Ignacy Nowak

WERSJA EKSPERYMENTALNA SYSTEMU MZT (MOWY Z TEKSTU) Z ZEWNĘTRZNYM STEROWANIEM FO

14/1994

WARSZAWA 1994

ISSN 0208-5658

Praca wpłynęła do Redakcji dnia 21 grudnia 1993 r.

Na prawach rękopisu

Instytut Podstawowych Problemów Techniki PAN
Nakład 100 egz. Ark.wyd. 1,5 Ark.druk. 1,60
Oddano do drukarni w kwietniu 1994 r.

Wydawnictwo Spółdzielcze sp. z o.o.
Warszawa, ul.Jasna 1

Ignacy Nowak
Zakład Fonetyki Akustycznej IPPT PAN
Poznań

WERSJA EKSPERYMENTALNA SYSTEMU MZT (MOWY Z TEKSTU) Z ZEWNĘTRZNYM STEROWANIEM FO.[1]

Streszczenie.

Wersja systemu mowy polskiej z tekstu, przedstawiona w niniejszej pracy, utworzona jest na bazie systemu MZT opisanego w [4]. W system wbudowano dodatkowe funkcje, pozwalające na wpisywanie do edytowanego tekstu ortograficznego komend sterujących składową frazową oraz parametrami ruchów akcentowych. Umożliwia to tworzenie serii zmodyfikowanych konturów intonacyjnych w wygłaszanych przez system tekstach. Kontrolę uzyskanych efektów ułatwia ilustracja graficzna przebiegu częstotliwości podstawowej w zdaniach ostatnio "wypowiedzianych" przez system. Dana wersja systemu mowy z tekstu zaprojektowana jest jako narzędzie badawcze, wspomagające poszerzanie i udoskonalanie zbioru reguł automatycznego tworzenia konturów intonacyjnych, co pozwoli na generowanie w systemie MZT w pełni zautomatyzowanym [4] wypowiedzi o bardziej urozmaiconym i precyzyjniej ukształtowanym przebiegu częstotliwości podstawowej.

1. Wstęp

W system syntezy ciągłej mowy polskiej z tekstu ortograficznego, przedstawiony w [4], wbudowany jest algorytm automatycznego tworzenia konturów intonacyjnych dla podstawowych typów zdań. Jednakże rezultaty działania tego algorytmu nie w pełni satysfakcjonują autorów; zapewne mogą też nie spełniać oczekiwań przyszłych użytkowników danego systemu MZT. Ograniczenia w działaniu obecnej wersji algorytmu związane były z zamierzonym utrzymaniem konwencji działania syntezy w czasie rzeczywistym na przeciętnym sprzęcie komputerowym typu IBM PC. Oznaczało to między innymi:

a) rezygnację z wykorzystania w szerszym zakresie informacji

[1] Praca wykonana w ramach zlecenia IPPT nr 372.

leksykalnej, syntaktycznej i semantycznej dotyczącej wygłaszanych zdań (informacja ta byłaby zresztą potrzebna także przy bardziej precyzyjnym kształtowaniu innych cech suprasegmentalnych generowanej mowy);

b) zastąpienie w zastosowanym modelu kształtowania częstotliwości podstawowej (model wg Fujisaki, [1]), funkcji ciągłych przez dane stablicowane;

c) pominięcie w przebiegu F0 cech mikroprozodycznych.

W związku z tymi ograniczeniami, system MZT w dotychczasowej postaci oferował względnie niewielką liczbę schematów przebiegu F0. Osiągnięto wskutek tego jedynie skromny podzbiór z tego bogactwa środków wyrazu, jakie intonacja zapewnia mowie naturalnej. Przy tym skala zmian F0 dla wielu sekwencji w porównaniu z mową naturalną musiała zostać zmniejszona - w przeciwnym razie niedokładne dopasowanie konturu intonacyjnego do konkretnego zdania/frazy mogłoby dawać zaskakujące czy wręcz komiczne efekty.

Zamierzając udoskonalić i poszerzyć zbiór reguł kształtowania intonacji w systemie MZT, opartym na bazie [4], należy brać pod uwagę dwa aspekty danego problemu.

1°. Dylematem niemożliwym do rozstrzygnięcia bez analizy tekstu co najmniej na poziomie leksykalnym i syntaktycznym, jest wyznaczenie granic między grupami podmiotu i orzeczenia, a także ustalenie, które wyrazy powinny w ramach tych grup uzyskać akcent intonacyjny - nawet we względnie nieskomplikowanych zdaniach. Jako tymczasowe rozwiązanie tej kwestii przyjęto, iż granice między frazami wyznaczać będą znaki interpunkcyjne (jest to bardzo odległe przybliżenie względem stanu faktycznego), a w ramach frazy każdy wyraz "leksykalny" (tj. nie będący wyrazem posiłkowym) i mający co najmniej dwie sylaby, będzie wyróżniony akcentem intonacyjnym (oraz iloczasowym). Znacznie bardziej rozbudowany pod tym względem jest układ decyzyjny dla wyrazów jednosylabowych i ich sekwencji - tutaj duże niebezpieczeństwo popełnienia "grubego" błędu spowodowało konieczność zastosowania w programie przynajmniej kryteriów leksykalnych.

2^0. Jednakże nawet wówczas, gdy w danym zdaniu właściwie określono granice między frazami i w każdej frazie wybrano wszystkie te sylaby, z którymi związane mają być akcenty intonacyjne, poprawność utworzonego konturu intonacyjnego zależy od wyboru linii deklinacyjnych dla poszczególnych fraz oraz od parametrów ruchów intonacyjnych związanych z akcentami (oczywiście, odnosi się to do wersji sterowania F0 przedstawionej w [4], opartej na modelu Fujisaki; w alternatywnych rozwiązaniach tego problemu wystąpić może inny zestaw komponentów modelu opisującego sterowanie przebiegiem częstotliwości podstawowej).

2. Struktura konturu intonacyjnego frazy: model Fujisaki i jego modyfikacje w związku z zastosowaniem w systemie MZT.

W celu przeprowadzenia dyskusji uwarunkowań, związanych z omawianym zagadnieniem, niezbędna jest krótka prezentacja modelu sterowania intonacją we w pełni zautomatyzowanym systemie MZT, opisanym w [4]. Zastosowano tam bowiem model różniący się od dotychczas wdrożonych (np. [3] lub [8]).

