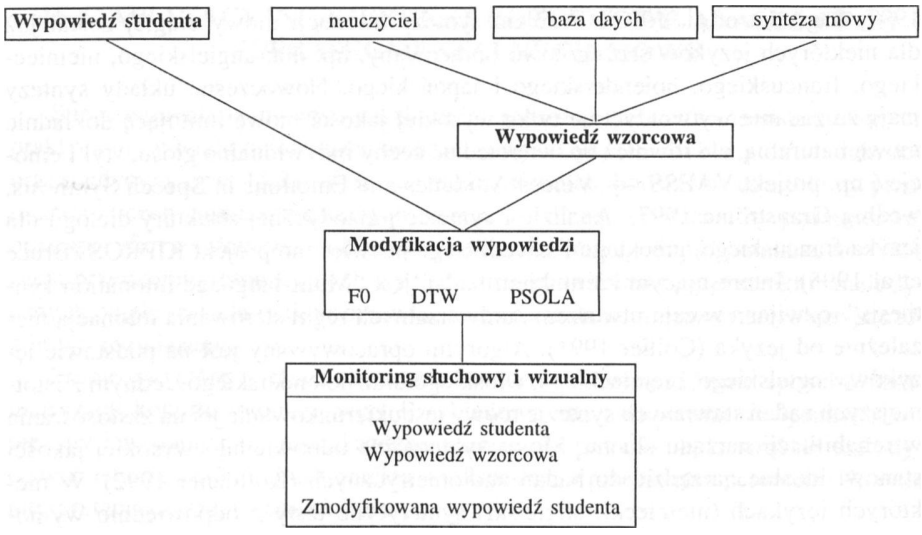
Metody akustyczne mogą być przydatne zarówno w diagnozie, jak i rehabilitacji zaburzeń narządu głosu i słuchu. Są nieinwazyjne, pozwalają na obiektywizację badań klinicznych w różnych zaburzeniach głosu i mowy, umożliwiają śledzenie postępów w rehabilitacji i mogą potencjalnie zostać wykorzystane w foniatrycznych oraz audiologicznych badaniach masowych.

15.2. JĘZYKOZNAWSTWO

Dla potrzeb lingwistyki podstawowym zagadnieniem jest ustalenie wzorców melodycznych oraz ich związków ze strukturą danego języka. Transkrypcja struktur melodycznych ciągle jest przedmiotem dyskusji specjalistów z różnych dziedzin zarówno lingwistyki, akustyki, jak i techniki. Aktualne systemy transkrypcji między innymi ToBI (np. Silverman et al. 1992, Grice et al. 1996), I.P.A (I.P.A Report 1989) oraz IviE (Grabe 1998) nie stanowią uniwersalnego narzędzia możliwego do wykorzystania w praktycznych aplikacjach. Szereg opracowań dotyczy lingwistycznych aspektów analizy struktury przebiegu melodii. Z tej dziedziny dla języka polskiego przeprowadzała głównie prace Steffen-Batogowa (np. 1963, 1996) oraz Jassem (np. 1962, 1989). Interesującym nowym kierunkiem prac z tego zakresu jest analiza porównawcza intonacji różnych języków (np. Collier 1990).

Popularne w ostatnich latach stały się komputerowe systemy nauki intonacji (np. system Visi Pitch). Poprzez jednoczesne wyświetlanie w czasie rzeczywistym wzorcowych intonacji w wypowiedziach nauczyciela oraz imitacji wzorców w głosie studenta przeprowadza się korektę. Najlepsze rezultaty osiąga się poprzez jednoczesną prezentację akustyczną i graficzną sygnału (Hess 1983, Allen et al. 1991). Tego rodzaju opracowania są bardzo złożone, wymagają wizualizacji i normalizacji nie tylko częstotliwości podstawowej i poziomu sygnału, ale również segmentacji sygnału.

Poniższa rycina (ryc. 15.1) ilustruje przykładowy system do nauki intonacji (za Granström 1997).



Ryc. 15.1. System nauki intonacji

Resyntezę wypowiedzi z zadanym wzorcem intonacyjnym umożliwia technika PSOLA. Synteza PSOLA wykorzystuje metodę konkatenacyjną (pitch synchronous overlap-and-add), sumuje, łączy krótkie fragmenty (zwykle difony). Synteza ta cechuje się dość dobrą jakością, ale posiada istotne wady: wymaga manualne-

go wyznaczania początku każdego okresu, wygładzania na łączonych segmentach oraz dużej pamięci. Synteza metodą PSOLA jest przeprowadzana dla określonego głosu. Po normalizacji w zakresie wysokości tonu (F0) i tempa wypowiedzi (DTW — Dynamic Time Warping, por. rozdz. 11) porównywany jest wzorzec intonacyjny i jego imitacja. System ten może służyć również do rehabilitacji słuchowej osób głuchych. Poprzez bisensoryczną stymulację (jednoczesne współdziałanie analizatora wzrokowego i słuchowego) odbierane bodźce wzajemnie się wzmacniają i uzupełniają. Tego rodzaju synergizm wzrokowo-słuchowy daje istotne efekty w rehabilitacji osób z uszkodzeniami słuchu i ułatwia porozumiewanie się z pacjentem. Powyższy przykład ilustruje ścisłe związki między lingwistycznymi a rehabilitacyjnymi aspektami zastosowań nauki o cechach suprasegmentalnych mowy.

15.3. TECHNIKA

Od końca lat 60. stale wzrasta zainteresowanie suprasegmentaliami w technologii mowy, głównie w syntezie i rozpoznawaniu. Praktycznie wdrożono modelowanie suprasegmentaliów do wszystkich aktualnych systemów technicznych wykorzystujących syntezę mowy (Willems et al. 1988, Teranishi 1989, Morlec et al. 1997, Sagisaka et al. 1997). Problem syntezy intonacji mowy ciągłej został już dla niektórych języków szczegółowo opracowany, np. dla: angielskiego, niemieckiego, francuskiego, holenderskiego i japońskiego. Nowoczesne układy syntezy mają za zadanie wytworzyć nie tylko wysokiej jakości mowę imitującą dokładnie mowę naturalną, ale również odzwierciedlać cechy indywidualne głosu, styl i emocje (np. projekt VAESS — Voices Attidudes and Emotions in Speech Synthesis, według Granströma 1997). Analizie i syntezie prozodycznej struktury dialogu dla języka francuskiego, greckiego i szwedzkiego poświęcono projekt KIPROS (Bruce et al 1995). Interesującym kierunkiem badań jest „Multi-language intonation synthesis" rozwijana w celu utworzenia uniwersalnych reguł sterowania intonacją niezależnie od języka (Collier 1991). Algorytm opracowywany jest na podstawie języków: angielskiego, niemieckiego, włoskiego oraz holenderskiego. Jednym z istotniejszych zadań stawianych syntezie mowy jest ukierunkowanie jej na zastosowania w rehabilitacji narządu słuchu. Mowa syntetyczna odpowiednio wysokiej jakości stanowi idealne narzędzie do badań audiometrycznych (Kollmeier 1992). W niektórych językach (niemiecki, angielski) syntetyczne testy z odpowiednio wymodelowanymi cechami segmentalnymi oraz suprasegmentalnymi wchodzą w skład standardowych zestawów testów audiometrycznych (np. MAC — Minimal Auditory Capability czy TAPS — testy przeznaczone do wykorzystania u pacjentów z wszczepami ślimakowymi, dla języka polskiego opracowano wstępne wersje programów TAPS oraz MAC, por. Pruszewicz 1994, Demenko et al. 1994, 1996).

Dla modelowania intonacji języka polskiego zaimplementowano jedynie podstawowe reguły, sterujące cechami segmentalnymi oraz suprasegmentalnymi, które pozwoliły uzyskać mowę zrozumiałą (Imiołczyk et al. 1994).

Złożoność zagadnienia wykorzystania suprasegmentaliów w rozpoznawaniu mowy przedstawiono w podstawowym zarysie w obecnej pracy. W dziedzinie tej dominują rozwiązania teoretyczne. Pierwszą i jak na razie jedyną udaną implementację analizy suprasegmentaliów w rozpoznawaniu mowy prezentuje system Verbmobil (Batliner 1997, Batliner et al. 1996, Niemann et al. 1997, 1998). Rozpoznawanie granic frazowych (poprawne dla języka niemieckiego w 94%) oparto na technice klasycznych sieci neuronowych MLP. W stosunkowo wąskim zakresie wykorzystywane są suprasegmentalia w identyfikacji głosu. Przykładowo Atal (1972) na podstawie analizy 10 głosów uzyskał średnio 97% poprawną identyfikację. W nowszym opracowaniu Parris i Carrey (1996) zwracają uwagę na przy-

datność cech suprasegmentalnych mowy w rozpoznawaniu mówców. Dla języka polskiego podjęto tylko pilotażowe analizy cech osobniczych na podstawie statystycznych badań zmienności tonu (np. Steffen-Batóg et al. 1970, Demenko 1984).

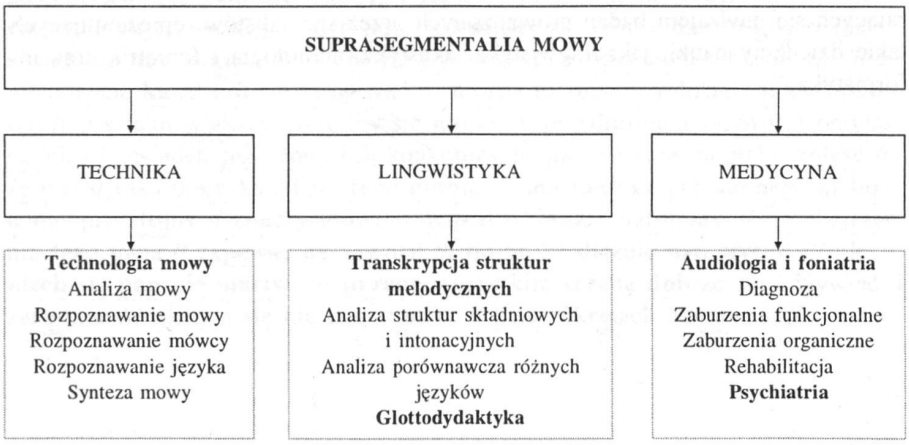
Na potrzeby telekomunikacji interesującą dziedziną jest identyfikacja języka. W tej dziedzinie istnieje obecnie niewiele opracowań. Dotychczasowe badania uwzględniają głównie języki: angielski, chiński, japoński oraz hiszpański (za Granström 1997).

15.4. APLIKACYJNE KIERUNKI ROZWOJOWE ANALIZ CECH SUPRASEGMENTALNYCH MOWY

Suprasegmentalia mowy są nośnikiem różnorodnych informacji przydatnych praktycznie w wielu dziedzinach techniki, lingwistyki (fonetyki i glottodydaktyki) oraz medycyny (ryc. 15.2). Wyróżnić można wspólne, podstawowe problemy przetwarzania cech suprasegmentalnych mowy, niezależne od dziedziny podstawowej oraz specyficzne, dotyczące określonej aplikacji. Wiarygodność pomiaru, normalizacja przebiegów intonacyjnych oraz automatyczna transkrypcja struktur melodycznych ciągle jest dla praktycznych zastosowań zadaniem niewystarczająco szczegółowo rozwiązanym.

Problemy, które narzuca konkretna aplikacja dotyczą głównie sposobu parametryzacji konturu intonacyjnego. Przykładowo, dla zastosowań medycznych konieczny jest pomiar długości każdego okresu, dla potrzeb nauki struktur melodycznych istotna jest natomiast analiza wygładzonych przebiegów parametru F0 oraz interpretacja zmian wysokości tonu.

W najnowszym modelu spektrografu Kay 4305 wyspecjalizowanym do analiz patologii mowy opracowano program MDVP (Multi Dimensional Voice Program). Jego zadaniem jest wszechstronna ocena mowy z zastosowaniem 33 parametrów akustycznych (w tym kilkunastu opisujących cechy suprasegmentalne wypowiedzi) oceniających zmianę częstotliwości, amplitudy oraz stosunek sygnału do szumu. Brak natomiast jest normy określającej cechy prawidłowej/nieprawidłowej emisji głosu. Jak wykazuje literatura przedmiotu, badania w tym zakresie prowadzone są dla różnych języków już od 30 lat i jak na razie postęp w tej dziedzinie jest



Ryc. 15.2. Zastosowania suprasegmentaliów mowy

stosunkowo powolny. Można przypuszczać, że rozwój narzędzi technicznych, takich jak program MDVP przyspieszy rozwój prac również w tej dziedzinie.

Dla systemów syntezy mowy z tekstu fundamentalne znaczenie ma analiza struktur składniowych i intonacyjnych języka. Na podstawie dowolnego tekstu ortograficznego system musi wygenerować sygnał mowy o akceptowalnych cechach prozodycznych. Automatyczna analiza gramatyczna jest konieczna w celu dokonania podziału wypowiedzi na frazy oraz w celu właściwego rozmieszczenia pauz. Nie wszystkie bowiem znaki interpunkcyjne generują granice międzyfrazowe i nie wszystkie granice międzyfrazowe są określone znakami interpunkcyjnymi. Na podstawie informacji zawartej w bazie słownikowej kolejne wyrazy tekstu poddawane są, zgodnie z analizą morfologiczną, dekompozycji (na: rdzenie, przyrostki, przedrostki, końcówki fleksyjne). Następnie określa się przynależność do kategorii części mowy oraz funkcje wyrazu w zdaniu. Jednym z najpopularniejszych formalizmów mających na celu reprezentację informacji składniowej oraz semantycznej na potrzeby automatycznej analizy mowy jest gramatyka HPSG (Head-driven Phrase Structure Grammar). Prozodyczna informacja określona symbolem PSCB (Prosodic Syntactic Clause Boundary) oraz informacja składniowa i semantyczna wykorzystywane są obecnie na świecie w rozwijanych systemach dialogowych (por. np. *Verbmobil* — Niemann et al. 1997). Dla języka polskiego, jak dotąd, brak szerszych opracowań w tym zakresie.

Zorganizowaną w 1995 roku w Kyoto Konferencję — *Obliczeniowe Aspekty w Przetwarzaniu Prozodii Mowy* (Sagisaka et al. 1997) poświęcono wszechstronnej ocenie aktualnego stanu wiedzy i dalszych kierunków rozwojowych analiz suprasegmentaliów w różnych dziedzinach nauki.

Niezależnie od zastosowań większość prac powinna być przeprowadzana na podstawie empirycznych faktów ujmujących akustyczne i biologiczne obserwacje. Zrozumienie złożonych związków w analizie suprasegmentaliów wymaga wspierających się nawzajem badań prowadzonych przez specjalistów reprezentujących takie dziedziny nauki, jak: lingwistyka, akustyka, audiologia i foniatria oraz informatyka.

ZAŁĄCZNIKI

ZAŁĄCZNIK 1

z telewizyjnego kina nocnego // wciąż z sympatią wspominam // człowieka który się zmniejszał ///film z gatunku fikcji naukowej ///a więc raczej fantastyczny // i raczej nie naukowy /// ale za to pobudzający wyobraźnię ///była to opowieść o panu ///który zaczął się zmniejszać /// oczywiście / poza tym panem // wszyscy to bagatelizowali /// żona powtarzała mu // że jest wcale duży jak na małego /// bo dużo zarabiał ///a lekarz kazał mu dobrze się odżywiać // nic się nie przejmować // i na zakrętach hamować /// bo każdy dobrze życzy małemu /// który dobrze zarabia /// życie pana który malał // stawało się więc coraz trudniejsze /// dostał wprawdzie specjalny domek trzypokojowy pod krzesłem /// i może nawet żyłby tam szczęśliwie /// gdyby nie jego własny kot /// który nie zmalał /// a więc w efekcie przerósł go o głowę /// i najpierw chciał go zjeść jako mysz /// a ostatecznie wrzucił go do piwnicy /// gdzie ten pan ze zgryzoty zaczął maleć tak dalece /// że prawie zjadł go drobny pajączek /// lecz mimo tych przygód żył nadal /// choć wciąż się zmniejszał /// dzięki czemu jest obecnie jednym z atomów /// który tym się różni od pozostałych /// że ma wyraz twarzy /// ręce // i nogi /// bardzo to był niegłupi film /// istotnie jest absolutnym przypadkiem /// że wzrost nasz wynosi powiedzmy sto siedemdziesiąt /// można żyć i mieć kłopoty nawet ze wzrostem sto dwadzieścia /// i można być szczęśliwym // zupełnie bez wzrostu /// jednocześnie jednak film ten nasuwał różne myśli i spostrzeżenia szersze /// że mianowicie los tego pana był jednym z lepszych /// jako atom // wiedzie on obecnie życie chyba spokojne /// o ileż paskudniejsze byłoby ono /// gdyby los // świadcząc mu psikusa /// skierował jego rozwój w kierunku odwrotnym /// a mianowicie kazał mu się zwiększać /// można to sobie wyobrazić tak /// pan ten // już jako większy /// stałby się najpierw przedmiotem zazdrości przyjaciółek / i sąsiadek jego żony lub konkubiny /// panie te nie mogłyby znieść // że ona // taka pokraka / i do tego głupia /// ma takiego przystojnego /// bo u nas przystojny // znaczy głównie drągal /// lekarz natomiast // po obejrzeniu tego pana // zapewne by zawołał /// ho-ho /// ślicznie pan wyrósł /// ale niech się pan nie martwi /// przede wszystkim trzeba dobrze się odżywiać i zara-biać /// niczym się nie przejmować // i na zakrętach hamować ///

prokuratura // skierowała do sądu wojewódzkiego w płocku /// akt oskarżenia przeciwko trzem pracownikom mostostalu z zabrza /// zarzucając im nieumyślne spowodowanie /// podczas prac remontowych ósmego sierpnia ubiegłego roku /// zawalenie się masztu radiowego w konstantynowie koło gąbina /// poinformowano piętnastego miesiąca // w prokuraturze wojewódzkiej w płocku /// korzystając z okazji /// jaką było spotkanie z grupą dziennikarzy włoskich /// pułkownik muamar kadafi oświadczył /// że nie wyklucza swojego udziału // w następnych wyborach prezydenckich we włoszech /// zdziwionym rozmówcom wyjaśnił ///że urodził się jako obywatel włoski /// i nikt mu oficjalnie tego obywatelstwa nie odebrał /// kadafi przedstawił też zarys swojego programu // / najważniejszym punktem // okazało się wypędzenie amerykanów /// których imperialistyczne zapędy ograniczają jego zdaniem wolność włoch /// przez jeden dzień / pasażerowie podróżujący moskiewskim metrem /// nie musieli uiszczać żadnej opłaty za przejazd ///stało się tak // za sprawą umowy /// jaką dyrekcja metra // podpisała z pewną amerykańską firmą /// która zgodziła się opłacić koszty jednego dnia eksploatacji / w zamian za ciągłe nadawanie programu reklamowego /// wysokość transakcji nie została ujawniona /// jednak amerykańscy przedsiębiorcy /// po otrzymaniu wyników badania skuteczności oddziaływania tej formy reklamy /// zdecydowali /// że do końca roku opłacą jeszcze dwa dni darmowej jazdy /// w dużych miastach na wschodnim wybrzeżu stanów zjednoczonych / pojawiła się ostatnio znaczna liczba sokołów /// ptaki te // ginęły masowo w latach pięćdziesiątych i sześćdziesiątych / w wyniku stosowania ddt // jako środka ochrony roślin w rolnictwie /// w ciągu ostatniego dziesięciolecia /// pewna ich ilość / dzięki ornitologom / zagnieździła się w nowym jorku /// ale nikt nie sądził /// że nowe środowisko // okaże się dla sokołów aż tak atrakcyjne /// ptaki polują na gryzonie / i trzebią nadmiernie rozmnożone gołębie /// ekolodzy obawiają się jednak /// że z braku naturalnych wrogów / mogą z czasem same stać się plagą /// londyn ma zostać jedyną stolicą europejską /nie posiadającą ogrodu zoologicznego ///mimo ciągnących się od dwóch miesięcy dyskusji /// nadal nie ma źródła /// z którego można by finansować działalność tej liczącej sto pięćdziesiąt lat instytucji ///władze miasta nie mają brakującej sumy /// około dwudziestu jeden milionów dolarów /// a rząd odmawia subwencji /// w ogrodzie przebywa osiem tysięcy zwierząt / z których większość / w razie likwidacji zoo /// zostanie uśpiona /// gdyż w związku z zastojem na międzynarodowym rynku zoologicznym /nie ma na nie nabywców ///

Oznaczenia:

/ — granica frazowa słaba
// — granica frazowa średnia
/// — granica frazowa silna

═══ — akcent silny
─── — akcent średni
..... — akcent słaby

ZAŁĄCZNIK 2

W poszczególnych 16 wierszach tabeli umieszczono kolejno wyrazy z następujących 16 wypowiedzi:

1. Wiesz, czym go obrzucili? **Jajem.**
2. Mnie się nie spieszy. **Ja jem.**
3. Czym go obrzucili — pomidorem czy **jajem?**
4. Podnieś te worki i wrzuć je **na wóz.**
5. W tamtych workach jest piasek, a w tych — **nawóz.**
6. Co jest w tych workach — piasek czy **nawóz?**
7. Czy podnieść te worki i wrzucić je **na wóz?**
8. Tam są najlepsi lekarze i to jest najlepsza **poradnia.**
9. Czy tam są najlepsi lekarze i czy to jest najlepsza **poradnia?**
10. Było już dość jasno, więc to była najlepsza **pora dnia.**
11. Czy to była najlepsza **pora dnia?**
12. Czy klocki rozrzucaliście, czy **zbieraliście?**
13. Czy Tomek rozrzuca, czy **zbiera liście?**
14. Czy to wszystko wczoraj **zbieraliście,** czy dzisiaj?
15. To wszystko wczoraj **zbieraliście,** a nie dzisiaj.

16. Nie rozrzucaliście klocków, tylko je **zbieraliście.**

ZAŁĄCZNIK 3

Tabela 1.

Wyniki przeprowadzonego rozpoznania 16 wyrazów kluczowych. W kolumnach umieszczono rezultaty odsłuchu dla wyrazu wypowiedzianego przez każdą z 15 osób. W kolejnych wierszach tabeli umieszczono odpowiednio: liczbę rozpoznanych znaczeń wyrazu, wyniki testu χ2 oraz poziom istotności. Przykładowo, wyraz *jajem* wypowiedziany przez głos MM, 18 osób rozpoznało prawidłowo, 2 osoby błędnie. Wartość statystyki dla tego przykładu χ2 = 12,8, a więc różnice między oceną znaczeń nie są istotne

L.p.	Wyraz	MM	WT	AN	MR	Ol	KM	MK	AK	MZ	PW	GD	JO	KK	RD	JK
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	jajem	20	20	20	20	20	20	19	18	18	20	20	19	19	18	20
	ja jem							1	2	2			1	1	2	
	X2	20,0	20,0	20,0	20,0	20,0	20,0	16,2	12,8	12,8	20,0	20,0	16,2	16,2	12,8	20,00
	a	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0003	0,0003	0,0000	0,0000	0,0000	0,0000	0,0003	0,0000
2	jajem	2													1	1
	ja jem	18	20	20	20	20	20	20	20	20	20	20	20	20	19	19
	X2	12,8	20,00	20,00	20,00	20,00	20,00	20,00	20,00	20,00	20,00	20,00	20,00	20,00	16,2	16,2
	a	0,0003	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
3	ja jem	1	2		2	1	2		2	2	2		1		2	1
	jajem	19	18	20	18	19	18	20	18	18	18	20	19	20	18	19
	X2	16,2	12,8	20,00	12,8	16,2	12,8	20,00	12,8	12,8	12,8	20,00	16,2	20,00	12,8	16,2
	α	0,0000	0,0003	0,0000	0,0003	0,0000	0,0003	0,0000	0,0003	0,0003	0,0003	0,0000	0,0000	0,0000	0,0003	0,0000
4	nawóz	7	6		5	6	3	2	2	4	4	6	2	1	2	1
	na wóz	13	14	20	15	14	17	18	18	16	16	14	18	19	18	19
	X2	1,8	3,2	20	5,0	3,2	9,8	12,8	12,8	7,2	7,2	3,2	12,8	16,2	12,8	16,2
	a	0,1797	0,0736	0,0000	0,0035	0,0736	0,0017	0,0003	0,0003	0,0072	0,0072	0,0072	0,0003	0,0000	0,0003	0,0000

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
5	nawóz	20	19	20	20	20	20	20	20	20	20	20	20	20	19	20
	na wóz		1												1	
	x2	20,00	16,2	20,00	20,00	20,00	20,00	20,00	20,00	20,00	20,00	20,00	20,00	20,00	16,2	20,00
	a	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
6	na wóz	1	1	1	1						1					1
	nawóz	19	19	19	19	20	20	20	20	20	19	20	20	20	20	19
	χ2	16,2	16,2	16,2	16,2	20,00	20,00	20,00	20,00	20,00	16,2	20,00	20,00	20,00	20,00	16,2
	a	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
7	na wóz	4	12	4	20	5	20	20	19	20	7	4	20	20	7	14
	nawóz	16	8	16		15			1		13	16			13	6
	X2	7,2	0,8	7.2	20,00	5,0	20,00	20,00	16,2	20,00	1,8	7,2	20,00	20,00	1,8	3,2
		0,0072	0,3711	0,0072	0,0000	0,0035	0,0000	0,0000	0,0000	0,0000	0,1797	0,0072	0,0000	0,0000	0,1797	0,0072
8	poradnia	20	20	20	20	19	20	16	20	19	20	19	20	16	20	20
	pora dnia					1		4		1		1		4		
	X2	20	20	20	20	16,2	20	7,2	20	16,2	20	16,2	20	7,2	20	20
	α	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0736	0,0000	0,0000	0,0000	0,0000	0,0000	0,0736	0,0000	0,0000
9	pora dnia	8	8	14	5	11	8	11	5	10	5	10	9	11	6	8
	poradnia	12	12	6	15	9	12	9	15	10	15	10	11	9	14	12
	χ2	0,8	0,8	3,2	5,0	0,2	0,8	0,2	5,0	0,0	5,0	0,0	0,2	04	3,2	0,8
	α	0,3711	0,3711	0,0736	0,0253	0,6547	0,3711	0,6547	0,0253	1,0000	0,0253	1,0000	0,6547	0,6547	0,0736	0,3711
10	pora dnia	20	20	20	20	20	20	20	18	20	20	20	20	19	18	20
	poradnia								2					1	2	
	X2	20	20	20	20	20	20	20	12,8	20	20	20	20	16,2	12,8	20
	α	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
11	poradnia	3	1	1			2	1	2	1			2	2	1	4
	pora dnia	17	19	19	20	20	18	19	18	19	20	20	18	18	19	16
	Z2	9,8	16,2	16,2	20	20	12,8	16,2	12,8	16,2	20	20	12,8	12,8	16,2	7,2
	α	0,0017	0,0000	0,0000	0,0000	0,0000	0,0003	0,0000	0,0003	0,0000	0,0000	0,0000	0,0003	0,0003	0,0000	0,0072
12	zbieraliście	16	9	7	18	18	11	19	18	16	18	18	11	18	19	5
	zbiera liście	4	11	13	2	2	9	1	2	4	2	2	9	2	1	15
	z2	7,2	0,2	1,8	12,8	12,8	0,2	16,2	12,8	7,2	12,8	12,8	0,2	12,8	16,2	5,0
	α	0,0072	0,6547	0,1797	0,0003	0,0003	0,6547	0,0000	0,0003	0,0073	0,0003	0,0003	0,6547	0,0003	0,0000	0,025
13	zbieraliście	8	10	1	7	4	5	3	8	6	7	4	3	4	8	6
	zbiera liście	12	10	19	13	16	15	17	12	14	13	16	17	16	12	14
	z2	0,8	0,0	16,2	1,8	7,2	5,0	9,8	0,8	3,2	1,8	7,2	9,8	7,2	0,8	3,2
	a	0,3711	1,0000	0,0000	0,1797	0,0073	0,0253	0,0017	0,3711	0,0736	0,1797	0,0073	0,0017	0,0073	0,3711	0,0736
14	zbieraliście	10	14	12	11	13	13	12	10	8	11	14	12	12	10	6
	zbiera liście	10	6	8	9	7	7	8	10	12	9	6	8	8	10	14
	z2	0,0	3,2	0,8	0,2	1,8	1,8	0,8	0,0	0,8	0,2	3,2	0,8	0,8	0,0	3,2
	α	1,0000	0,0736	0,3711	0,6547	0,1797	0,1797	0,3711	1,0000	0,3711	0,6547	0,0736	0,3711	0,3711	1,0000	0,0736
15	zbieraliście	9	9	8	11	14	10	11	13	13	11	13	10	11	13	16
	zbiera liście	11	11	12	9	6	10	9	7	7	9	7	10	9	7	4
	Z2	03	0,2	0,8	0,2	3,2	0,0	0,2	1,8	1,8	0,2	1,8	0,0	0,2	1,8	7,2
	a	0,6547	0,6547	0,3711	0,6547	0,0736	1,0000	0,6547	0,1797	0,1797	0,6547	0,1797	1,0000	0,6547	0,1797	0,0073
16	zbiera liście	3	2	9	2	3	16	6	1		2	4	16	5	1	11
	zbieraliście	17	18	11	18	17	4	14	19	20	18	16	4	15	19	9
	χ2	9,8	12,8	0,2	12,8	9,8	7,2	3,2	16,2	20	12,8	7,2	7,2	5,0	16,2	0,2
	a	0,0017	0,0003	0,6547	0,0000	0,0017	0,0072	0,0072	0,0000	0,0000	0,0000	0,0072	0,0072	0,0253	0,0000	0,6547

ZAŁĄCZNIK 4

Tabela 2 Udział cech akustycznych w klasyfikacji akcentu

	Zestawienie analizy funkcji dyskryminacyjnej. Grupująca 5 zmiennych w modelu Lambda Wilksa: 0,33831 przybl. F (5,174) = 68,065 p< 0,0001			
Cechy	Lambda Wilksa	Cząstk. Wilksa	F usun. (1,174)	poziom p
ΔDvi	0,401683	0,842223	32,59540	0,000001
ΔI	0,338315	0,999978	0,00380	0,950944
ΔFi	0,418120	0,809117	41,04926	0,000000
AFr	0,363026	0,931911	12,71315	0,000469
ΔFmax	0,346420	0,976585	4,17215	0,042602

ZAŁĄCZNIK 5

Frazy wzorcowe z oznaczonym typem akcentu rdzennego

Lp.	Fraza	Typ akcentu
1	2	3
1	znów	ML
2	znowu	ML
3	znowu on	ML
4	znowu ona	ML
5	znowu ten wariat	ML
6	znów	LM
7	znowu	LM
8	znowu on	LM
9	znowu ona	LM
10	znowu ten wariat	LM
11	znak	ML
12	ten znak	ML
13	jakiś znak	ML
14	tu jest jakiś znak	ML
15	znak	xL
16	zły znak	xL
17	bardzo zły znak	xL
18	bardzo zły znak	xL
19	znak	HL
20	jakiś znak	HL
21	tu jest jakiś znak	HL
22	bardzo niedobry znak	HL
23	to jest bardzo niedobry znak	HL
24	bardzo zły człowiek	HL
25	bardzo zły człowiek	ML
26	co	LH
27	proszę	LH
28	co mówiłeś	LH
29	co ona mówiła	LH
30	to był całkiem niezły i uczciwy człowiek	ML
31	to był całkiem niezły i uczciwy człowiek	HL

1	2	3
32	róże	HM
33	goździki	HM
34	tulipany	HM
35	tak	MM
36	owszem	MM
37	może	MM
38	możliwe	MM
39	nie marudź	MM
40	znów	MH
41	znowu	MH
42	znowu on	MH
43	znowu ona	MH
44	znowu ten wariat	MH
45	o	LHL
46	ale	LHL
47	jeszcze	LHL
48	wspaniale	LHL
49	kolosalnie	LHL
50	dla miasta	ML
51	dwa miasta	ML
52	dzień dobry	HL
53	dzień dobry	Ml
54	dzień dobry	HM
55	dzień dobry	LH
56	dzień dobry	LM
57	dzień dobry	LH
58	dzień dobry	LHL
59	dzień dobry	xL
60	dzień dobry	MM

ZAŁĄCZNIK 6

Wyrazy kluczowe: ***Marek***, ***Czarek***, ***Darek***.

Struktura segmentalna wyrazów kluczowych:
Samogłoska akcentowana *a* przed *r*
nieakcentowana *e* przed *k*

Kontekst:

Wyraz następujący po kluczowym rozpoczyna się od spółgłoski bezdźwięcznej.