Model Fujisaki (p.[1]) stosowano dotychczas raczej w algorytmach aproksymacji konturów intonacyjnych wyekstrahowanych z wypowiedzi naturalnych (np. [2], [6]). Tym niemniej jego założenia predestynują go również do wykorzystania w komputerowych systemach typu tekst-mowa (ang. *Text-to-Speech*), pod warunkiem dokonania uproszczeń, pozwalających na wykonanie przez program obliczeń w tzw. czasie rzeczywistym. W przypadku syntezy mowy z tekstu, utrzymanie takiej konwencji działania oznacza, że przerwy międzyzdaniowe oraz pauzy między frazami w trakcie wygłaszania tekstu mogą jedynie symulować nabieranie oddechu przez mówiącego człowieka - w tym czasie powinna zostać przeprowadzona kompletna obróbka parametrów syntezy dla kolejnego zdania. W praktyce oznacza to zastąpienie - wszędzie, gdzie to jest możliwe - operacji zmiennoprzecinkowych przez stałoprzecinkowe, oraz wykorzystanie danych stablicowanych zamiast opisu analitycznego funkcji ciągłych.

Zasadniczymi cechami modelu Fujisaki są: a) przedstawienie

konturu intonacyjnego frazy w postaci sumy *składowej frazowej* (linii deklinacyjnej) i pewnej liczby *akcentów* (ich liczba, rozmieszczenie i struktura zależą od wielu czynników) oraz b) wybór wartości minimalnej $F0_{min}$, względem której przeprowadzane są obliczenia. Akcent realizowany jest na ogół za pomocą dwóch *ruchów akcentowych* (przeważnie pierwszy z nich oznacza wzrost, a drugi - spadek F0).

Przebieg składowej frazowej $F0_p$ związany jest z $F0_{min}$ następującą zależnością:

$$\ln F0_p(t) = \ln F0_{min} + G_p(t) \qquad (1)$$

przy czym funkcja $G_p(t)$ ma postać:

$$G_p(t) = A_p \cdot \alpha \cdot t \cdot \exp(-\alpha \cdot t) \qquad (2)$$

gdzie A_p oznacza współczynnik wzmocnienia, α - współczynnik tłumienia, t - czas.

Ruchy akcentowe (wzrosty bądź spadki F0), umieszczone na linii deklinacyjnej opisanej równaniem (1) w taki sposób, że ich początki przypadają w punktach $T_0, T_1, \ldots T_I$, dają w efekcie krzywą F0(t) dającą się opisać następująco:

$$\ln F0(t) = \ln F0_p(t) + \sum_{i=0,1,\ldots,I} \pm G_{ai}(t-T_i) \qquad (3)$$

przy czym funkcje $G_{ai}(\tau)$ mają postać:

$$G_{ai}(\tau) = A_{ai} \cdot (1-(1+\beta_i \cdot \tau) \cdot \exp(-\beta_i \cdot \tau)) \qquad (4)$$

dla $\tau \geq T_i$

oraz

$G_{ai}(\tau) = 0$

dla $0 \leq \tau < T_i$,

gdzie A_{ai} są współczynnikami wzmocnienia danego ruchu akcentowego, β_i - współczynnikami tłumienia; t,τ oznaczają czas.

Należy zwrócić uwagę na to, że funkcje G_{ai} działają od miejsca włączenia T_i do końca frazy. Ich logistyczny kształt powoduje jednak, że po względnie niedługim czasie od miejsca włączenia wartość danej funkcji G_{ai} staje się niemal stała i równa A_{ai}.

W algorytmie kształtowania przebiegu F0 dla omawianego tutaj sytemu tekst-mowa, zamiast modelu Fujisaki w ścisłej postaci anali-

tycznej, opisanej równaniami (1)-(4), zastosowano wywodzący się z niego model sekwencyjny, w którym:

1°. Każdy ruch akcentowy działa na kontur intonacyjny frazy przez ściśle określony czas; wartość F0 osiągnięta na końcu tego odcinka staje się wartością wyjściową dla następnej operacji.

2°. Na każdym odcinku konturu intonacyjnego frazy współwystępuje oddziaływanie linii deklinacyjnej z co najwyżej jednym ruchem akcentowym.

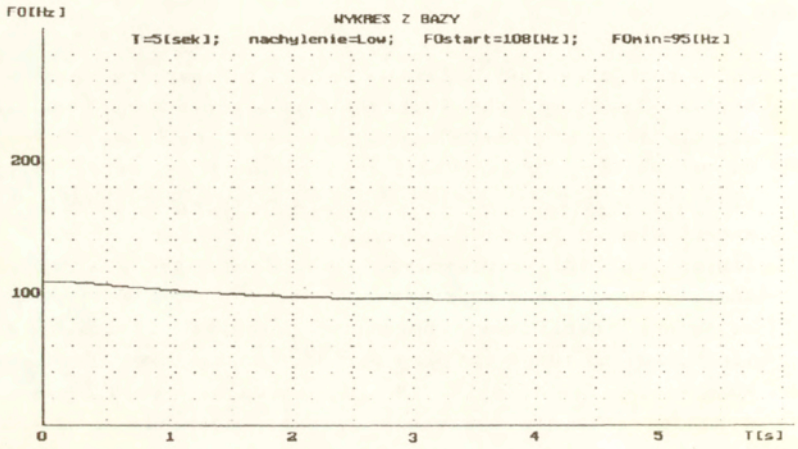
3°. Zamiast obliczania przebiegu linii deklinacyjnych ze wzorów analitycznych w samym programie syntezy, wprowadzono jako dane wejściowe stablicowane uprzednio przebiegi 3 rodzin takich linii, według tempa spadku i wartości początkowej F0 sklasyfikowane jako *Low* (niskie), *Medium* (średnie) i *High* (wysokie); w każdej rodzinie występuje 7 linii odpowiadających różnym długościom frazy: do 1.5, 1.5-2.5, 2.5-3.5, 3.5-4.5, 4.5-5.5, 5.5-6.5 oraz ponad 6.5 sekundy. Tablice przygotowano przyjmując $F0_{min}$ = 95 Hz. Łącznie program dysponuje więc 3•7=21 wzorcami linii deklinacyjnych. Przykłady wzorców ze zbioru bazowego ukazują Ryc.1-3.