Struktura rytmiczna

Wyrazy kluczowe stanowią odrębny zestrój akcentowy.

Struktura prozodyczna:

1. Środek frazy, brak akcentu frazowego
2. Koniec frazy, akcent frazowy
3. Koniec frazy, brak akcentu frazowego
4. Środek frazy, akcent frazowy

Tekst I.

Zawsze przekręcasz moje słowa. Nie powiedziałem, że Marek sprzedaje warzywa. Powiedziałem, że teraz ***Marek***[1] *sprzedaje pieczywo. Wczoraj znalazłem stare zdjęcia. To jest chyba* ***Czarek***[2]*. Popatrz, jak się zmienił. Źle patrzysz. Nie w środku, tylko z boku jest widoczny* ***Czarek***[3]*. Tu jest Ewa Piotrowska, a tu* ***Darek***[4] *Piotrowski.*

Tekst II.

Źle mnie zrozumiałeś. Nie powiedziałem, że Czarek czeka od drugiej. Powiedziałem, że pewnie ***Czarek***[1] *czeka od czwartej. Właśnie oglądałem stare zdjęcia. To jest chyba* ***Darek***[2]*, przy nim chyba Dorota. Źle patrzysz. Nie w środku, tylko z boku widoczny jest* ***Darek.***[3] *Tu jest Ewa Piotrowska, a tu jest* ***Marek***[4] *Piotrowski.*

Tekst III.

Nastąpiło nieporozumienie. Nie powiedziałem, że Darek poszedł do domu. Powiedziałem, że znowu ***Darek***[1] *poszedł do kina. Spójrz na te stare zdjęcia. To jest chyba* ***Marek***[2]*, poprzedni mąż Zosi. Źle patrzysz. Nie w środku, tylko z boku jest widoczny* ***Marek***[3]*. Tu jest Ewa Piotrowska, a tu jest* ***Czarek***[4] *Piotrowski.*

ZAŁĄCZNIK 7

A.Co słychać u Zosi? Jak wygląda?

B.*Dobrze* *wygląda.*

A.A *ja słyszałam że mocno zmizerniała.*

B.*Nieprawda, Zosia* ***dobrze*** *wygląda.*

A.*Zosia (wygląda* ***dobrze)*** *prosiła, by Cię pozdrowić.*

B.*Jak się czujesz* ***? Dobrze?***

A.*Czuję się* ***dobrze*** *ale wyglądam okropnie.*

B.*O co ci chodzi?! Wyglądasz* ***dobrze!***

A. *Jak wygląda Janek po chorobie?*

B. *Janek wygląda dobrze, chociaż wydaje mi się, że zeszczuplał.*

A.*A ty jak uważasz ? Janek wygląda* ***dobrze?***

B.*Myślę, że tak. Wczoraj zdałam trudny egzamin. Uzyskana ocena* ***(dobrze)*** *podniosła mnie na duchu.*

A. *Co powiesz o Zosi?*

B. *Zosia wygląda dobrze — zdrowe i wesołe dziecko. Martwię się natomiast Jankiem. Wygląda bardzo źle.*

A. *To nieprawda! Janek wygląda* ***dobrze!***

A. *Wiadomość (Janek wygląda dobrze) była pocieszająca. Podniesiona na duchu kupiłam kapelusz. Jak w nim wyglądam?*

B.*Wyglądasz* ***dobrze,*** *chociaż wole cię w berecie.*

A. *Jak znajdujesz Krystynę?*

B.*Wygląda* ***dobrze —*** *zadbana i elegancka.*

A. *Musisz to powiedzieć głośno i wyraźnie.*

B.*Dobrze —* *głośno i wyraźnie. Jak mnie słyszysz*

A. *Dobrze.*

A. *Jak wyglądam w tej sukience ?*

B.*Wyglądasz* ***dobrze.***

A.*A ty jak myślisz ? Wyglądam* ***dobrze?***

B. *Moim zdaniem doskonale, Zosiu. Przyjdź jutro wieczorem.*

A. Dobrze.

A. *Wydaje mi się, że dziewczynki wyrosły już z tych sukienek.*

B.*Nie przesadzaj. Ania wygląda* ***dobrze.***

A. *Idę do domu.*

B.*Dobrze,* *chociaż wolałbym, żebyś jeszcze chwilę posiedziała.*

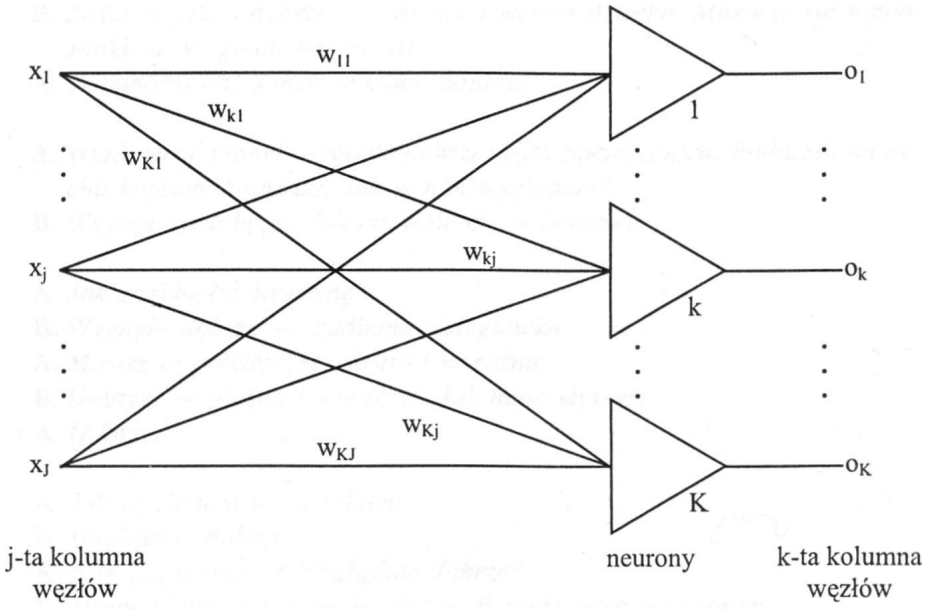
ZAŁĄCZNIK 8

ALGORYTM PROPAGACJI WSTECZNEJ

Algorytm propagacji wstecznej błędów (backpropagation) opracowany przez Rumelharta (1986) umożliwia praktyczne uczenie wielowarstwowej sieci. Metoda polega na tym, że mając wyznaczony błąd δ(m)(j) występujący podczas realizacji j-tego kroku procesu uczenia w neuronie m — można rzutować ten błąd wstecz do wszystkich tych neuronów, których sygnały stanowiły wejścia dla m-tego neuronu.

Wejściowe i wyjściowe sygnały sieci oznaczono odpowiednio przez xi oraz ok. Waga wkl związane jest wyjściem j-tego neuronu (j = 1, 2, ... J) oraz wejściem k--tego neuronu. Przy oznaczeniach jak na rys. 9.1, sygnał wyjściowy sieci określony jest zależnością 1.1.

$$\mathbf{o} = \Gamma\,[\,\mathbf{WX}\,] \qquad (1.1)$$



Ryc. 1. Jednowarstwowa sieć z neuronami ciągłymi gdzie wejście i wyjście oznaczono wektorami:

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_J \end{bmatrix}, \quad \mathbf{O} = \begin{bmatrix} O_1 \\ O_2 \\ \vdots \\ O_K \end{bmatrix}$$

wektor wag jako macierz:

$$\mathbf{W} = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1J} \\ w_{21} & w_{22} & \cdots & w_{2J} \\ \vdots & \vdots & \cdots & \vdots \\ w_{K1} & w_{K2} & \cdots & w_{KJ} \end{bmatrix}$$

Nieliniowy operatorem Γ [·] zdefiniowano jako macierz:

$$\Gamma[\cdot] = \begin{bmatrix} f(\cdot) & 0 & \dots & 0 \\ 0 & f(\cdot) & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & f(\cdot) \end{bmatrix}$$

gdzie f( ) są funkcjami aktywacji neuronów.

Pożądany sygnał wyjściowy z — określono wektorem:

$$\mathbf{Z} \triangleq \begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_K \end{bmatrix} \tag{1.2}$$

Wektor aktywacji netk neuronu k może być wyrażony jako:

$$\mathbf{net_k} = \bar{\mathbf{W}}\ \bar{\mathbf{X}} \tag{1.3}$$

Błąd klasyfikacji jednego obrazu wejściowego y p (gdzie p = 1, 2 ... P) jest sumą kwadratów błędów na wszystkich wyjściach k= 1, 2, ... K wynosi (zależność 1.4)

$$Q_p = \frac{1}{2}\sum_{k=1}^{K}(z_{pk} - o_{pk})^2 = \frac{1}{2}\| \mathbf{z_p} - \mathbf{o_p} \|^2 \tag{1.4}$$

Dla y_J zakłada się stałą wartość wejścia sieci. Wagi w_{kJ} (dla k = 1,2...K) są w trakcie uczenia traktowane jak pozostałe.

Korekcja wag określona jest następująco (1.5)

$$\Delta w_{kj} = -\eta \frac{\delta Q}{\delta w_{kj}} \tag{1.5}$$

Dla każdego połączenia *k* obowiązuje (1.6)

$$net_k = \sum_{j=1}^{J} w_{kj}\, x_j \tag{1.6}$$

Wyjście neuronu opisuje zależność 1.7.

$$o_k = f(net_k) \tag{1.7}$$

Błąd *delta* pochodzący z k-tego neuronu zdefiniowany jest następująco (1.8)

$$\delta_{ok} \stackrel{\Delta}{=} -\frac{\delta Q}{\delta(net_k)} \tag{1.8}$$

Składowa gradientu (1.9) zależy od netk pojedynczego neuronu:

$$\frac{\delta Q}{\delta w_{kj}} = \frac{\delta Q}{\delta(net_k)} \cdot \frac{\delta(net_k)}{\delta w_{kj}} \tag{1.9}$$

Wartości xj dla j= 1,2...J są stałe (1.10)

$$\frac{\delta(net_k)}{\delta w_{kj}} = x_j \tag{1.10}$$

Uwzględnienie wzorów 1.8 i 1.10 prowadzi do następującej zależności (1.11):

$$\frac{\delta Q}{\delta w_{kj}} = -\delta_{ok} x_j \tag{1.11}$$

Generalną regułę delta określa wyrażenie (1.12)

$$\Delta w_{kj} = \eta \delta_{ok} x_j, \quad \text{dla } k = 1, 2, ..., K \text{ i } j = 1, 2, ..., J \tag{1.12}$$

Metodę uczenia delta dla sieci jednowarstwowych ilustruje wyrażenie 1.13

$$w'_{kj} = w_{kj} + \eta(z_k - o_k) f'(net_k) x_j \tag{1.13}$$

Regułę 1.13 można uogólnić dla sieci wielowarstwowych. Zasadę korekcji wag w warstwie ukrytej (warstwie, której wyjścia nie są bezpośrednio dostępne) określa zależność 1.14

$$v'_{ji} = v_{ji} + \eta f'_j(net_j) x_i \sum_{k=1}^{K} \delta_{ok} w_{kj} \tag{1.14}$$

dla j = 1,2,...J - 1, i = 1,2,...,I

Wzór ten wyraża uogólnioną regułę delta. Korekcja wag dochodzących do j-tego neuronu w warstwie ukrytej jest proporcjonalna do sumy ważonej wszystkich wartości δ w warstwie następnej. Metodę tę określa się nazwą propagacja wsteczna (backpropagation).

ZAŁĄCZNIK 9

Fragment wyników klasyfikacji 9 akcentów rdzennych. Sieć probabilistyczna
KL — klasyfikacja otrzymana,
T.KL — klasyfikacja oczekiwana,
E.KL — decyzja,
Error — błąd klasyfikacji

KL	T. KL	E. KL	Error	KL	T. KL	E. KL	Error
1	2	3	4	5	6	7	8
LM	LM	Right	1.9e-17	HL	HL	Right	2.823e-11
LM	LM	Right	0.004755	LH	LH	Right	3.728e-11
ML	ML	Right	1.17e-06	LH	LH	Right	0.0007528
LL	LL	Right	0.02152	ML	ML	Right	2.077e-06
LL	LL	Right	3.619e-07	7	ML	7	0.07839
HL	HL	Right	1.84e-09	LM	LM	Right	0.0004151
LH	LH	Right	3.535e-07	7	LM	7	0.02487
LH	LH	Right	2.632e-18	ML	ML	Right	4.362e-05
ML	ML	Right	0.002651	ML	ML	Right	2.329e-08
ML	ML	Right	2.148e-07	LL	LL	Right	4.252e-06
LM	LM	Right	0.0001475	LL	LL	Right	0.00809
LM	LM	Right	0.0003623	HL	HL	Right	2.2e-18
LL	LL	Right	2.527e-08	LH	LH	Right	5.116e-16
ML	ML	Right	0.000515	LH	LH	Right	3.982e-19
ML	ML	Right	5.213e-06	ML	ML	Right	2.273e-05
LL	LL	Right	0.002513	ML	ML	Right	0.0009937
7	HL	7	0.03309	LM	LM	Right	7.409e-19
LH	LH	Right	0.005556	LM	LM	Right	5.998e-09
LH	LH	Right	7.338e-14	ML	ML	Right	0.0199
ML	ML	Right	6.838e-11	LM	LM	Right	7.409e-19
LL	LL	Right	5.441e-06	LM	LM	Right	1.974e-05
HL	HL	Right	4.579e-20	LM	LM	Right	3.088e-08
LH	LH	Right	1.023e-16	LM	LM	Right	6.392e-07
LH	LH	Right	4.987e-18	ML	ML	Right	0.01312
ML	ML	Right	6.038e-13	7	LL	7	0.07266

1	2	3	4	5	6	7	8
ML	ML	Right	2.063e-09	LL	LL	Right	0.0006577
LM	LM	Right	1.408e-11	HL	HL	Right	4.043e-19
LM	LM	Right	4.443e-14	LH	LH	Right	1.665e-05
ML	ML	Right	2.379e-13	LH	LH	Right	1.904e-16
ML	ML	Right	2.593e-08	?	ML	?	0.0906
?	LL	?	0.03112	?	ML	?	0.03334
LL	LL	Right	9.661e-10	LM	LM	Right	5.546e-09

ZAŁĄCZNIK 10

Fragment wyników klasyfikacji 12 struktur akcentowych. Sieć MLP

KL	T. KL	E. KL	Error	KL	T. KL	E. KL	Error
1	2	3	4	5	6	7	8
ML	ML	Right	6.459e-06	ML	ML	Right	8.007e-35
ML	ML	Right	2.576e-12	P	P	Right	1.218e-07
ML	ML	Right	2.079e-13	L	L	Right	0.0002757
LM	LM	Right	7.396e-16	L	L	Right	6.859e-08
LM	LM	Right	3.563e-18	L	L	Right	5.069e-11
LM	LM	Right	1.821e-08	HL	HL	Right	1.198e-12
P	P	Right	1.117e-07	ML	ML	Right	8.972e-06
ML	ML	Right	0.0001123	ML	ML	Right	1.397e-19
P	P	Right	1.584e-08	ML	ML	Right	2.405e-14
ML	ML	Right	1.17e-12	LM	LM	Right	7.145e-16
H	H	Right	0.0001409	LM	LM	Right	4.482e-20
H	H	Right	1.155e-10	LM	LM	Right	4.145e-11
LL	LL	Right	1.015e-30	P	P	Right	0.007407
H	H	Right	1.348e-10	ML	ML	Right	7.814e-06
LL	LL	Right	0	7	P	?	0.1302
L	L	Right	0.0002763	ML	ML	Right	6.188e-05
HL	HL	Right	4.29e-18	H	H	Right	9.773e-13
L	L	Right	6.229e-08	?	H	?	0.02695
HL	HL	Right	1.055e-13	LL	LL	Right	2.615e-17
L	L	Right	7.132e-05	H	H	Right	1.278e-07
L	L	Right	7.704e-06	LL	LL	Right	1.067e-07
HL	HL	Right	2.215e-28	L	L	Right	8.848e-08
P	P	Right	2.269e-06	HL	HL	Right	2.61e-10
L	L	Right	0.001112	L	L	Right	2.69e-09
L	L	Right	9.421e-10	HL	HL	Right	3.628e-07
HL	HL	Right	1.703e-35	L	L	Right	3.595e-11
H	H	Right	1.578e-06	L	L	Right	1.171e-07
H	H	Right	4.849e-07	HL	HL	Right	2.612e-05
HL	HL	Right	1.936e-40	?	P	7	0.01977
H	H	Right	1.336e-07	L	L	Right	3.29e-15
ML	ML	Right	2.018e-11	L	L	Right	0.0001917