4°. Stablicowano w podobny sposób przebiegi ruchów akcentowych, wybierając skalę 6,12,18,...84 Hz oraz tempo określone umownie jako *Slow* (wolne), *Fast* (szybkie) oraz *Sfast* (bardzo szybkie). Tempa dobrano w taki sposób, aby spełniony był (w przybliżeniu) warunek:

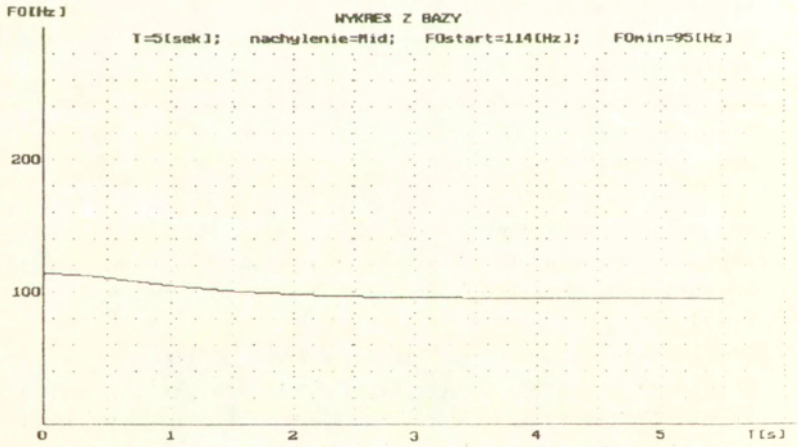
$$4 \cdot T_{sfast} = 2 \cdot T_{fast} = T_{slow} \quad (*),$$

gdzie T_x oznacza czas potrzebny na realizację ruchu o wybranej skali w tempie *x*. W ten sposób uzyskano 3•14=42 wzorce ruchów akcentowych. Przykłady wzorców ze zbioru bazowego ukazują Ryc.4-6.

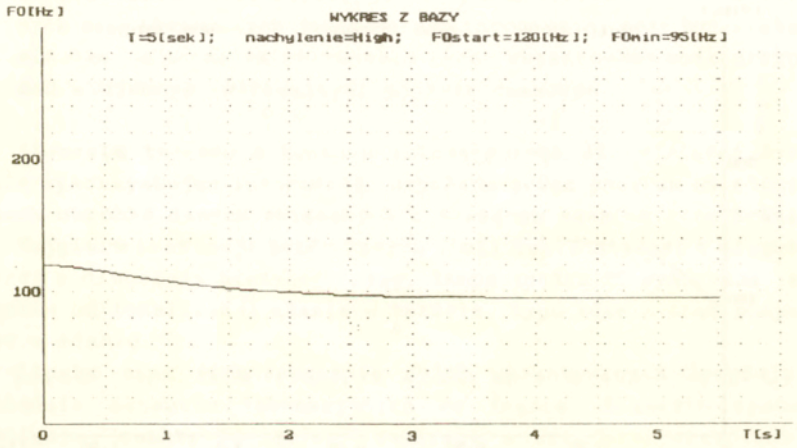
5°. Zdefiniowano dodatkowe pojęcie *wzrostu początkowego*: jest to dodatkowa wielkość, o jaką należy zwiększyć wartość początkową F0 dla danej frazy ze względu na specyficzną strukturę pierwszej grupy akcentowej w tej frazie. Szczególnie istotne okazuje się jego zastosowanie w przypadkach, gdy grupa ta jest wyjątkowo krótka, bądź przeciwnie - wiele sylab początkowych jest nieakcentowanych.



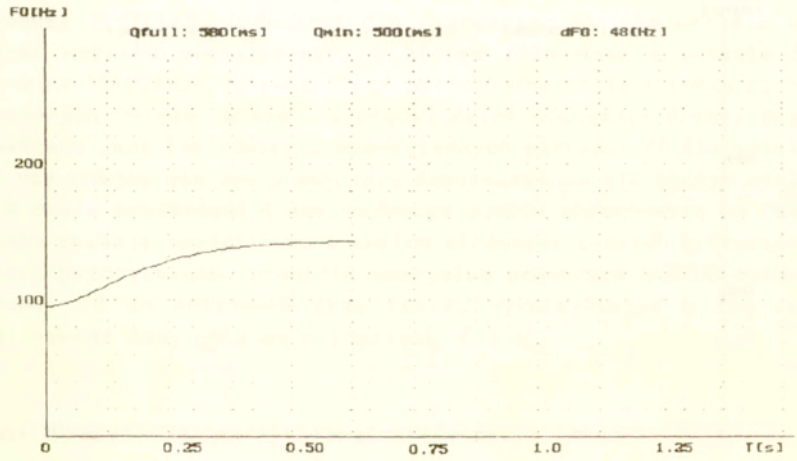
Ryc.1 Wzorzec linii deklinacyjnej: fraza 5 sek, nachylenie *Low*.



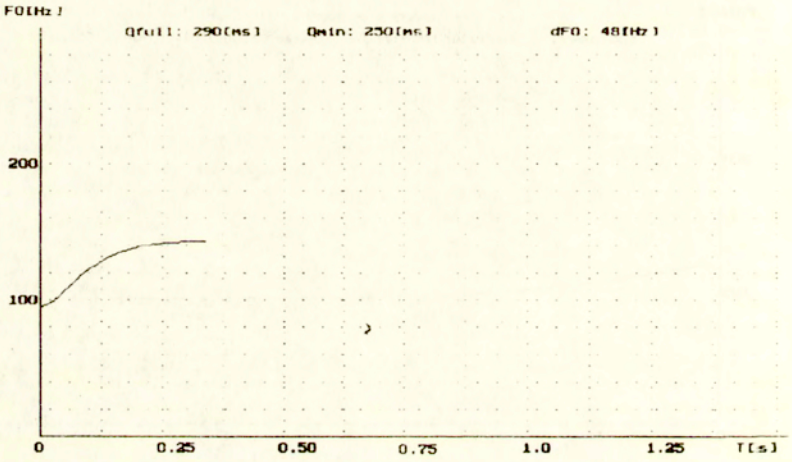
Ryc.2 Wzorzec linii deklinacyjnej: fraza 5 sek, nachylenie *Mid*.



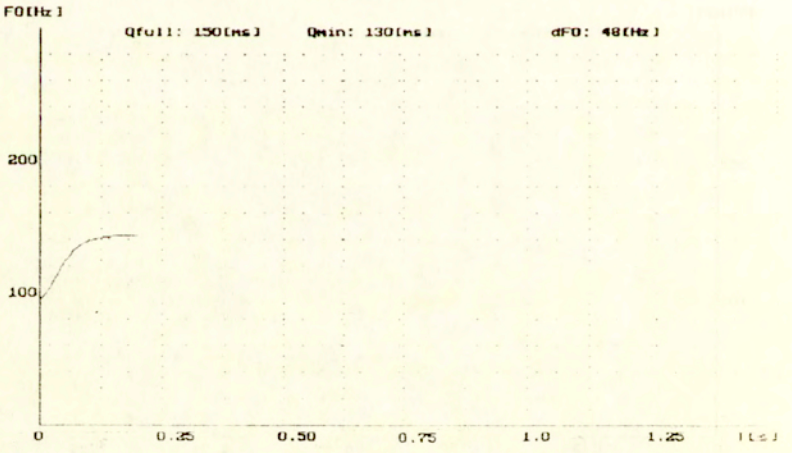
Ryc.3 Wzorzec linii deklinacyjnej: fraza 5 sek, nachylenie *High*.



Ryc.4 Wzorzec ruchu akcentowego: zakres 8 (48Hz), tempo *Slow*.



Ryc.5 Wzorzec ruchu akcentowego: zakres 8 (48 Hz); tempo *Fast*.



Ryc.6 Wzorzec ruchu akcentowego: zakres 8 (48Hz); tempo *Sfast*.

W zamian za to zrezygnowano z rozwiązania przyjmowanego w zastosowaniach aproksymacyjnych, gdzie liczba ruchów akcentowych współtworzących jeden akcent intonacyjny może być większa niż dwa, a początek pierwszego ruchu akcentowego może przypadać w ujemnym (wirtualnym) punkcie czasowym.

Algorytm tworzenia konturu intonacyjnego dla bieżącej frazy działa wykorzystując informację uzyskaną przez program na różnych etapach obróbki danych związanych z bieżącym zdaniem i tą frazą.

Najpierw następuje dobór wzorca linii deklinacyjnej o długości zgodnej z długością bieżącej frazy. Tempo spadku F0 wybierane jest zależnie od lokalizacji zdania w tekście, typu zdania oraz pozycji frazy w zdaniu.

Liczba oraz rozmieszczenie sylab akcentowanych decydują o rozkładzie akcentów intonacyjnych we frazie. Algorytm syntezy lokalizuje "punkty kontrolne" związane z kolejnymi akcentami z dokładnością do framy; punkty te dzielą frazę na odcinki iloczasowe, z których każdy może realizować co najwyżej dwa ruchy akcentowe. Wyznaczeniem kierunku, tempa oraz dopuszczalnej w danym kontekście skali ruchów akcentowych na kolejnych odcinkach zajmuje się procedura POSTULAT. Dodatkowo dla pierwszego odcinka określa ona wartość wzrostu początkowego, o którym była mowa w punkcie 5^8. Procedura ADAPTACJA sprawdza, czy postulowane skale i tempa ruchów akcentowych "mieszczą się" na odpowiednich odcinkach frazy; przeprowadzana jest też kontrola bezwzględnych wartości F0 dla ustalenia, czy znajdą się one w zakresie dopuszczalnym dla danego odcinka. W razie konieczności zakres/tempo ruchów akcentowych na danym odcinku zostają zmodyfikowane według złożonego zestawu kryteriów o różnych priorytetach. Wreszcie specjalna procedura KONTUR wpisuje wartości F0 do kolejnych fram frazy, wykorzystując w tym celu stablicowane dane opisane w punktach 3^6 i 4^6.

3. Uwarunkowania związane z zewnętrznym sterowaniem F0.

Głównym celem prac przedstawionych w niniejszej prepublikacji było stworzenie narzędzia badawczego, za pomocą którego można

będzie udoskonalić kryteria wykorzystywane przez:
- algorytm wyboru linii deklinacyjnej,
- wymienione wyżej procedury POSTULAT oraz ADAPTACJA,
- procedury decydujące o lokalizacji punktów czasowych związanych z akcentami intonacyjnymi we frazie (lokalizacja względem początku samogłoski akcentowanej w danej grupie akcentowej).

Należy zwrócić uwagę na fakt, iż realizacja celu tak sformułowanego pozwala na dokładniejsze zbadanie prawidłowości sterowania przebiegiem F0 sklasyfikowanych w punkcie 2^0 Wstępu (zagadnienia związane z p.1^0 Wstępu powinny stać się przedmiotem oddzielnych badań).

Wersję programu syntezy realizującą wskazany cel zaprojektowano przy niezmiennym założeniu, że podstawowe funkcje w modelu Fujisaki, reprezentujące składowe konturu intonacyjnego, zostały stablicowane. Możliwe jest przy tym generowanie zdań opartych na tym samym tekście ortograficznym, lecz o konturze intonacyjnym utworzonym na jeden z następujących sposobów:
- wyznaczonym całkowicie przez zbiór reguł automatycznych,
- w pełni określonym przez zbiór komend wpisanych do tekstu ortograficznego,
- generowanym na podstawie splotu reguł automatycznych i komend.

W ramach interakcji między regułami automatycznymi i komendami wpisywanymi do tekstu zrealizowane zostały następujące założenia:

a) Jeśli w dany tekst ortograficzny nie wpisano żadnej komendy, to kontur intonacyjny wygenerowanej wypowiedzi powinien być identyczny z konturem utworzonym przez system w pełni zautomatyzowany.
b) Wpisanie komendy odnoszącej się do miejsca w tekście, które jest punktem kontrolnym ustaleń automatycznych, przesłania te ustalenia, nie tworząc nowego punktu kontrolnego. Zależnie od treści wpisanej komendy, przesłonięcie może dotyczyć wszystkich ustaleń lub tylko części z nich.
c) Wpis komendy w miejscu tekstu innym niż podano w b), tworzy

nowy punkt kontrolny. Zbiór ustaleń wpisany w komendzie, jeśli nie jest kompletny, uzupełniony zostaje ustaleniami wytworzonymi przez reguły automatyczne. W skrajnym przypadku można za pomocą komendy tylko zaznaczyć nowy punkt kontrolny, pozostawiając wybór wszystkich parametrów regułom automatycznym.

Z powyższego wyszczególnienia wynika, że w algorytmie zastosowano ogólną zasadę: wszędzie, gdzie obowiązują ustalenia wynikające z (poprawnie) wpisanych komend, mają one priorytet przed regułami automatycznymi - te ostatnie służą do "wypełnienia przestrzeni decyzyjnej". Należy przy tym zwrócić uwagę na fakt, że ustalenia reguł automatycznych zastosowanych w danym miejscu frazy zależą od "historii konturu", określonej przez zbiór wszystkich punktów kontrolnych poprzedzających punkt bieżący, niezależnie od źródła ustaleń dla tych punktów, zatem ustalenia te zależą także od wpisanych komend. Na przykład, jeśli końcowy odcinek ostatniej frazy zdania oznajmującego, na którym za pomocą reguł automatycznych realizowany jest spadek F0 do poziomu linii deklinacyjnej, poprzedzony został odcinkiem, na którym za pomocą odpowiedniej komendy zrealizowano wzrost częstotliwości podstawowej większy niż to wynika z reguł automatycznych, to dany "automatyczny" spadek końcowy będzie miał większą skalę, dostosowaną do aktualnej wartości F0.