1	2	3	4	5	6	7	8
LH	LH	Right	1.741e-12	HL	HL	Right	2.551e-23
LH	LH	Right	1.26e-41	H	H	Right	3.147e-23
P	P	Right	0.007674	H	H	Right	2.302e-08
H	H	Right	8.903e-07	HL	HL	Right	1.468e-26
H	H	Right	1.518e-05	H	H	Right	2.521e-08
H	H	Right	0.0002795	ML	ML	Right	0.0001098

ZAŁĄCZNIK 11

Fragment wyników klasyfikacji akcentów w tekście czytanym

N — samogłoski nieakcentowane

A — samogłoski akcentowane

F — samogłoski ostatnie przed granicą frazy

Sylaby pożądane	Klasyfikacje			Klasyfikacje z sieci neuronowej			Błędy
	N	A	F	N	A	F	
1	2	3	4	5	6	7	8
wi	1	0	0	0.999140	0.000000	0	1
zyj	1	0	0	0.270711	0.000000	0.001453	
ne	0	1	0	0.061830	0.854130	0.029656	
go	1	0	0	0.996365	0.000040	0	
ki	0	1	0	0.078490	0.852090	0.031074	
nna	1	0	0	0.995943	0.000066	0	
noc	1	0	0	0.815399	0.178767	0	
ne	0	0	1	0.231000	0.000000	0.969692	
go	0	0	1	0.000000	0.000000	1	
zsym	1	0	0	0.996357	0.000050	0	
pa	0	1	0	0.117367	0.798471	0.003704	
tja	1	0	0	0.995634	0.000072	0	
wspo	1	0	0	0.891583	0.000030	0	
mi	0	0	1	0.789550	0.000000	0	3
nam	0	0	1	0.216625	0.000000	0.987578	
czlo	1	0	0	0.997438	0.000000	0	
wje	0	0	1	0.799093	0.000000	0	3
ka	0	0	1	0.000072	0.000000	1	
ktu	0	1	0	0.082234	0.848338	0.031757	
ry	1	0	0	0.995934	0.000067	0	
sie	1	0	0	0.994960	0.000103	0	
zmniej	0	0	1	0.000000	0.000000	1	
szal	0	0	1	0.013068	0.000000	1	
film	0	1	0	0.000000	0.996440	0.001342	
zga	1	0	0	0.995543	0.000000	0	
tun	0	1	0	0.000000	1.000000	0	

1	2	3	4	5	6	7	8
ku	1	0	0	0.996370	0.000053	0	
fi	0	1	0	0.082264	0.849118	0.031585	
kcji	1	0	0	0.996374	0.000053	0	
nna	1	0	0	0.934174	0.019122	0	
ko	0	0	1	0.235810	0.000000	0.996099	
wej	0	0	1	0.000000	0.000000	1	
A	1	0	0	0.996064	0.000062	0	
wienc	0	1	0	0.000000	0.000000	0.999971	2
ra	0	1	0	0.080455	0.850007	0.031443	
czejj	1	0	0	0.966530	0.000000	0.00001	
fan	1	0	0	0.869872	0.086106	0	
ta	1	0	0	0.814502	0.179708	0	
sty	0	0	1	0.426213	0.502255	0	3
czny	0	0	1	0.000027	0.000000	0.999742	
i	1	0	0	0.996018	0.000057	0	
ra	0	1	0	0.000000	1.000000	0	
czej	1	0	0	0.991900	0.000098	0	
nie	0	1	0	0.002005	0.844790	0.009233	
nna	1	0	0	0.995605	0.000077	0	
ko	0	0	1	0.228063	0.000000	0.999883	
wy	0	0	1	0.000000	0.000000	1	
A	0	1	0	0.000000	1.000000	0	
le	1	0	0	0.816054	0.177644	0	
za	0	1	0	0.000000	1.000000	0	
to	1	0	0	0.991332	0.000115	0	
po	1	0	0	0.814200	0.178759	0	
bu	1	0	0	0.996371	0.000053	0	
dza	1	0	0	0.870416	0.085575	0	
jon	0	i	0	0.000153	1.000000	0	
cy	1	0	0	0.996249	0.000056	0	
wy	1	0	0	0.996371	0.000053	0	
bra	0	0	l	0.000000	0.000000	1	
znie	0	0	1	0.000000	0.000000	1	
by	0	1	0	0.000000	1.000000	0	
la	1	0	0	0.885677	0.060984	0	
to	0	1	0	0.000000	1.000000	0	
0	1	0	0	0.995251	0.000076	0	
po	0	1	0	0.000000	1.000000	0	
wiesc	1	0	0	0.996336	0.000054	0	

I	2	3	4	5	6	7	8
0	1	0	0	0.996369	0.000053	0	
pa	0	0	I	0.225843	0.000000	0.999959	
nu	0	0	1	0.000000	0.007442	0.977443	
ktu	0	1	0	0.000000	1.000000	0	
ry	1	0	0	0.992654	0.000176	0	
za	0	1	0	0.025069	0.976432	0.008807	
czol	1	0	0	1.000000	0.000000	0	
sie	1	0	0	0.996335	0.000054	0	
zmniej	0	0	1	0.002757	0.913692	0.026527	3
szac	0	0	1	0.006142	0.000000	1	
0	1	0	0	0.820854	0.105778	0	
czy	1	0	0	0.996349	0.000054	0	
wi	0	1	0	0.000000	0.999396	0.000272	
scie	1	0	0	0.995597	0.000078	0	
po	0	1	0	0.082356	0.849034	0.031508	
za	1	0	0	0.995517	0.000000	0	
tym	1	0	0	0.995765	0.000072	0	
pa	0	0	1	0.225667	0.000000	0.999959	
nem	0	0	1	0.070837	0.000000	1	
wszy	0	1	0	0.082440	0.848913	0.031303	
scy	1	0	0	0.984974	0.000947	0	
to	0	1	0	0.083186	0.847781	0.02972	
ba	1	0	0	0.993870	0.000153	0	
ga	1	0	0	0.800097	0.000011	0	
te	1	0	0	0.851289	0.114261	0	
li	1	0	0	0.760278	0.219084	0	
zo	1	0	0	0.995762	0.000060	0	
wa	0	0	1	0.000000	0.000166	0.995945	
li	0	0	1	0.065731	0.000000	0.999952	
ZO	0	1	0	0.048063	0.382365	0.15282	
nna	1	0	0	0.984484	0.001008	0	
po	1	0	0	0.996368	0.000053	0	
wta	1	0	0	0.828701	0.075962	0	
rza	0	1	0	0.000000	0.999578	0.000082	
la	0	0	1	0.272905	0.000000	0.230781	3
mu	0	0	1	0.000000	0.000073	0.996183	

ZAŁĄCZNIK 12

Teksty czytane

Zapoczątkowana przed czterdziestoma laty odbudowa zamku ma się ku końcowi. Ten wspaniały obraz architektonicznego geniuszu uczy pokory i dziś, kiedy technologia budowlana daje dużo większe możliwości. Zamek przetrwał wiele burz historii, kilka pożarów oraz próby nieudolnej przebudowy i zniszczenia. Dziś każdy, kto zna się na rzeczy, nie może nie przyznać, że dawni mistrzowie dali dowód swych wielkich talentów, wznosząc budowlę tak monumentalną i zniewalającą siłą oddziaływania. Od stuleci ich dzieło cieszy oczy licznych rzesz laików i znawców.

Moja żaba

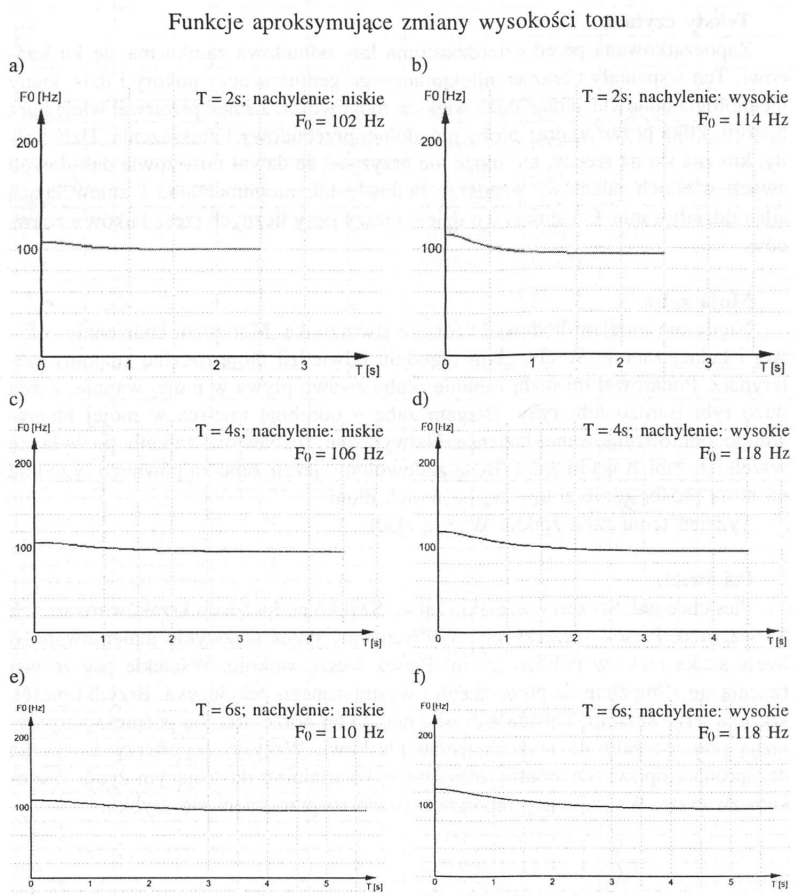
Nigdy nie mogłem hodować żadnego stworzonka. Marzyłem codziennie o żywej i ładnej żabuni. W ubiegłym tygodniu odwiedził moją rodzinę znajomy weterynarz. Podarował mi małą żabunię. Żaba żwawo pływa w mojej wannie. Zjada dużo ryb. Bardzo lubi ryby. Błagam żabę o odrobinę miejsca w mojej nieprawdopodobnie zrujnowanej łazience. Najwyraźniej rozbawiona żabunia pozbyła się wszelkich żabich wad i zalet. Będę zadowolony, jeżeli żabunia pewnego tygodnia na moją prośbę ufnie zbliży się do moich dłoni.

Tydzień temu żaba znikła. Wąż ją zjadł.

Psi świat.

Pies chce pić. Szczerzy wściekle ząbki. Szybko podpełza do krzaków rosnących na wzgórzu. Piesek szczerzy kły. Wytresowany przez niezwykle utalentowanego hycla szuka rzeki w pobliżu trzcin. Piesek węszy wokoło. Wściekłe psy ze wsi rzucają się z impetem na piszczącego i wystraszonego pekińczyka. Brzydki piesek szczeka teraz smutno. Zapada wczesna noc. O tej porze nikt nie pośpieszy z przyjazną pomocą żałośnie szczekającemu pieskowi. Wszyscy reporterzy z wysoką dezaprobatą opiszą ten smutny incydent w kwartalniku ilustrującym życie szkockich zwierząt. Wstrząsający reportaż o psach rozprzestrzeni się szybko.

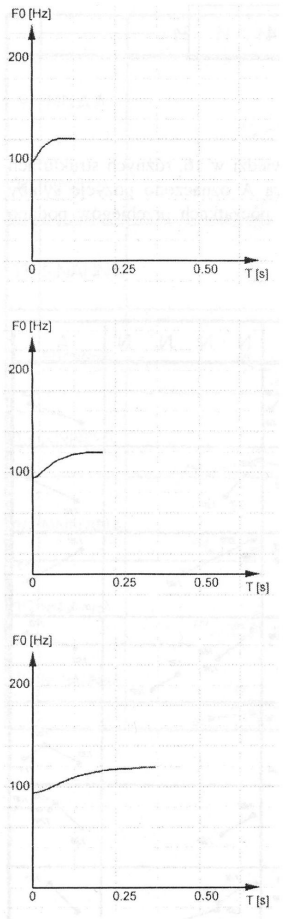
ZAŁĄCZNIK 13



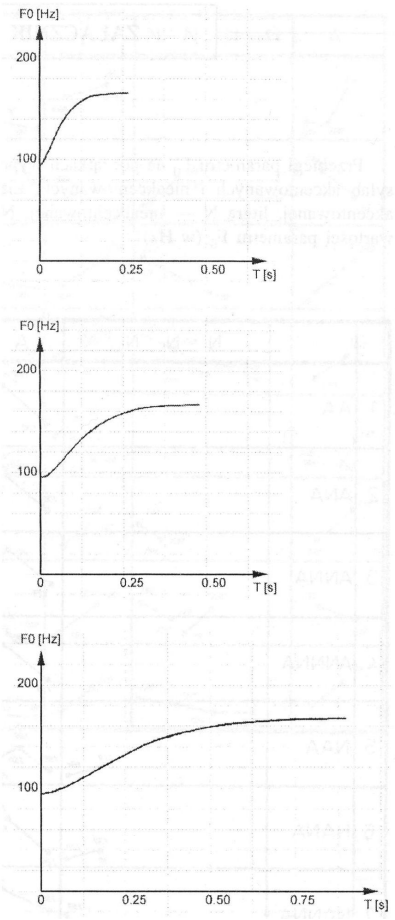
Ryc. 1. Funkcje aproksymujące zmiany wysokości tonu. Składowa frazowa

a) w zdaniu 2s (niskie) c) w zdaniu 4s (niskie) e) w zdaniu 6s (niskie)

b) w zdaniu 2s (wysokie) d) w zdaniu 4s (wysokie) f) w zdaniu 6s (wysokie)



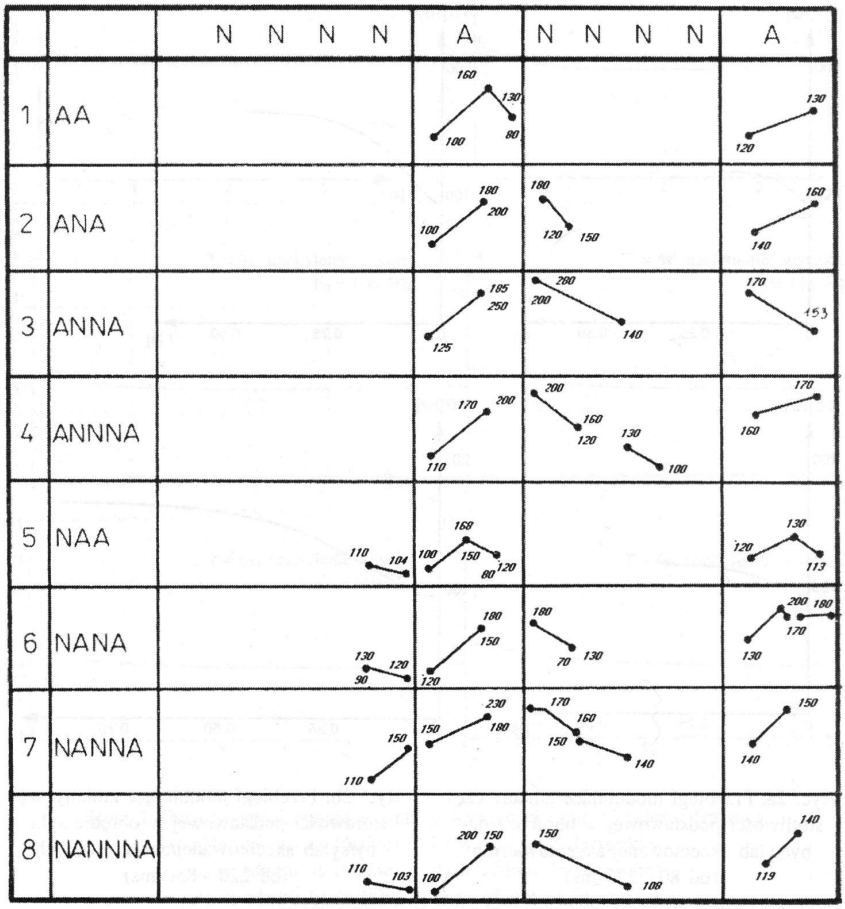
Ryc. 2a. Przebiegi modulujące zmiany częstotliwości podstawowej w obrębie sylaby/sylab akcentowanej/akcentowanych (od 80 - 220 ms)