Istotnym problemem przy projektowaniu interakcji między zbiorem reguł automatycznych i systemem komend jest stopień swobody użytkownika/operatora w kształtowaniu poszczególnych fragmentów konturu. W systemie przewidziano dwa stopnie swobody użytkownika, związane z funkcjonowaniem procedury ADAPTACJA.

Przy włączonym działaniu tej procedury wszystkie wartości w obrębie komend są korygowane w analogiczny sposób jak w algorytmie w pełni automatycznego wyznaczania przebiegu F0. Oznacza to, że algorytm:

a) kontroluje zakres zmienności częstotliwości podstawowej, sprawdzając, czy wartości F0, które byłyby uzyskane wskutek ruchów określonych w komendzie, mieszczą się w przedziale właściwym dla danego odcinka frazy (sam zakres wartości dopuszczalnych może być parametrem komendy);

b) po ewentualnych korektach, wynikających z punktu a), sprawdza możliwość realizacji zmian na bieżącym odcinku frazy.

W razie konieczności skala omawianych ruchów zostaje zmniejszona, bądź - jeśli to w danej sytuacji jest dopuszczalne - ich tempo zostaje zwiększone (zwiększenie tempa zmian skraca odcinek czasu potrzebny dla realizacji ruchu akcentowego - zob. zależność (*) na str.5).

Wyłączenie adaptacji powoduje, że program "dba" tylko o eliminację błędów, które mogłyby spowodować zawieszenie jego działania. Oznacza to na przykład, iż ruch akcentowy, którego skala bądź tempo nie są zgodne z możliwościami systemu na danym odcinku frazy, zostanie wykonany tylko częściowo - w efekcie wystąpić może różnica między planowaną i faktycznie uzyskaną wartością F0 na końcu frazy. Ogólnie, wystąpienie znaczących różnic tego rodzaju (ponad 5 Hz), jest przy wyłączonej adaptacji sygnałem "przesterowania" parametrów.

4. Komendy sterujące F0 w tekście ortograficznym - struktura, lokalizacja, wykonanie

Komendy sterujące przechodzą dwustopniową weryfikację. Najpierw sprawdzane jest ich rozmieszczenie w tekście - błędna lokalizacja danej komendy powoduje jej zignorowanie. Na drugim etapie kontroli przyjmowane są tylko takie parametry (lub grupy parametrów), które z punktu widzenia algorytmu mają sens. Wszelkie parametry spoza dopuszczalnego zakresu wartości, bądź występujące w grupach niekompletnych (podział parametrów na grupy w komendzie akcentowej będzie przedstawiony poniżej), są zastępowane ustaleniami "automatycznymi".

Dokładną strukturę komend sterujących przedstawiono poniżej w postaci definicji, dotyczących kolejno komendy frazowej i komendy akcentowej. W definicjach tych (po prawej stronie znaku równości):

- ujęcie wyrażenia w nawiasy kwadratowe oznacza, że jego występowanie w danym ciągu napisów jest opcjonalne;

- symbole pisane kursywą reprezentują liczby, natomiast pozostałe symbole reprezentują same siebie.

/KOMENDA FRAZOWA/ = @[S*s*][W*w*]@ (i)

przy czym:

S sygnalizuje, że następująca po nim liczba oznacza stromość linii deklinacyjnej;

s jest liczbą o wartości ze zbioru { 0,1,2 }, identyfikującą stromość (nachylenie) linii deklinacyjnej. 0 odpowiada linii *Low*, 1 - *Medium*, 2 - *High*;

W sygnalizuje, że następująca po nim liczba oznacza wzrost początkowy;

w jest liczbą o wartości ze zbioru { 0,1,2,...,9 }. Określa wielkość *wzrostu początkowego* w jednostkach 6-hercowych; wartość 0 oznacza brak wzrostu początkowego, 1 - wzrost o 6 Hz,..., 9 - wzrost o 54 Hz.

Uwaga: przypomnieć należy, że cały system kontroli przebiegu częstotliwości podstawowej wyskalowany jest w tych jednostkach, natomiast wpis wartości F0 jest oczywiście dokładny.

Komenda frazowa może zostać wpisana bezpośrednio po dowolnym wyrazie tekstu. Podane w niej (opcjonalnie) wartości parametrów stromości linii deklinacyjnej i/lub wzrostu początkowego dotyczą frazy trwającej od zakończenia frazy poprzedniej (lub od początku zdania, gdy nie było fraz poprzedzających w danym zdaniu) do miejsca wpisania komendy, które może stać się w ten sposób nową bieżącą granicą frazową (zob. *3.Uwarunkowania...*, punkty a)-c)). Niepoprawne składniowo jest natomiast wpisanie komendy frazowej wewnątrz wyrazu (tzn. w taki sposób, że bezpośrednio po niej i przed nią występują litery ortograficzne, nie zaś odstęp lub znak interpunkcyjny), bądź też na samym początku zdania.

Powyższe ustalenia zilustrowane są wyczerpująco w podanych na następnej stronie przykładach.

Niech zdaniem "bazowym" będzie:
Ta mała śliczna dziewczynka ma bardzo wesołą mamę.

Wówczas pojedyncze komendy frazowe o rozmaitej lokalizacji i treści, wpisane do tego zdania, są interpretowane następująco:

Ta mała śliczna dziewczynka@S2W6@ ma bardzo wesołą mamę.
/Dwie frazy, w tym pierwsza o parametrach linii deklinacyjnej podanych w komendzie: stromość *High*, start 6•6Hz=36Hz ponad wartością początkową dla wzorca tej linii; parametry linii deklinacyjnej w drugiej frazie wybrane automatycznie./

Ta mała śliczna dziewczynka@S2@ ma bardzo wesołą mamę.
/Ustalenia j.w., z tym, że w pierwszej frazie wartość wzrostu początkowego jest określana automatycznie./

Ta mała śliczna dziewczynka@W6@ ma bardzo wesołą mamę.
/J.w., z tym, że ustalenia automatyczne w pierwszej frazie dotyczą stromości linii deklinacyjnej./

Ta mała śliczna dziewczynka@@ ma bardzo wesołą mamę.
/J.w., wszystkie ustalenia dotyczące linii deklinacyjnych w obydwu frazach wykonane automatycznie./

Ta mała śliczna dziewczynka ma bardzo wesołą mamę@S2W6@.
/Tylko jedna fraza, ustalenia automatyczne przesłonięte wartościami wpisanymi w komendzie./

Ta mała śliczna dziewczynka ma bardzo wesołą mamę@@.
/Komenda formalnie poprawna, lecz bez znaczenia - zarówno podział automatyczny na frazy (1 fraza), jak i ustalenia parametrów linii deklinacyjnej nie zostały zmodyfikowane./