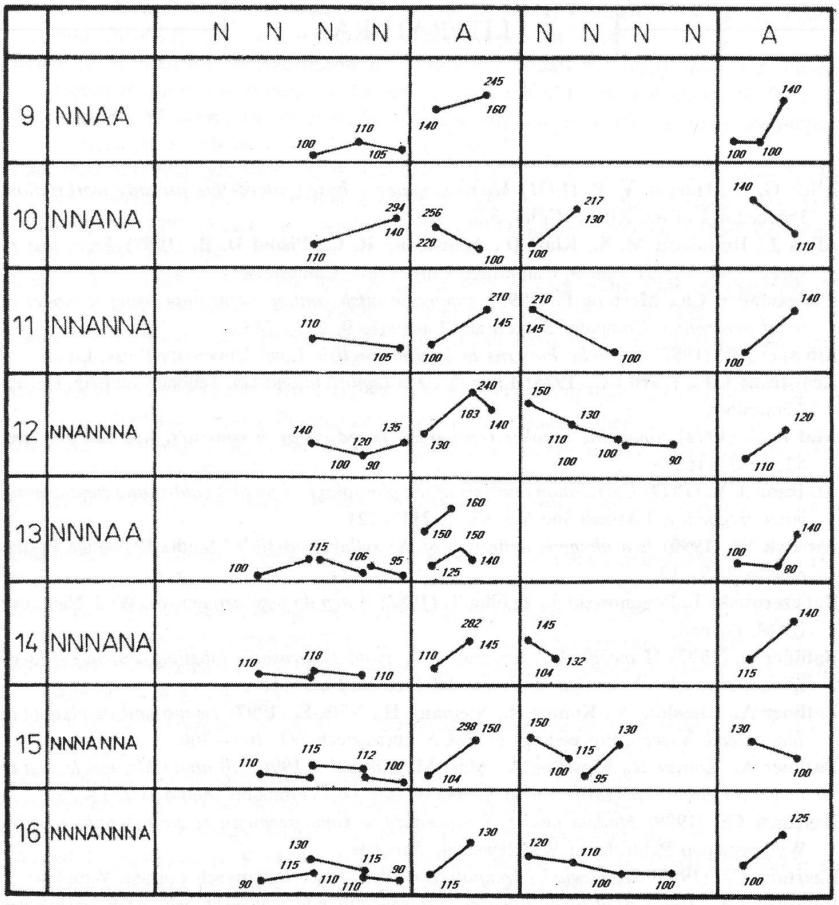


Ryc. 2b. Przebiegi modulujące zmiany częstotliwości podstawowej w obrębie sylaby/sylab akcentowanej/akcentowanych (od 220 - 850 ms)

ZAŁĄCZNIK 14

Przebiegi parametru F0 na początkach wypowiedzi w 16. różnych strukturach sylab akcentowanych i nieakcentowanych. Literą A oznaczono pozycję sylaby akcentowanej, literą N — nieakcentowanej. Na początkach przebiegów podano wartości parametru F0 (w Hz)




N N N N
A
N N N N
A
9 NNAA
10 NNANA
11 NNANNA
12 NNANNNA
13 NNNAA
14 NNNANA
15 NNNANNA
16 NNNANNNA

LITERATURA

Allen G. D., Harper V. P. (1991) *Microcomputer - based interactive prosody workstation,* Proceedings of the XII th ICPhS, 322-324.

Allen J., Hunnicutt M. S., Klatt D., Armstrong R. C., Pisoni D. B. (1987) *From text to speech: the MITalk system,* Cambridge Univ. Press, Cambridge.

d'Alessandro Ch., Mertens P. (1995) *Automatic pitch contour stylization using a model of tonal perception,* Computer Speech and Language 9, 257 - 288.

Altenberg B. (1987) *Prosodic Patterns in Spoken English,* Lund University Press, Lund.

Armstrong L.E., Ward I.C. (1926) *A handbook of English intonation,* Teubner, Leipzig, Heffer, Cambridge.

Atal B. S. (1972) *Automatic speaker recognition based on pitch contours,* J.Acoust.Soc.Am. 52, 1687- 1697.

Atkinson J. E. (1977) *Correlation analysis of the physiological factors controlling fundamental voice frequency,* J.Acoust.Soc.Am. 63, 1, 211 -221.

Awedyk W. (1990) *Is a phonetic definition of the syllable possible?* Studia Phonetica Posnaniensia 2, 5-12.

Bańczerowski J., Pogonowski J., Zgółka T. (1982) *Wstęp do językoznawstwa,* Wyd. Naukowe UAM, Poznań.

Batliner A. (1997) *M specified: A revision of the syntactic-prosodic labelling system for large spontaneous speech databases,* Verbmobil Memo 124, 1 - 16.

Batliner A., Kiessling A., Kompe R., Niemann H., Noth E. (1997) *Tempo and its change in spontaneous speech,* Proceedings of ESCA Eurospeech '97, 763 - 766.

Batliner A., Kompe R., Kiessling A., Mast M., Noth E. (1996) *All about Ms and Is, not to forget As, and a comparison with Bs and Ss and Ds,* Verbmobil Memo 102, 1 -29.

Basztura Cz. (1989) *Modele analizy i procedury w komputerowym rozpoznawaniu głosów,* Wydawnictwo Politechniki Wrocławskiej, Wrocław.

Basztura Cz. (1993) *Rozmawiać z komputerem,* Wyd. Prac Naukowych Format, Wrocław.

Basztura Cz. (1994) *Analiza możliwości zastosowania podstawowych jednostek językowych w automatycznym rozpoznawaniu mowy,* w *Analiza, synteza i rozpoznawanie sygnału mowy dla celów automatyki, informatyki i lingwistyki, medycyny,* Materiały z Seminarium PTFon, Poznań 1993, 56-79.

Berkovits R. (1993) *Utterance - final lengthening and the duration of final stop closures,* Journal of Phonetics, vol. 21, 4, 479-489.

Berkovits R. (1994) *Durational effects in final lengthening, gapping, and contrastive stress,* Language and Speech, vol. 37, 3, 237 - 250.

Bolinger D. (1958) *A theory of pitch accent in English,* Word 14, 109- 149.

Bolinger D. (1989) *Intonation and its uses: Melody in grammar and discourse,* London, Edward Arnold.

Brandt S. (1998) *Analiza danych,* PWN, Warszawa.

Bruce G. (1995) *Modelling Swedish intonation for read and spontaneous speech,* Proceedings of the ICPhS '95 Stockholm, vol. 2, 28 - 33.

Bruce G., Granström B., Gustafson K., House D. (1991) *Prosodic Phrasing in Swedish,* Lund University, Dept, of Linguistics, Working Papers 38, 5 - 17.

Bruce G., Granström B., Gustafson K., House D., Touati P. (1995) *Modelling Swedish prosody in a dialogue framework,* Proceedings of the ICSLP '94, Yokohama, 1099- 1102.

Butzberger J. W. (1990) *Statistical methods for analysis and recognition of intonation patterns in speech,* B. S., Miami University Reports 1990.

Campbell N. (1992) *Syllable-based segmental duration,* in *Talking Machines: Theories, Models and Designs,* G. Bailly, C. Benoit and T.R. Sawalis, ed., Amsterdam, Elsevier Science.

Campbell N. (1993) *Automatic detection of prosodic boundary in speech,* Speech Communication 13, 345-354.

Campbell N. (1997) *Synthesing Spontaneous Speech* in *Computing Prosody* (Sagisaka Y., Campbell N., Higuchi N. ed.), Springer-Verlag New York, Inc., 165-185.

Campbell N., Isard S. D. (1991) *Segment durations in a syllable frame,* Journal of Phonetics 19, 37-47.

Campione E., Flachaire E., Hirst D., Veronis J. (1997) *Stylization and symbolic coding of Fo: a quantitative model,* Proceedings of ESCA Workshop, Athens, 71-73.

Chen S., H., Wang Y. R. (1995) *Tone Recognition of Continuous Mandarin Speech based on Neural Networks,* IEEE Transactions on Speech and Audio Processing, vol.3, 146- 151.

Collier R. (1975) *Physiological correlates of intonation patterns,* J.Acoust.Soc.Am. 58, 1, 249- -255.

Collier R. (1990) *On the perceptual analysis of intonation,* Speech Communication 9, 443 - 451. Collier R. (1991) *Multi-language intonation synthesis,* Journal of Phonetics, 19, 61 -73.

Collier R. (1992) A *comment on the prediction of prosody,* in *Talking Machines, Theories, Models and Designs,* G. Bailly, C. Benoit and T. R. Sawallis, ed., Elsevier Science Publishers, 205 - 207.

Cooper W. E., Sorensen J. M. (1977) *Fundamental frequency contours at syntactic boundaries,* J.Acoust.Soc.Am, vol. 62, 682 - 692.

Cooper W. E., Sorensen J. M. (1981) *Fundamental frequency in sentence production,* Springer, New York.

Cruttenden A. (1986) *Intonation,* Cambridge Univ. Press, Cambridge.

Cruttenden A. (1997) *Intonation,* Cambridge Univ. Press, Cambridge.

Crystal D. (1969) *Prosodic Systems and Intonation in English,* Cambridge.

Crystal T. H., House A. S. (1990) *Articulation rate and the duration of syllables and stress groups in connected speech,* J.Acoust.Soc.Am, vol. 88, 1, 101-111.

Delattre P. (1966) *A comparison of syllable length conditioning among languages,* International Review of Applied Linguistics 4, 183 - 198.

Delattre P., Poenack E., Olsen C. (1965) *Some characteristics of German intonation for the expression of continuity and finality,* Phonetica 13, 134- 161.

Demenko G. (1983) *Aproksymacja częstotliwości podstawowej w zdaniach,* Warszawa, Prace IPPT, 38/83, 1 - 39.

Demenko G. (1984) *Analiza matematyczna cech osobniczych głosu w zakresie parametru F0,* Warszawa, Prace IPPT, 24/84, 1 - 52.

Demenko G. (1986) *Formalizacja matematyczna oraz rozpoznawanie automatyczne i percepcyjne elementów ograniczonego zbioru przebiegów częstotliwości podstawowej w mowie polskiej,* rozprawa doktorska, Poznań.

Demenko G. (1987) *Wizualizacja częstotliwości podstawowej (parametru F0) mowy* (Visible F0 in Speech) w *Wizualizacja mowy i jej zastosowania,* (ed. W. Jassem), Warszawa, 65 - 88.

Demenko G., Pruszewicz A., Wika T. (1989) *Analiza periodyczności w mowie osób głuchych,* Warszawa, Prace IPPT, 22/89.

Demenko G. (1995a) *Baza danych tekstowych dla analizy intonacji języka polskiego,* Prace IPPT, 20/95.

Demenko G. (1995b) *Synteza podstawowych typów przebiegów intonacyjnych,* Prace IPPT, 19/95, 1 -25.

Demenko G. (1995c) *Struktury prozodyczne* w *systemach rozpoznawania mowy,* Materiały I Krajowej Konferencji: Głosowa Komunikacja Człowiek - Komputer, 207-210.

Demenko G. (1997) *Wyznaczniki fonetyczno-akustyczne granicy frazy i akcentu w języku polskim,* in *Speech and Language Technology,* vol. 1, Wyd. PTFon, Wrocław, 101 - 125.

Demenko G. (1998) *Acoustic classification and automatic recognition of accent and phrase boundary in speech signal,* Archives of Acoustics, vol.23, 2, 159- 178.

Demenko G., Jassem W., Krzyśko M. (1988) *Classification of basic F0 patterns using discriminant functions,* Phonetica 41, 1- 12.

Demenko G., Jassem W. (1999) *A text-to-speech oriented comparison of English and Polish intonation,* Acta Acustica, vol. 85, 51-52.

Demenko G., Möbius B., Pätzoldt M. (1990) *Statistische Analyse zur Klassifizierung von Intonationsverlaufen,* Proceedings of DAGA '90, Wiedeń, 1051 - 1054.

Demenko G., Nowak I., Imiołczyk J. (1993) *Analysis and Synthesis of Pitch Movements in a Read Polish Text,* Proceedings of the 3rd Eurospeech '93, Berlin, vol. 2, 797 - 800.

Demenko G., Pruszewicz A. (1994) *Audiometria mowy* w *Zarys Audiologii klinicznej* pod red. A. Pruszewicza, Poznan, 219 - 244.

Demenko G., Richter L., Pruszewicz A., Szyfter W., Woźnica B. (1996a) *Testy do badania słuchowej percepcji mowy (TBSPM),* Otolaryngologia Polska, 50, 628 - 632.

Demenko G., Richter L., Serwat P. (1996b) *Analiza statystyczna wyników percepcyjnej oceny granic frazowych i akcentu w języku polskim,* Materiały XLIII Otwartego Seminarium z Akustyki OSA '96, 167 - 173.

De Pijper J. R. (1983) *Modelling British English Intonation,* Foris Publication, Dordrecht.

De Pijper J. R., Sanderman A. A. (1993) *Prosodic cues to the perception of constituent boundaries,* Proc. of Eurospeech '93, Berlin, 1211 - 1214.

De Pijper J. R., Sandermann A. A. (1994) *On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues*, J.Acoust.Soc.Am., vol.96, 2037 - 2047.

Di Cristo A., Hirst D. J. (1986) *Modelling French micromelody: analysis and synthesis,* Phonetica 43, 11- 30.

Dłuska M. (1957) *Akcenty i atona w języku polskim,* w Studia z filologii polskiej i słowiańskiej, t.2, Warszawa.

Dłuska M. (1976) *Prozodia języka polskiego,* PWN, Warszawa.

Dukiewicz L. (1978) *Intonacja wypowiedzi polskich,* Prace Instytutu Języka Polskiego, wyd. PAN, Wrocław.

Dukiewicz L. (1995a) *Fonetyka* w *Gramatyka współczesnego języka polskiego. Fonetyka i fonologia* (W. Wróbel), Wyd. Instytutu Jęz. Polskiego PAN, Kraków.

Dukiewicz L. (1995b) *Czy sylaba jest hybrydą?* w: *Eufonia i Logos* (J. Pogonowski red.), Wyd. Naukowe UAM, Poznań, 107- 114.

Dukiewicz L., Sawicka I. (1995) *Fonetyka i fonologia,* Kraków.

Dumouchel P., O'Shaughnessy D. (1993) *Prosody and Continuous Speech Recognition,* Proceedings of the 3rd Eurospeech '93, Berlin, 2195-2198.

Dziubalska-Kołaczyk K. (1995) *Phonology without the syllable, a Study in the Natural framework,* Motivex, Poznan.

von Essen O. (1956) *Grundzuge der hochdeutschen Satzintonation,* Henn, Ratingen.

Fant G., Kruckenberg A. (1989) *Preliminaries to the study of Swedish prose reading and reading style,* STL-QPR Reports, May 1989, Stockholm, 1 -83.

Fant G., Kruckenberg A. (1994) *Notes on stress and word accent in Swedish*, STL-QPSR Reports, 2-3/1994.

Fant G., Kruckenberg A. (1995) *The voice source in prosody,* Proceedings of ICPhS '95, vol.2, 622 - 625.

Fant G., Hertegard S., Kruckenberg A., Liljencrants J. (1997) *Covariation of subglottal pressure, F0 and glottal parameters,* Proceedings of Eurospeech '97, 453 -456.

Farley G. (1994) *A biomechanical laryngeal model of voice F0 and glottal width control,* J.Acoust.Soc.Am., vol. 100, 6, 3794 - 3808.

Farley G. R. (1996) *Control of voice Fo by an artificial neural network,* J.Acoust.Soc.Am., vol. 96, 3, 1374-378.

Frąckowiak-Richter L. (1973) *The duration of Polish vowels,* Speech Analysis and Synthesis III, PWN, Warszawa, 87-115.

Freij G., Fallside F. (1988) *Lexical stress recognition using hidden Markov models,* IEEE S3.10, 135-138.

Fry D. B. (1955) *Duration and intensity as physical correlates of stress,* J.Acoust.Soc.Am. 27, 765 -768.

Fry D. B. (1958) *Experiments in the perception of stress,* Language and Speech, 1, 126-152. Fujisaki H. (1981) *Dynamic characteristics of voice Fundamental Frequency* in *Speech and singing. Acoustical analysis and physiological interpretations,* Proceedings of the fourth FASE Symposium on Acoustics and Speech 1981, vol. 2, 55-70.

Fujisaki H. (1983) *Dynamic characteristics of voice fundamental frequency in speech and singing,* in. *The production of speech,* P. F. MacNeilage, Hg, Springer, New York, 39- -55.

Fujisaki H. (1988) *A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour,* in *Vocal physiology: voice production, mechanisms and functions,* O. Fujimura (Hg.), Raven, New York, 347-355.

Fujisaki H. (1997) *Prosody, Models, and Spontaneous Speech,* in *Computing Prosody,* Sagisaka.Y., Campbell, N., Higuchi, N. ed.), Springer-Verlag, New York, Inc. 27-41.

Fukada T., Komori Y., Aso T., Ohora Y. (1994) *A study on pitch pattern generation using HMM-based statistical information,* Proceedings of the ICSLP '94, 723-726.

Gärdning E. (1983) *A generative model of intonation,* in *Prosody: models and measurements,* A. Cutler, D. R. Ladd (hg.), Springer Verlag, Berlin, 11-25.

Gärdning E., Eriksson L. (1989) *Perceptual cues to some Swedish prosodic phrase patterns - a peak shift experiment,* STL-QPSR Reports 1/1989, 13-31.

Gärdning E., Eriksson L. (1991) *On the perception of Prosodic Phrase Patterns,* Lund University, Dept, of Linguistics, Working Papers 38, 45 - 70.

Gatnar E. (1998) *Symboliczne metody klasyfikacji danych,* PWN, Warszawa.

Goldberg D. E. (1998) *Algorytmy genetyczne i ich zastosowania,* WNT, Warszawa.

Grabe E. (1998) *IviE Labelling Guide Version 1.1,* University of Cambridge, Cambridge.

Granström B. (1997) *Applications of Intonation - An Overview,* Proceedings of ESCA Workshop on Intonation Theory, Athens, 21 - 25.

Grice M., Reyelt M., Benzmuller R., Mayer J., Batliner A. (1996) *Consistency in Transcription and Labelling of German Intonation with GToBI,* Report 153, October 1, 1-12.

Grover C., Terken J. (1994) *Rhythmic constraints in durational control,* Proceedings of ICSLP'94, Yokohama, 363-366.

Grocholewski S. (1995a) *Założenia akustycznej bazy danych dla języka polskiego na nośniku CD-ROM,* Materiały I Krajowej Konferencji: Głosowa Komunikacja Człowiek-Komputer, Wrocław, 22 - 28.

Grocholewski S. (1995b) *Niezależny od mówcy system rozpoznawania izolowanych słów,* Materiały I Krajowej Konferencji „Głosowa Komunikacja Człowiek--Komputer", Wrocław, 173 —

- 177.

Grocholewski S. (1997) *CORPORA — Speech Database for Polish Diphones,* Proceedings of the Eurospeech '97, 1735- 1738.

Gubrynowicz R. (1967) *Kwantyzacja częstotliwościowa sygnału mowy do celów automatycznego rozpoznawania samogłosek,* Archiwum Akustyki 2, 255 - 266.

Gubrynowicz R. (1968) *Zastosowanie pasmowej analizy widmowej do określania cech osobniczych głosu,* Prace IPPT, 26/1968, Warszawa.

Gubrynowicz R. (1998) *The Polish database of spoken language,* Proc. International Conference on Language Resources and Evaluation, Grenada, 1031 - 1038.

Gubrynowicz R., Mikiel W., Żarnecki P. (1980) *An acoustic method for the evaluation of the state of the larunx source in cases involving pathological changes in the vocal folds,* Arch. Acoustics 5,3- 30.

Gubrynowicz R., Kacprowski J., Mikiel W., Żarnecki P. (1981) *Detection and evaluation of laryngeal pahology based on pitch period measurements in continuous speech,* Proceedings of the 4th F.A.S.E Symposium, 131 - 134.

Gubrynowicz R. Marasek K., Mikiel W., Więźlak W. (1990) *Simplified system for isolated word recognition,* Arch.of Acoustics, vol. 15, 287-300.

Gussenhoven C., Rietveld A. C. M. (1992) *Intonation contours, prosodic structure and preboundary lengthening,* Journal of Phonetics 20, 283 - 303.

Halliday M. A. K. (1967) *Intonation and Grammar in British English* (the Hague).

Harris M. S., Umeda N. (1987) *Difference limens for fundamental frequency contours in sentences,* J.Acoust.Soc.Am. 8, 4, 1139 - 1145.

Harris M. O., Umeda N., Bourne J. (1981) *Boundary perception in fluent speech,* Journal of Phonetics, vol. 9, 1, 1-18.

't Hart J. (1976) *Psychoacoustic backgrounds of pitch contour stylization,* IPO Annual Progress Report 14, 61 -65.

't Hart J. (1981) *Differential sensitivity to pitch distance, particulary in speech,* J.Acoust.Soc.Am. 69, 811 -821.

't Hart J. (1991) *Fo stylisation in speech: straight lines versus parabolas,* J.Acoust.Soc.Am. 90, 6, 3368-3370.

't Hart J., Collier, R. (1975) *Integrating different levels of intonation analysis,* Journal of Phonetics, 235 -255.

't Hart J., Collier, R., Cohen, A. (1990) *A perceptual study of intonation. An experimental*

-phonetic approach to speech melody, Cambridge University Press, Cambridge.

Hasegawa Y., Hata K. (1992) *Fundamental frequency as an acoustic cue to accent perception,* Language and Speech, vol. 31, part 1-2.

Heffner R. M. S. (1949) *General phonetics,* Wisconsin.

Helfrich H. (1985) *Satzmelodie und Sprachwahrnehmung; Psycholinguistische Untersuchungen zur Grudfrequenz,* R. Posener, G. Meggle De Gruyter, Berlin.

Hermes D, J. (1995) *Timing of pitch movements and accentuation of syllables,* IPO Annual Progress Report 30, 38 - 44.

Hermes D. J., Rump H. (1994) *Perception of prominence in speech intonation induced rising and falling pitch movements,* J.Acoust.Soc.Am. 96, 1, 83-92.

Hermes D., van Gestel J. C. (1991) *The frequency scale of speech intonation,* J.Acoust.Soc.Am. 90, 97- 102.

Hess W. (1983) *Pitch determination of speech signals - algorithms and devices,* Springer Verlag, Berlin.

Hess W. (1992) *Prosodiebezogene Interaktion,* Verbmobil-Arbeitspaket 15.5, Uni Bonn.

Hess W., Batliner A., Kiessling A., Kompe R., Noth E., Petzold A., Reyelt M., Strom V. (1997) *Prosodic Modules for speech Recognition and Understanding in VERBMOBIL,* in *Computing Prosody,* Sagisaka, Y., Campbell, N., Higuchi, N. ed., Springer-Verlag New York, Inc., 361 -381.

Hill D. R., Reid N. A. (1977) *An experiment on the perception of intonation features,* International Journal of Man-Machine Studies 9, 337 - 347.

Hirose K., Sakurai A., Konno H. (1994) *Use of prosodic features in the recognition of continuous speech,* Proceedings of the ICSLP '94, Yokohama, 1123- 1126.

Hirose K. (1997) *Disambiguating Recognition Results by Prosodic Features,* in *Computing Prosody,* Sagisaka Y., Campbell N., Higuchi N. ed., Springer-Verlag New York, Inc. 328 - 341.

Hirschberg *J. (1995) Prosodic and Other Acoustic Cues to Speaking Style in Spontaneous Read Speech,* Proceedings of ICPhS '95, vol.2, 36-43.

Hirst D., Nicolas P., Espesser R. (1991) *Coding the Fo of a continuous text in French: an Experimental Approach,* Proceedings of 12th ICPhS, 5, 234-237.

House D. (1995) *The influence of silence on perceiving the preceding tonal contour,* Proceedings of the ICPhS 95, Stockholm, 123- 126.

House D., Hermes D., Beaugendre F. (1997) *Temporal-Alignment Categories of Accent-lending Rises and Falls,* Proceedings of the ESCA Eurospeech '97, 879--882.

Huber D. (1989) *A statistical approach to the segmentation and broad classification of continuous speech into phrase- sized information units,* IEEE, 600-603.

Hunt A. (1994) A *prosodic recognition module based on linear discriminant analysis,* Proceedings of the ICSLP'94, Yokohama, 1119 - 1121.

Huggins A. W. F (1972) *On the perception of temporal phenomena in speech,* J.Acoust.Soc.Am., vol.51, 9, 1279- 1290.

Imiołczyk J., Nowak I., Demenko G. (1993) *A Text-to Speech System for Polish,* Proceedings of the Eurospeech '93, vol.2, 885-889.

Imiolczyk J., Nowak L, Demenko G. (1994) *High-Intelligibility Text-to-speech Synthesis for Polish,* Arch, of Acoustics, vol. 19, 2, 161 - 172.

I.P.A. (1989) *Report on the 1989 Kiel convention,* Journal of the International Phonetic Association 19 (2), 67-80.

Isăcenko A. V., Schädlich H. J. (1966) *Untersuchungen über die deutsche Satzintonation, Akademie-Verlag,* Berlin, 7-67.

Izworski A. (1995) *Rozpoznawanie fonemów w klasyfikacji wyrazów izolowanych sieciami neuronowymi o różnych strukturach,* I Krajowa Konferencja: Głosowa Komunikacja Człowiek- Komputer, Wrocław 27-28 październik 1995, 27-32.

Izworski A., Wszołek W. (1999) *Wykorzystanie metod sztucznej inteligencji* w *diagnostyce i przetwarzaniu patologicznych sygnałów akustycznych,* in *Speech and Language Technology,* vol.3 wyd. PTFon, Poznań.

Jafari M., Wong K.H., Behbehani K., Kondraske V. (1988) *Performance characterization of human pitch control system: An acoustic approach,* J.Acoust.Soc.Am. 85, 3, 1322 - 1328.

Jassem W. (1949) *Indication of speech rhythm in the transcription of Educated Southern English,* Le matre phontique 92, 22 - 24.

Jassem W. (1952) *Intonation of Conversational English (Educated Southern British),* Wrocław.

Jassem W. (1962) *Akcent języka polskiego,* Kom. Językoznawstwa PAN, Prace językoznawcze 31, Wrocław.

Jassem W. (1973) *Podstawy fonetyki akustycznej,* PWN, Warszawa.

Jassem W. (1996a) *A quantitative analysis of standard British-English Nuclear Tones,* Journal of Quantitative Linguistics, 3, 239 - 243.

Jassem W. (1996b) *„Tłumaczenie przy pomocy programu Text-Assist" Podstawowe założenia fonetyczne i techniczne tłumaczenia różnojęzycznego w czasie rzeczywistym,* ed. Katarzyna Dobrogowska, Czesław Basztura, Wyd. PTFon, Poznań, 65 - 84.

Jassem W. (1984) *Isochrony in English speech,* in: *Intonation, Accent and Rhythm* (D. Gibbon and H. Richter, red.) de Gruyter, Berlin, 203-225.

Jassem W. (1999) *English Stress, accent and Intonation Revisited,* Speech and Language Technology, Wyd. PTFon, Poznań, 33-50.

Jassem K. (1997) *Polen a Machine Translation System Based on an Electronic Dictionary,* in *Speech and Language Technology,* vol. 1, Wyd. PTFon, Wroclaw, 161 - 194.

Jassem W., Demenko G. (1986) *On Extracting Linguistic Information from F0 traces,* w *Intonation in Discourse* (C. Johns-Lewis ed), Croom Helm, London, 1-18.

Jassem W., Demenko G. (1989) *Zależność przebiegu parametru Fo od długości frazy i dźwięczności segmentalnej,* Prace IPPT 29/1989, Warszawa.

Jassem W., Dommelen W. (1990) *Perception of Polish Accent In A Re-synthesised Speech Signal,* Archives of Acoustics XV, 3-4, 325-348.

Jassem W., Krzyśko M., Stolarski P. (1981) *Regresyjny model izochronizmu zestrojowego* w *sygnale mowy,* Prace IPPT 33/1981, Warszawa.

Jassem W., Morton J., Steffen-Batóg M. (1968) *The perception of stress in synthetic speech -like stimuli by polish listeners,* Speech Analysis and Synthesis I, 289 - 308.

Jones D. (1956) *Outline of English Phonetics,* Heffer, Cambridge.

Kacprowski J. (1965) *Zastosowanie analizy i syntezy mowy w telekomunikacji i automatyce,* Rozprawy elektrotechniczne 11, 479 - 491.

Kacprowski J., Gubrynowicz R. (1970) *Automatic recognition of Polish Vowels Using a Method of Spectrum Segmentation,* Speech Analysis and Synthesis 2, 51 - 70.

Kacprowski J., Mikiel W. (1968) *Realizacja procesu syntezy mowy za pomocą syntezatora Synfor II,* Prace IPPT, 25/1968, Warszawa.

Katae N., Matsumoto T., Kimura S. (1995) *High-quality Japanese Text-To-Speech System: Narsys,* Proceedings of the Eurospeech '95, 1861 - 1864.

Kato H., Tsuzaki M., Sagisaka Y. (1997) *Measuring temporal compensation effect un speech perception,* in *Computing Prosody,* Sagisaka, Y., Campbell, N., Higuchi, N. ed., Springer- Verlag New York, Inc., 251 - 269.

Kiessling A., Kompe R., Batliner A., Niemann H., Noth E. (1996) *Classification of Boundaries and Accents in Spotaneous Speech,* report 156, 1 -5.

Klatt D. H. (1976) *Linguistic uses of segmental duration in English: Acoustic and perceptual evidence,* J.Acoust.Soc.Am., vol. 59, 5.

Klinghardt H. (1925) *Sprachmelodie und Sprechtakt,* Elwert, Marburg.

Kohler K. J. (1983) *Prosodic Boundary Signals in German,* Phonetica 40, 89-134.

Kohler K. J. (1991) *Prosody in speech synthesis, the interplay between basic research and TTS application,* Journal of Phonetics 19, 121-138.

Kohler K. J. (1995) *Prolab - The Kiel System of prosodic labelling,* Proceedings of '95 ICPhS, vol.3, 162 -165.

Kohler K. J. (1997) *Modelling Prosody in spontaneous Speech,* in *Computing Prosody* (Sagisaka, Y., Campbell, N., Higuchi, N. ed.), Springer-Verlag New York, Inc., 187 - 209.

Kollmeier B. (1992) *Moderne Verfahren der Sprachaudiometry,* Audiologische Akustik, Killisch-Hom GmbH.

Komatsu A., Ohira E., Ichikawa A. (1986) *Prosodic Aids to structural analysis of Conversional speech,* ICASSP '86, 2283 - 2285.

Kompe R., Kiessling A., Kuhn T., Mast M., Niemann H., Noth E., Ott K., Batliner A. (1993) *Prosody takes over: a prosodically guided dialog system,* Proceedings of the 3rd European Conference on Speech and Technology, Eurospeech '93 Berlin, 2003 -2006.

Kubzdela H. (1986) *Metoda globalnego rozpoznawania wyrazów na podstawie spektrogramów binarnych,* Prace IPPT 28/1986 Warszawa.

Kubzdela H. (1997) *Automatyczna identyfikacja samogłosek w mowie ciągłej,* in *Speech and Language Technology,* vol. 1, Wyd. PTFon, Wroclaw, 23 - 35.

Kuijk D., Boves L. (1997) *Acoustic Characteristics of Lexical Stress in Continuous Speech,* IEEE 1997, 1655-1658.

Kvale K. (1993) *Segmentation and labelling of Speech,* Norwegian Institute of Technology. Lachenbruch P. A. (1975) *Discriminant analysis,* Hafner Press, London.

Ladd D. R. (1983) *Peak features and overall slope* in *Prosody: Models and Measurements, A.* Cutler and D.R. Ladd ed., Berlin, Springer-Verlag, 39 - 52.

Ladd D. R. (1996) *Intonational Phonology.* Cambridge Univ. Press, Cambridge.

Ladd D. R., Silverman K. E. A., Tolkmitt F., Bergmann G., Scherer R. (1985) *Evidence for the independent function of intonation contour type, voice quality, and F0 range in signalling speaker affect,* J.Acoust.Soc.Am.78, 2, 435-444.

Ladd D. R., Verhoven J., Jacobs K. (1994) *Influence of adjacent pitch accents on each others perceived prominence: Two contradictory effects,* Journal of Phonetics, 22, 87 - 99.

Ladefoged P. (1975) *A Course in Phonetics,* Harcourt, Bruce, Ivanovich Inc, New York.