Ta mała śliczna dziewczynka@S3W10@ ma bardzo wesołą mamę.
/Dwie frazy, dobór wszystkich ustaleń automatyczny, gdyż wyspecyfikowane wartości S oraz W przekraczają dopuszczalny zakres./

@S2W6@Ta mała śliczna dziewczynka ma bardzo wesołą mamę.
/Komenda zignorowana, gdyż nie "odcina" żadnej frazy./

Ta mała śliczna dziewczynka@S2W6@ma bardzo wesołą mamę.
/Komenda zignorowana jako umieszczona wewnątrz wyrazu (brak odstępu po komendzie frazowej) - jest tak oczywiście tylko z punktu widzenia algorytmu, który nie ma wbudowanego słownika i w związku z tym oczekuje odstępu lub znaku interpunkcyjnego jako separatora wyrazów./

/KOMENDA AKCENTOWA/ = #[T*t*][S*s1*D*d1*S*s2*D*d2*][P*p*F*f*]# (ii)

przy czym:

T sygnalizuje, że następująca po nim liczba określa położenie punktu kontrolnego związanego z bieżącym akcentem intonacyjnym;

t jest liczbą o wartości ze zbioru { 0,±1,...,±14 }, określającą wyrażone we framach przemieszczenie punktu kontrolnego względem początku samogłoski następującej po danej komendzie;

S i D oznaczają, że następujące po nich liczby określają odpowiednio tempo i skalę ruchu akcentowego;

s1,s2 są liczbami o wartościach ze zbioru { 0,1,2 }, określającymi tempo ruchów akcentowych według klucza: 0-*Slow*, 1-*Fast*, 2-*Sfast*;

d1,d2 są liczbami o wartościach ze zbioru {0,±1,...,±14}, określającymi skalę ruchów akcentowych w jednostkach 6-hercowych (zob. Uwaga do definicji komendy frazowej). Wartość 0 oznacza wyłączenie danego ruchu akcentowego, wartość dodatnia - wzrost, ujemna - spadek F0;

P i F oznaczają, że następujące po nich liczby określają odpowiednio dolną i górną granicę zmienności F0 względem linii deklinacyjnej na bieżącym odcinku frazy;

p jest liczbą o wartości ze zbioru { 0,1,...,4 }, określającą w jednostkach 6-hercowych dolną granicę zmienności F0 .

f jest liczbą o wartości ze zbioru { 5,...,14 }, określającą w jednostkach 6-hercowych górną granicę zmienności F0 .

Komenda akcentowa może być umieszczona w tekście ortograficznym jedynie bezpośrednio przed dowolną samogłoską, albo bezpośrednio po ostatnim wyrazie frazy. Inna lokalizacja komendy powoduje jej zignorowanie. Ustalenia komendy dotyczą odcinka: od poprzedniego punktu kontrolnego związanego z ruchami akcentowymi (punkt ten może być wyznaczony automatycznie lub przez poprzednią komendę akcentową w danej frazie), bądź też od początku frazy - do punktu określonego parametrem T w bieżącej komendzie (jeśli parametr ten nie jest wpisany, obowiązują ustalenia automatyczne).

Jak wynika z definicji, w ramach komendy akcentowej wyróżniono trzy grupy parametrów. Pierwszą - jednoparametrową - tworzy lokalizacja na osi czasowej. Drugą - cztery parametry, określające tempo i skalę pary ruchów akcentowych na bieżącym odcinku frazy (wyłączenie jednego lub obydwu ruchów odbywa się w sposób jawny, przez podanie skali o wartości 0). W trzeciej grupie definiuje się zakres dopuszczalnej zmienności F0. Jeśli procedura ADAPTACJA jest wyłączona, to parametry z trzeciej grupy nie wpływają na przebieg F0 na odcinku zdefiniowanym przez daną komendę akcentową, bez względu na to, czy wyspecyfikowano je jawnie, czy też pozostawiono do ustalenia regułom automatycznym, o ile tylko ruchy akcentowe zostały w tej komendzie zdefiniowane *explicite*.

Zgodnie z opisem podanym w części *3.Uwarunkowania...*, należy podawać kompletne grupy parametrów. W przeciwnym razie, dla danej grupy parametrów zostaną przyjęte ustalenia automatyczne, a częściowy wpis będzie zignorowany. Taki sam skutek ma podanie błędnej wartości któregokolwiek parametru w danej grupie.

Podobnie jak w przypadku komend frazowych, seria przykładów opartych na tym samym zdaniu bazowym, w którym zaznaczono samogłoski w sylabach akcentowanych wyznaczonych metodą automatyczną, zilustruje najlepiej podane powyżej reguły.

Ta mała śliczna dzi#T3S1D-2S1D2P3F12#ewczynka ma bardzo wesołą mamę.

/Zamiast jednego odcinka kontrolnego (od [i] w wyrazie *śliczna* do [ɨ] w wyrazie *dziewczynka*, powstały dwa odcinki, z których

pierwszy ma wszystkie parametry pochodzące z komendy akcentowej (drugi jest obsługiwany przez reguły automatyczne)./

Ta mała śliczna dzi#T3#ewczynka ma bardzo wesołą mamę.
/Układ akcentów j.w., lecz oprócz lokalizacji punktu kontrolnego (po trzech framach głoski [e]), wszystkie inne ustalenia pozostawiono regułom automatycznym./

Ta mała śliczna dzi#T3S1D0S1D2#ewczynka ma bardzo wesołą mamę.
/Układ akcentów j.w.; na rozpatrywanym odcinku wyłączono pierwszy ruch akcentowy (d1=0)./

Ta mała śliczna dziewczynka ma b#T5S1D-3S1D4#ardzo wesołą mamę.
/Na odcinku istniejącym już na skutek podziału automatycznego komenda ustaliła nową dokładną lokalizację punktu kontrolnego, skalę i tempo ruchów akcentowych./

Ta mała śliczna dzi##ewczynka ma bardzo wesołą mamę.
/Wyznaczony został dodatkowy punkt kontrolny; wybór ustaleń pozostawiono regułom automatycznym./

Ta mała śliczna dziewczynka ma bardzo wesołą mamę#T3S1D-6S1D-1#.
/Zdefiniowano nową strukturę spadku końcowego w danym zdaniu oznajmującym; wartość parametru T na końcu zdania nie jest brana pod uwagę./

Ta mała śliczna dzie#T3S1D-3S1D3P3F12#wczynka ma bardzo wesołą mamę.
/Omyłkowa lokalizacja komendy - nie przed samogłoską; komenda zostanie zignorowana./

Ta mała śliczna dzi#T3S2D-4#ewczynka ma bardzo wesołą mamę.
/Nowy punkt kontrolny; lokalizacja według parametru komendy; grupa definiująca ruchy akcentowe niekompletna - zostanie pominięta i zastąpiona ustaleniami automatycznymi./

5. Program syntezy MZT z zewnętrznym sterowaniem F0 - struktura, menu, działanie.

Program MZT z zewnętrznym sterowaniem F0 powstał na bazie systemu w pełni zautomatyzowanego, przedstawionego w [4]. Napisany został w TURBO PASCALu, v.5.5 (opis języka np. [5]).

Program wykorzystuje bez zmian jeden z modułów automatycznej transkrypcji fonematycznej (zob.[7]), utworzonych na podstawie pracy [9], realizujący północno-wschodnią odmianę polszczyzny.

Moduł edytora został zmodyfikowany tylko w nieznacznym stopniu, tak by umożliwić wpisywanie komend sterujących intonacją do tekstu ortograficznego. Ze względu na eksperymentalny charakter programu, nie ma konieczności redagowania w edytorze długich tekstów. W związku z tym zmniejszono bufor tekstu do około dwóch stron ekranowych.

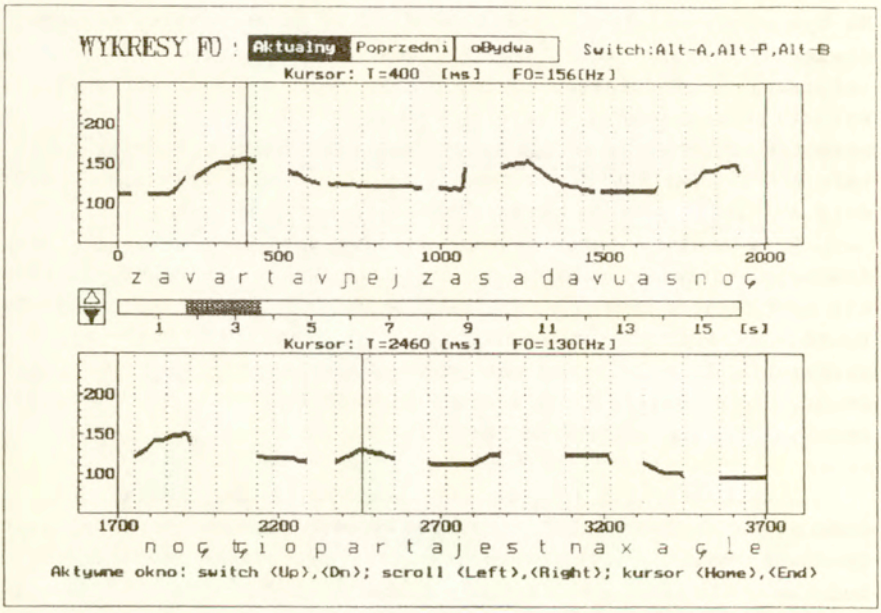
Pewnym udogodnieniem działania programu jest wprowadzenie rozwijalnego menu wspomagającego jego obsługę. "Naturalna" struktura tego menu nie wymaga szczegółowych objaśnień; wystarczy informacja, że dotarcie do głównych opcji możliwe jest po naciśnięciu klawisza <F8>, wyborze klawiszami strzałkowymi odpowiedniej grupy operacji i naciśnięciu klawisza <Enter>. Możliwy jest przyspieszony wybór głównych opcji za pomocą naciśnięcia kombinacji klawisza <Alt> i wyróżnionej litery w nazwie danej opcji. Program pozwala też na wykonywanie niektórych operacji metodą "gorących kluczy" - listę czynności objętych nimi (oraz listę operacji edycyjnych przy pisaniu tekstu) można obejrzeć po naciśnięciu klawisza <F1> (Help).

Największe zmiany nastąpiły w module syntezy. Dokonano gruntownej przebudowy systemu procedur realizujących sterowanie intonacją. Przetwarzanie informacji związanej z kształtowaniem konturu intonacyjnego dla kolejnej frazy przebiega najpierw dwutorowo: jedna z linii działania obejmuje przygotowanie danych związanych z komendami sterującymi; druga polega na utworzeniu analogicznego zbioru danych związanego z regułami automatycznymi. Obydwie linie realizowane są - w celu zminimalizowania czasu działania systemu - w toku innych czynności związanych z obróbką parametrów dla danej frazy. Operacje te kontynuowane są aż do momentu przygotowania bufora zawierającego komplet parametrów dla całej bieżącej frazy.

Na tym etapie działania systemu wartość F0 na wszystkich odcinkach dźwięcznych frazy jest stała. W momencie tym następuje połączenie informacji pochodzącej z obydwu linii, wybór linii deklinacyjnej, wzrostu początkowego i kolejnych ruchów akcentowych. Zachowana jest przy tym trójetapowa metoda pracy algorytmu (nowe procedury POSTULAT, ADAPTACJA i KONTUR), z tym, że działanie zmodyfikowanej procedury ADAPTACJA podlega ograniczeniom opisanym w części 3.

Wpis konturu intonacyjnego realizowany jest zatem w cyklu frazowym, natomiast sama operacja wpisu przebiega również cyklicznie - od jednego punktu kontrolnego do następnego. W efekcie liczba przebiegów algorytmu dla kolejnych fram frazy (dwa przebiegi), w porównaniu do w pełni automatycznego systemu MZT, pozostała bez zmiany, co pozwoliło na utrzymanie konwencji pracy systemu w czasie rzeczywistym na komputerze IBM PC/AT 286.

Dodatkową opcją programu jest możliwość obejrzenia konturów intonacyjnych dwóch zdań ostatnio wypowiedzianych przez system. Specjalny moduł grafiki dysponuje w tym celu dwoma 16-sekundowymi buforami; dla zdania dłuższego informacja obcinana jest do początkowych 16 sekund. Ekspozycja wykresów częstotliwości podstawowej następuje w dwóch oknach graficznych. Wywołanie grafiki ilustrującej przebieg F0 odbywa się za pomocą opcji menu lub gorących kluczy, przy czym wybrać trzeba tryb prezentacji. Obydwa okna mogą być przeznaczone dla tego samego zdania (ostatnio wygłoszonego bądź poprzedniego), lub też: górne okno - dla zdania ostatnio wygłoszonego, a dolne - dla poprzedniego. Okno obsługuje odcinek 2-sekundowy, lecz przewijanie wykresu umożliwia dostęp do całego 16-sekundowego bufora zdaniowego. Wykres przewijany jest wraz z tekstem fonematycznym odpowiadającym danemu odcinkowi wypowiedzi, co znacznie ułatwia użytkownikowi orientację w strukturze konturu. W danym momencie tylko jedno z z okien jest aktywne i może być przewijane. Do przełączania aktywności okien służą klawisze strzałkowe <Up> i <Dn>, do przewijania wykresu - klawisze <Left> i <Right>. Poza tym, w aktywnym oknie można odczytać dokładną wartość F0 w wybranym punkcie za pomocą kursora sterowanego klawiszami <Home> i <End>. Ryc.7 ukazuje obraz ekranu podczas pracy modułu grafiki.