Ladefoged P., Broadbent D. E. (1957) *Information conveyed by vowels,* Journal of the Acoust. Soc. of Am. 29, 98-104.

Ladefoged P., Draper H. M. and Whitteridge D. (1958) *Syllables and stress,* Miscellanea Phonetica III/, 1-14.

Lea W. A. (1979). *Prosodic aids to speech recognition,* in *Trends in Speech Recognition,* Lea W. A. ed. Prentice-Hall, Englewood Cliffs NJ, 166-205.

Lee T., Ching P. C., Chan L. W., Cheng Y. H., Mak B. (1995) *Tone Recognition of Isolated Cantonese Syllables,* IEEE Transactions On Speech And Audio Processing, vol.3, 204 - 209.

Lehiste I. (1970) *Suprasegmentals,* M.I.T. Press, Cambridge.

Lehiste I. (1977) *Isochrony reconsidered,* Journal of Phonetics 5, 253 - 265.

Lehning M. (1996a) *Statistical methods for the automatic labelling of German prosody,* Verbmobil Report 159, 1-8.

Lehning M. (1996b) *Prosodische Etikettirung und Segmentierung deutscher Spontansprache,* Verbmobil Report 160, 1-12.

Levitt H., Rabiner L. R. (1971) *Analysis of fundamental frequency contours in speech,* J.Acoust.Soc.Am. 49, 569-581.

Lieberman P. (1967) *Intonation, perception and language,* (M.I.T. Press, Cambridge).

Lieberman P., Katz W., Jongman A., Zimmerman R., Miller M. (1985) *Measures of the sentence intonation of read and spontaneous speech in American English,* J.Acoust.Soc.Am. *'ll,* 2, 657-676.

Lieberman M. Y., Pierrehumbert J. (1984) *Intonational invariance under changes in pitch range and length,* in *Language Sound Structure,* M. Aronoff, R. Oehrle ed., M.I.T. Press, Cambridge, 157-233.

Lindblom B. F. E. (1975) *Some temporal regularities in spoken Swedish, Auditory Analysis and Perception of Speech,* Academic Press London.

Lyberg B. (1981) *Some consequences of a model for segment duration based on Fo - dependence,* Journal of Phonetics 9, 1, 97 - 103.

Lyberg B. (1984) *Some fundamental frequency perturbations in a sentence context,* Journal of Phonetics 12, 307-317.

Lyons J. (1992) *Introduction to theoretical linguistics,* Cambridge Univ. Press, Cambridge, 68 - -70, 170ff, 194 ff„ passim.

Łobacz P. (1976) *Objective and subjective speech tempo in Polish, Speech Analysis and Synthesis,* vol. 4, 173 -187.

Maeda S. (1976) *A characterisation of American English intonation,* Ph.D.diss., M.I.T., Cambridge.

Majewski W. (1994) *Automatyczne rozpoznawanie izolowanych wyrazów - współczesne tendencje,* Materiały z Seminarium PTFon - Poznań 1993, 95- 112.

Majewski W., Blasdell R. (1969) *Influence of fundamental Frequency Cues on the Perception of some Synthetic Intonation Contours,* J.Acoust.Soc.Am. 45, 450 — 457.

Majewski W., Zalewski J. (1973) *Rola częstotliwości podstawowej w procesie percepcji syntetycznych sygnałów dźwiękowych mowy,* Prace Naukowe ITA Politechniki Wrocławskiej, 13, 37 - 50.

Mast M., Kompe R., Harbeck S., Kiessling A., Niemann H., Noth E. (1996) *Dialog act classification with the help of prosody,* Verbmobil Report 130, 1- 4.

Masters T. (1993) *Sieci neuronowe* w *praktyce,* WNT, Warszawa.

Matuszkina O. (1976) *Wpływ artykulacji spółgłoskowej na przebieg częstotliwości podstawowej w sygnale mowy polskiej,* Prace IPPT, 37/1976, Warszawa.

McAllister J. (1991) *The processing of lexically stressed syllables in read and spontaneous speech,* Language and Speech 34, 1 - 26.

McClelland J.L. Rumelhart D.E. (1987) *Explorations in Parallel Distributed Processing,* MIT Press, Cambridge.

Mikrut Z. (1993) *Rozpoznawanie ręcznie pisanych cyfr za pomocą sieci neuronowych o różnych strukturach,* Zeszyty Naukowe AGH, Automatyka, z. 66, 31 -58.

Mixdorf H., Fujisaki H. (1997) *Automated Quantitative Analysis of F0 Contours of Utterances from a German ToBI-Labeled Speech Database,* Proceedings of Eurospeech '97, 187-190.

Mobius B. (1993) *Ein quantitatives Modeli der deutschen Intonation,* Niemeyer, Tubingen.

Mobius B., Demenko G., Patzoldt M. (1991a) *Parametrische Beschreibung von Intonationsverlaufen* w *Beitrage zur Angewandten und experimentallen Phonetik,* Steiner, Stuttgart, 111 - 124.

Möbius B., Demenko G., Pätzoldt M. (1991b) *Parametric description of German fundamental frequency contours,* Proceedings of the 12th ICPhS, Aix-en-Provence, vol. 5, 222-225.

Moore Ch. A., Cohn J. F., Katz G. S. (1994) *Quantitative description and differentiation of fundamental frequency contours,* Computer Speech and Language 8, 385 - 404.

Morgan D. P., Scofield Ch. (1992) *Neural Networks and Speech Processing,* Kluwer Academic Publishers, Boston/Dordrecht/London.

Morlec Y., Bailly G., Auberge V. (1997) *Generating the prosody of attitudes,* Proceedings of ESCA Workshop on Intonation, Athens, 251 -254.

Morrison D. F. (1990) *Wielowymiarowa analiza statystyczna,* PWN, Warszawa.

Myślecki W. (1979) *Reguły generacji pobudzenia krtaniowego w procesie syntezy fraz mowy polskiej,* Instytut Telekomunikacji i Akustyki Politechniki Wrocławskiej, Praca doktorska.

Nakai M., Singer H., Sagisaka Y., Shimodaira H. (1997) *Accent Phrase Segmentation by Fo Clustering Using Superpositional Modelling,* in *Computing Prosody,* Sagisaka, Y., Campbell, N., Higuchi, N. ed., Springer-Verlag New York, Inc., 343 -359.

Nakatani, L. H., O'Connor, K. D. Aston C. (1981) *Prosodic aspects of American English speech rhytm,* Phonetica, 38, 84- 105.

Niemann H., Noth E., Kiessling A., Kompe R., Batliner A. (1997) *Prosodic Processing and its Use in Verbmobil,* Verbmobil Report 209, 1 - 2.

Niemann H., Noth E., Batliner A., Buckow J., Gallwitz F., Huber R., Kießling A., Kompe R., Warnke V. (1998) *Using Prosodic Cues in Spoken Dialog Systems,* Proceedings of Workshop Speech and Computer, SPECOM98 St.-Petersburg, 17- -26.

Nishinuma Y., Di Cristo A., Espesser R. (1984) *How does vowel duration affect loudness in a CV syllable?,* Speech Communication 3, 39-47.

Noth E. (1991) *Prosodische Information in der automatischen Spracherkennung,* Niemeyer, Tubingen.

Noth E., Ott K., Batliner A. (1993) *Prosody takes over: a prosodically guided dialog system,* Proceedings of the Eurospeech '93, vol. 2, 2003 - 2006.

Nowakowska W. (1977) *Rola częstotliwości podstawowej i poziomu intensywności w percepcji akcentu w mowie polskiej,* Prace IPPT 74/1977, Warszawa.

Obrębowski A. (1982) *Badania kliniczne i elektroakustyczne nad egzogenną wirylizacją narządu głosu,* rozprawa habilitacyjna. Wyd. AM. Poznań.

O'Connor J. D. and Arnold G. F. (1973) *Intonation of colloquial English,* Longamn, London

Olive J. P. (1975) *Fundamental frequency rules for the synthesis of simple declarative English sentences,* J.Acoust.Soc.Am. 57, 476 - 482.

Öhman S. E. G. (1967) *Word and sentence intonation: a quantitative model,* Royal Inst. of Technology (Stockholm), STL-QPSR Reports 2 - 3, 20 - 54.

Oppenheim A. V. (1982) *Sygnały cyfrowe,* WNT, Warszawa.

Ostendorf M., Ross K. (1997) *A Multi-level Model for Recognition of Intonation Labels, in Computing Prosody,* Sagisaka, Y., Campbell, N., Higuchi, N. ed., Springer-Verlag New York, Inc. 291 - 307.

Pagel V., Carbonell N., Laprie Y., Vaissiere J. (1995) *Spotting prosodic boundaries in continuous speech in French,* Proceedings of the ICPhS '95 308-311.

Palmer H. (1922) *English intonation with systematic exercises,* Heffer, Cambridge.

Palmer H. (1933) *A new classification of english tones,* Institute for Research and English Teaching, Tokyo.

Parris E., Carey M. (1996) *Language Independent Gender Identification,* Proc. IEEE, Atlanta GE, 685-688.

Pike K. L. (1947) *Phonemics. A Technique for Reducing Languages to Writing,* Ann Arbor.

Pierrehumbert *J.* (1979) *The perception of fundamental frequency declination,* J.Acoust.Soc.Am. 66, 363-369.

Pierrehumbert J. B. (1983) *Linguistic units for Fo synthesis,* in *Abstract of the 11 th ICPhS,* A.Cohen, M.P.R. van den Broecke ed., Foris Dordrecht, 137- 144.

Pierrehumbert J., Steele S.S. (1989) *Categories of tonal alignment in English,* Phonetica 46, 181 - 196.

Pierrehumbert-Breckenridge J. (1980) *The phonology and phonetics of English intonation.* Massachussetts Institute of Technology, PhD Diss.

Portele T. (1997) *Perceptual evidence for accent categories: preliminaries and first results,* Procedings of ESCA Workshop on Intonation, Athens, 271 -274.

Portele T., Heuft B. (1997) *Towards a prominence-based synthesis system,* Speech Communication 21, 61 -72.

Price P., Ostendorf M., Shattuck-Hufnagel S., Fong C. (1991) *The use of prosody in syntactic disambiguation,* J.Acoust.Soc.Am., vol.90, 2956-2970.

Pruszewicz A. (1992) *Foniatria kliniczna,* PZWL, Warszawa.

Pruszewicz A., Demenko G., Wika T., (1993) *Variability Analysis of Fo Parameter in the Voice of Individuals with Hearing Disturbances,* Acta Otolaryngol (Stockholm) 113, 450- -454

Pruszewicz A., Obrębowski A., Świdzinski P., Demenko G., Wika T., Wojciechowska A. (1991) *Usefulness of Acoustic Studies on the Differential Diagnostics of Organic and Functional Dysphonia,* Acta Otolaryngol, Stockholm, 111, 414-419.

Pruszewicz A., Demenko G., Richter L., Wika T. (1994) *Nowe listy artykulacyjne do badań audiometrycznych,* cz. 1 i 2, Otolaryngologia Polska, tom XLVIII nr. 1, 50 - 62.

Pruszewicz A., Demenko G., Wika T., (1999) *Zastosowanie metod fonetycznych w diagnostyce i rehabilitacji zaburzeń głosu, słuchu i mowy,* Speech and Language Technology, wyd. PTFon, vol. 3. 289 - 299

Rabiner L. R. (1989) *A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,* Proc. of the IEEE, vol. 77, 2, 257 - 286.

Rakowski A. (1971) *Pitch discrimination of the threshold of hearing,* Proceedings of the 7th International Congress on Acoustics, Budapest, vol.3, 20, 373 - 376.

Rakowski A. (1991) *Badanie słuchu absolutnego* w *Problemy współczesnej akustyki,* IPPT PAN, Warszawa.

Rakowski A. (1999) *Similarities in the Phonological Systems of Music and Language,* Speech and Language Technology, Wyd. PTFon, Poznań, 11-20.

Ramachandran R. P., Mammone R. (1995) *Modern Methods of Speech Processing,* Kluwer Academic Publishers, Boston/Dordrecht/London.

Raphael L. J. (1971) *Preceding Vowel Duration as a Cue to the Perception of the Voicing Characteristic of Word-Final Consonants in American English,* J.Acoust.Soc.Am. 51, 4, 1296- - 1303.

Reichl W., Harengel S., Wolfertstetter F., Ruske G. (1995) *Neural networks for nonlinear discriminant analysis in continuous speech recognition,* Proceedings of Eurospeech '95, 2163 - 2166.

Renowski J. (1967a) *Badania wpływu zmian częstotliwości podstawowej w mowie naturalnej na wrażenie intonacjji,* Prace Naukowe ITA Politechniki Wrocławskiej, 147, 3- 13.

Renowski J. (1967b) *Wpływ obcięcia ostatniej sylaby w zdaniu pytającym na wrażenie intonacji,* Prace Naukowe ITA Politechniki Wrocławskiej, 147, 13-23.

Renowski J. (1967c) *Wpływ zmian niektórych czynników prozodycznych na wrażenie intonacji w oparciu o badania przeprowadzone dla języka francuskiego,* Prace Naukowe ITA Politechniki Wrocławskiej, 147, 125- 133.

Richter L. (1983) *Wstępna charakterystyka izochronizmu zestrojowego w języku polskim,* Prace IPPT, 4/1983, Warszawa.

Richter L. (1987) *Modelling of the rhythmic structure of utterances in Polish,* Studia Phonetica Posnaniensia, A. Mickiewicz University Press, vol. 1, 91-125.

Rietveld A, C. M., Gussenhoven C. (1992/93) *Scaling prominence,* Proc. Dept. of Language and Speech, University of Nijmegen, vols. 16/17, 86-90.

Rietveld A. C. M., Gussenhoven C. (1985) *On the relation between pitch excursion size and prominence,* Journal of Phonetics 13, 299-308.

Rietveld T., Gussenhoven C. (1995) *Aligning pitch targets in speech synthesis: effects of syllable structure,* Journal of Phonetics 23, 375 - 385.

Reinecke J., Lehning M. (1994) *Interpolation und Glattung von Sprachgrundfrequenzverlaufen durch bandbegrenzte Funktionen,* DAGA '94 Proceedings, 988- -992.

Roach P. (1994) *Conversion between prosodic transcription systems: Standard British and ToBI,* Speech Communication 15, 91 - 99.

Rose P. (1991) *How effective are long term mean and standard deviation as normalisation parameters for tonal fundamental frequency,* Speech Communication 10, 229-247.

Ross K., Ostendorf M. (1996) *Prediction of abstract prosodic labels for speech synthesis,* Computer Speech and Language 10, 155 - 185.

Rossi M. (1978) *The perception of non-repetitive intensity glides on vowels, Journal of Phonetics* 6, 9- 18.

Rump H., Collier R. (1995) *Pitch peak height and focus,* IPO Annual Progress Report 30, 45 -

-50.

Rutkowska D., Piliński M., Rutkowski L. (1997) *Sieci neuronowe, algorytmy genetyczne i systemy rozmyte,* PWN, Warszawa.

Sagisaka Y., Campbell N., Higuchi N. (1997) *Computing Prosody, Computational Models for Processing Spontaneous Speech,* Springer -Verlag, New York.

Sanderman A., Collier R. (1996) *Prosodic rules for the implementation of phrase boundaries in synthetic speech,* J.Acoust.Soc.Am. 100, 5, 3390-3396.

Santen J. P. H. (1997a) *Segmental Duration and Speech Timing, in Computing Prosody,* Sagisaka, Y., Campbell, N., Higuchi, N. ed., Springer-Verlag New York, Inc. 226- -249.

Santen J. P. H. (1997b) *Prosodic Modelling in Text-to Speech Synthesis,* Proceedings of Eu- rospeech'97, 19-28.

Scheffers M.T. M (1981) *Automatic stylisation of Fo contours,* Proceedings of the 4th FASE Symposium, 981 -987.

Scott D. (1982) *Duration as a cue to the perception of a phrase boundary,* J.Acoust.Soc.Am. 71, 996- 1007.

Sickert K. (1983) *Automatische Spracheingabe und Sprachausgabe,* München, Verlag, Markt und Technik.

Silverman K. (1986) *F0 Segmental Cues Depend on Intonation: The Case of the Rise after Voiced Stops,* Phonetica 43, 76-91.

Silverman K., Beckman M., Pitrellli J., Ostendorf M., Price P., Pierrehumbert J., Hirschberg J. (1992) *TOBI: a standard for labeling English prosody,* Proceedings of the 2th International Conference on Spoken Language Processings 2, 867 - 870.

Skorek J. (1997) *Fonologia suprasegmentalna języków rosyjskiego i polskiego,* Wyd. WSP, Zielona Góra.

Skorupka S. (1955) *Studia nad budową akustyczną samogłosek polskich,* Speech Analysis and Synthesis, t.l -4, W. Jassem ed., Warszawa.

Sobczak W., Malina W. (1985) *Metody selekcji i redukcji informacji,* WNT, Warszawa.

Sorin Ch. (1981) *Functions, roles and treatments of intensity in speech,* Journal of Phonetics 9, 4, 359 - 374.

Stangert E. (1997) *Relating prosody to syntax: boundary signalling in Swedish,* Proceedings of Eurospeech '97, 239 - 242.

Steffen-Batóg M. (1963) *Analiza struktury przebiegu melodii polskiego języka ogólnego,* rozprawa doktorska, Poznań.

Steffen-Batogowa M. (1966) *Versuch einer strukturellen Analyse der polnischen Aussagemelodie,* Zeitschrift fur Phonetik, Sprachwissenschaft und Kommunikationsforschung 19, 397 - 440.

Steffen-Batogowa M. (1970) *The influence of intrinsic vowel pitch on the differences in the realisation of intended intervals,* Speech Analysis and Synthesis II, Warszawa, 177- 194.

Steffen-Batogowa M. (1973) *The effect of consonant articulation and intonation on fundamental frequency in consonants,* Speech Analysis and Synthesis III, Warszawa, 121 - 134.

Steffen-Batogowa M. (1987) *Tempo of speech and stress strategies of Polish utterances,* Studia Phonetica Posnaniensia, Adam Mickiewicz University Press, vol. 1, 127- 147.

Steffen-Batóg M. (1990) *Structural Index FBSG/MS As A Feature Distinguishing Varieties Of Contemporary Spoken Polish,* Studia Phonetica Posnaniensia, Adam Mickiewicz University Press, vol. 2, 25 - 133.

Steffen-Batóg M. (1996) *Struktura przebiegu melodii polskiego języka ogólnego,* Wyd. UAM. Poznań.

Steffen-Batóg M., Jassem W., Gruszka-Kościelak H. (1970) *Statistical distribution of shortterm F0 values as a personal voice characteristic,* Speech Analysis and Synthesis II, 195 - -206.

Steffen-Batogowa M., Katulska K. (1984) *Individual differences in the Perception of Main Stress Group Boundaries in Polish,* Lingua Posnaniensis XXVII, 101 - 115.

Streefkerk B. (1997) *Acoustical correlates of Prominence: a design for research,* Proceedings 21, University of Amsterdam, 131 - 142.

Streefkerk B. M., Pols L. C. W. (1996) *Prominent accent and pitch movements,* IFA Proceedings 20, University of Amsterdam, 111 - 119.

Streeter L. A. (1978) *Acoustic determinants of phrase boundary perception,* J.Acoust.Soc.Am. 64, 6, 1582 -1592.

Sundberg J. (1979) *Maximum speed of pitch changes in singers and untrained subjects,* Journal of Phonetics 7, 71 -79.

Swerts M. (1997) *Prosodic features at discourse boundaries of different strength,* J.Acoust.Soc.Am. 101, 1, 514-520.

Swerts M., Don G., Collier, R. (1994) *Melodic cues to the perceived finality of utterances,* J.Acoust.Soc.Am. 96, 4, 2064 - 2075.

Świdziński P. (1999) *Przydatność analizy akustycznej w diagnostyce zaburzeń głosu,* Poznań, Wyd. Akademii Medycznej.

Tadeusiewicz R. (1988) *Sygnał mowy,* Wydawnictwo Komunikacji i Łączności, Warszawa.

Tadeusiewicz R. (1993) *Sieci neuronowe,* Akademicka Oficyna Wydawnicza, Warszawa.

Tadeusiewicz R. (1994) *Zastosowanie sieci neuronowych do rozpoznawania mowy,* Materiały z Seminarium PTFon, Poznań 1993, 137 - 150.

Tadeusiewicz R. (1998) *Elementarne wprowadzenie do techniki sieci neuronowych z przykładowymi programami,* Akademicka Oficyna Wydawnicza PLJ, Warszawa.

Tadeusiewicz R., Flasiński M. (1991) *Rozpoznawanie obrazów,* PWN, Warszawa.

Tadeusiewicz R., Mikrut Z. (1994) *Sieci neuronowe rozpoznające obrazy,* Materiały I Krajowej Konferencji Sieci Neuronowe i Ich Zastosowania, Częstochowa, 52 -71.

Tadeusiewicz R., Wszołek W., Izworski A. (1998) *Application of Neural Networks in Diagnosis of Pathological Speech,* w materiałach konferencji NC98, *International ICSC/IFAC Symposium on Neural Computation* Vienna, 23 - 25.

Takefuta Y. (1975) *Method of Acoustic Analysis of Intonation,* in *Measurement procedures in Speech, Hearing and Language,* University Park Press, 363 - 368.

Taylor P. (1993) *Automatic recognition of intonation from F0 contours using the Rise/Fall Connection model,* Proceedings of the Eurospeech '93, 789 - 792.

Taylor P. (1995) *Using neural networks to locate pitch accents,* ESCA, Eurospeech '95 Proceedings, 1345- 1348.

Taylor P., King S., Isard S., Wright H., Kowtko J. (1997) *Using intonation to constrain language models in speech recognition,* ESCA Eurospeech '97 Proceedings, 2763-2766.

Tarnowska C., Mozolewski E., Ziętek E., Richter L., Demenko G. (1997) *Perzeptiv-akustische Charakteristik von ösophagusstimme und - sprache nach plastischer Umformung des pharyngoösophagealen Abschittes der sogenannten PhOA--Plastik,* Otorhinolaryngol Nova, 7, 143- 149.

Titze I. R. (1993) *Current Topics in Voice Production Mechanisms,* Acta Otolaryngol (Stockh), 113, 421-427.

Terken J. (1993) *Synthesising natural- sounding intonation for Dutch: rules and perceptual evaluation,* Computer Speech and Language, 7, 27 - 48.

Terken J. (1997) *Variation of Accent Prominence within the Phrase: Models and Spontaneous Data,* in *Computing Prosody,* Sagisaka, Y., Campbell, N., Higuchi, N. ed.), Springer-Verlag, New York, Inc., 95 -115.

Teranishi R. (1989) *A text to Speech system having several prosody options,* Proceedings of the third Symposium on Advanced Man Machine Interface Through Spoken Language, 1989, 1 - 13.

Thorsen N. (1978) *An acoustical investigation of Danish intonation,* Journal of Phonetics, 151 - - 175.

Thorsen N. (1982) *On the Variability in F0 Patterning and the Function of Fo Timing in Languages where Pitch Cues Stress,* Phonetica 39, 302-316.

Thorsen N. (1988) *Standard Danish intonation,* Annual Report of the Institute of Phonetics (univ. of Copenhagen) AR1PVC22, 1 - 23.

Traber C. (1997) *Data-driven prosody generation using automatic learning procedures,* Recueil des Publications et Communications Externes du Departement RCP, CNET 1997, 159- 162.

Umeda N. (1982) *F0 declination is situation dependent,* Journal of Phonetics 10, 279-290.

Umeda N., Quinn, A. M. S. (1981) *Word duration as an acoustic measure of boundary perception,* Journal of Phonetics 9, 19 - 28.

Vaissiere J. (1983) *Language-independent prosodic features,* in *Prosody: models and measurements,* A. Cutler, D.R. Ladd (Hg.) Springer, Berlin, 53 - 66.

Vaissiere J. (1988) *The use of Prosodic Parameters in Automatic Speech Recognition* in *Recent Advances in Speech Understanding and Dialog Systems,* H. Niemann M. Lang, G. Sagerer ed., vol.46.

Vaissiere J. (1995) *Natural explanations for prosodic cross- languages similarities,* Proceedings of the ICPhS '95, vol. 4, 654-657.

Verhoeven J. (1994) *The discrimination of pitch movement alignment in Dutch,* Journal of Phonetics 22, 65 - 85.

Veronis J., Di Cristo P., Courtois F., Lagrue B. (1997) *A stochastic model of intonation for French text-to-speech synthesis,* Proceedings of Eurospeech '97, 2643-2646.

Volskaya N. B. (1998) *The influence of the Segmental Context on the Realisation of Rising F0 Patterns in Russian,* Proceedings of Workshop Speech and Computer, SPECOM98 St.- Petersburg, 343 - 346.

Waibel A. (1986) *Recognition of Lexical Stress in a Continuous Speech Understanding System - A Pattern Recognition Approach,* Proceedings of the ICASSP '86, 2287 - 2290.

Wang M. Q., Hirschberg J. (1992) *Automatic classification of intonational phrase boundaries,* Computer Speech and Language 6, 175- 196.

Węglarz J., Czogała E., Łęski J. (1997) *A new Fuzzy Inference System with Moving Consequents in If-then Rules. Application to Pattern Recognition,* Biulletin of the Polish Academy of Science, vol. 45, 4, 644-655.

Whalen D. H., Levitt A. G. (1995) *The universality of intrinsic Fo vowels,* Journal of Phonetics 23, 349 - 366.

Willems N., Collier R., 't Hart J. (1988) *A synthesis scheme for British English intonation,* J.Acoust.Soc.Am. 4, 84, 1250 - 1258.

Wightman C. W., Shattuck-Hufnagel S., Ostendorf M., Price P. J. (1992) *Segmental durations in the vicinity of prosodic phrase boundaries,* J.Acoust.Soc.Am. 91, 3, 1707- 1716.

Ying P. A., Ching P. C., Chan L. W. (1995) *Automatic Recognition of Cantonese Lexical Tones in Connected Speech by Multi-Layer perceptron,* ESCA Workshop Proceedings, Madrid, 2205 - 2207.

Zee E. (1978) *Duration and intensity as correlates of F0,* Journal of Phonetics 6, 213 -220.

Zitter A. D. (1992) *The perceptual salience of melodic variation: contour shape and peak height,* Journal of Phonetics 20, 181 - 188.

Żurada J. M. (1992) *Introduction to Artificial Neural systems,* West Publishing Company, St. Paul.

Żurada J., Barski M., Jędruch W. (1996) *Sztuczne sieci neuronowe,* PWN, Warszawa.

ANALYSIS OF POLISH SUPRASEGMENTALS FOR SPEECH TECHNOLOGY

Summary

The present dissertation presents problems arising in the analysis of suprasegmentals in speech, their modelling, classification, synthesis and automatic recognition. On the basis of linguistic premises related to the general theory of suprasegmentals and on empirical verification of given hypotheses at the acoustic, perceptual and structural level, a model of the Polish intonational phrase is proposed. The description of accent and intonation at the linguistic level is based on the main features of a British-English system developed essentially by O'Connor and Arnold (1973) and Jassem (1984), according to which an intonational phrase is defined in terms of a sequence of (optional) pre-ictic, (constitutive) ictic, and (optional) pos-tictic accents.

A Polish phrase includes only one ictic accent, which is the primary accent, whilst the pre- and post-ictic accents are secondary. The pre-ictic and the ictic accents are mainly determined by specific pitch relations, whilst the post-ictic accent (if any) is essentially durational. Two classes of pre-ictic accents: H (high) and L (low), and 9 classes of ictic accents: HL, ML, xL, HM, LM, MH, MM, and LHL have been distinguished, where H is High, M Medium, L Low and xL extra--Low relative to the particular speaker's average and mean-Low pitch; e. g., LH means "rising from Low to High", etc. Pre-ictic accents are also referred to as prenuclear, whilst the primary, ictic accent is nuclear. On the basis of perceptual and acoustic analyses of melodically simple and complex phrases, structurally typical for Polish, the distinctive acoustic features of the individual classes of accent were defined. In the acoustical description, pitch accent was determined by pitch variations in the successive vowels/syllables and the pitch relations between syllables. In order to ensure the discriminative validity of these features as well as to eliminate redundancies, the features were subjected to statistical verification.

In keeping with the most recent tendencies prevalent especially in the automatic analysis of suprasegmentals, a structural parametrization of the melodic units (the tunes) was performed on the basis of original measurements of the appropriate features of the speech signal, viz. fundamental frequency, vowel duration and energy. With a view to simplifying the description which could theoretically include hundreds of possible combinations of features defining, e.g., the pitch relations between the beginnings and the ends of vowels and/or syllables, an eleven--element vector of such features was postulated based on experience gained from numerous previous investigations of various languages as well as the author's own experience, having a direct relation to the perceptual detection of accent and speech melody. Two of our features define the direction of pitch movement (distinguishing, e.g., a falling intonation such as HL from a rising intonation such as LH), another distinguishes the range of pitch movement, such as LM vs. LH. Two parameters are related to the global features of fundamental frequency for the given voice (mean and minimum pitch). These features lead to a definition of the relative level of an intonation related to the frequency scale, distinguishing, e.g., tune LM from MH. Three parameters define features of the nuclear tunes distinguishing the ictus from a pre-ictic accent. Variations of vowel duration and the velocity of pitch movement in the last syllable of a tune contribute towards the detection of the end of the intonational phrase.

The basis for the classification of tunes was decided to relate to the total intonation pattern recurrent in the speech signal. Pitch movements viewed from one syllable to the next have been found to be insufficient for a unambiguous description.

Classification was performed on the following intonational structures extracted from isolated utterances which involved 9 nuclear tunes representing the most frequent types of ictus, i.e., the nuclear (primary) accent in Polish, two prenuclear tunes with pre-ictic (secondary) accents of type H and L, plus (an optional) anacrusis. The 11-element feature vector describing the structural suprasegmental information in the domain of a single vowel/syllable or a sequence of vowels/syllables

was the input to an artificial neural network. Several types of neural networks were tested: probabilistic, radialfunction, and classical-multilayer (of the multilayer perceptron type). The net architecture and the training process were developed and tested experimentally in keeping with the procedures recommended in the pertinent literature.

In the total of 3675 pitch patterns classified, in the testing set (including mono- and polysyllabic utterances) the score was on average 83% correct and for the testing set 80 % (using a classical single-layer MLP). The poorest results were obtained in the classification of unaccented initial syllables (the anacruses): only 67 % for the training set and 60% for the testing set. In general, the neural networks fared best in cases that were perceptually and acoustically most distinct, viz. the ictic LH, HL and LHL (on average 92% correct).

For connected speech, in view of the low numbers of some nuclear tunes (e.g., HM, MH and LHL), the following classification was assumed, based on perceptual experiments: H (pre-ictic High), L (pre-ictic Low), R (ictic rising), F (ictic falling), MM (ictic level), and initial unaccented (anacrusis). The modification, then, consisted in conflating the classes MH, LM and LH into class R (rising), and classes HL, HM, xL and ML into class F (falling). For testing the possibility of automatically detecting accents in continuous speech a neural network was used which had been trained on isolate polysyllabic utterances. The testing set, a sample of continuous speech, included excerpts from press reviews, each approx. 6 mins long, read by three speakers. For a total of 1027 different tunes in the testing set, the average score of correct automatic classification was between 70 and 83 percent, depending on the type of accent.

Testing neural networks on data not included in the training set (in the read texts) verified the generalization of new cases. A relatively high score of between 70 and 83 percent also confirms the basic applicability of the model developed for British English, in the analysis and description if Polish intonation.

Orienting the analyses towards practical applications made it possible to verify rules for the control of fundamental frequency in the synthesis of Polish speech which used preliminary approximations of nuclear tunes of class F (type HL and ML) by exponential-power functions. The possibility of modifying and implementing rules for modelling other nuclear and prenuclear (pre-ictic) accents, based on the results of the present study is also shown.

In view of the lack of a comprehensive treatment of the state-of-the-art in the studies of Polish suprasegmentals, a synthetic overview is presented characterising the basic universal problems related to that issue in various languages together with an evaluation of the current state of research in this area in various countries.

The application of the results of suprasegmental analysis in other academic areas than speech technology, such as medicine and linguistics is shown. On the basis of the authors own work, examples are given of applying, in audiology and speech therapy, descriptions of the temporal variations of fundamental frequency in the diagnostics of voice pathology and in the voice organs rehabilitation.

The results of a comprehensive analysis of the tunes in Polish speech may be directly used, above all, in systems of Automatic Speech Recognition and Text--to-Speech Synthesis, which are currently carried out in Poland with increasing intensity.