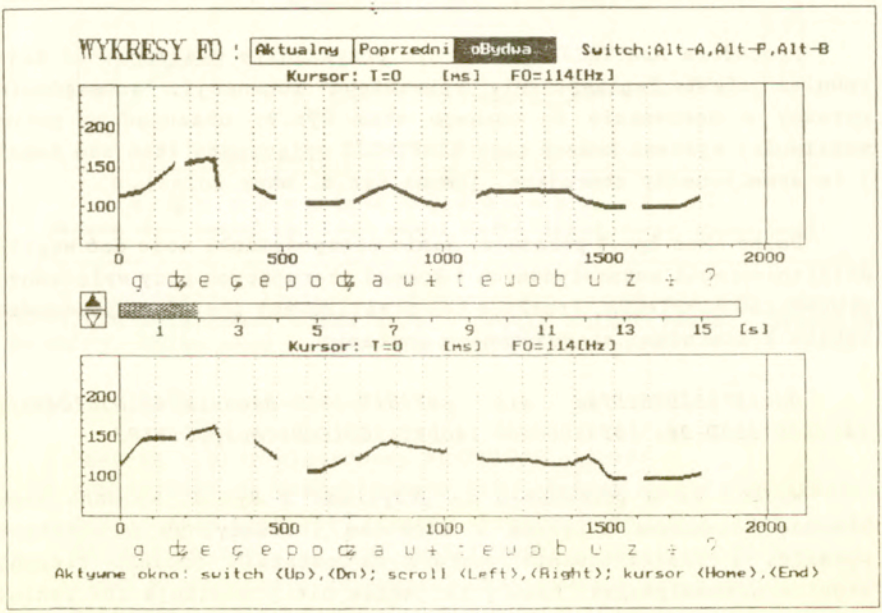
Ryc.7 Program MZT z zewnętrznym sterowaniem F0 - ogólny wygląd ekranu podczas wyświetlania konturu intonacyjnego.

Wzbogacenie możliwości systemu wydłużyło oczywiście kod programu oraz listę użytkowanych zbiorów pomocniczych (tablice linii deklinacyjnych i ruchów akcentowych). Przeprowadzona uprzednio optymalizacja kodu systemu automatycznego pozwoliła jednak w efekcie pozostać w granicach około 200 Kb pamięci operacyjnej potrzebnej dla skutecznego działania programu MZT z zewnętrznym sterowaniem F0, co dla takiego systemu wydaje się być wielkością w pełni akceptowalną.

6. Porównanie konturów intonacyjnych przykładowego zdania, wygłaszanego przez system MZT przy różnych opcjach działania.

Dla poglądowego przedstawienia zawartych w systemie możliwości sterowania kształtem konturu intonacyjnego za pomocą komend,

przedstawiono poniżej (Ryc.8 i 9) cztery przykładowe wykresy F0 pochodzące z przebiegów programu oraz ich interpretację. Wszystkie przykłady są realizacjami wypowiedzi opartej na tym samym tekście ortograficznym, którym jest pytanie: *Gdzie się podziały te łobuzy?* Daje to także okazję do pokazania typowych problemów, które mogą wyniknąć podczas pracy z wyłączoną procedurą ADAPTACJA. Celowo położono nacisk na problemy związane z komendami akcentowymi, komendy frazowe nie generują bowiem takiej rozmaitości efektów, jak komendy akcentowe.



Ryc.8 Realizacje konturu intonacyjnego przykładowego zdania. Górne okno - reguły automatyczne, dolne - komedy sterujące (p.tekst) - ADAPTACJA włączona.

Ryc. 8 przedstawia dwa przebiegi częstotliwości podstawowej podanego powyżej zdania. W górnym oknie znajduje się przebieg

wypowiedzi dla tekstu nie zawierającego komend sterujących, zatem zrealizowany wyłącznie za pomocą reguł automatycznych. Wyraźnie widoczna jest dosyć niska dynamika zmian F0, co czyni sformułowanie pytania bardzo uprzejmym i delikatnym - w sposób nieadekwatny do zawartej w nim treści, sygnalizującej irytację pytającego.

W dolnym oknie Ryc.8 widzimy przebieg prawie w całości ukształtowany za pomocą komend sterujących. Wpisany tekst wraz z komendami ma postać:

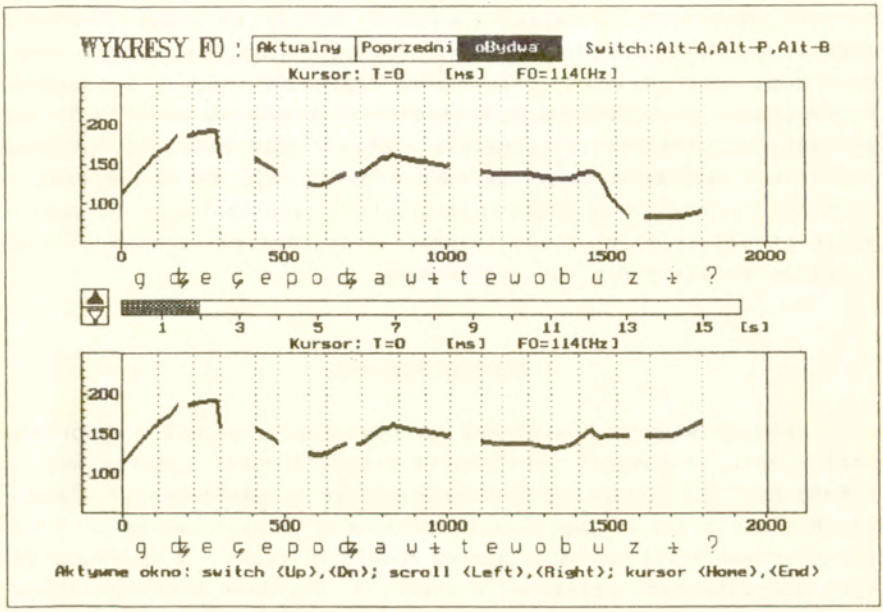
Gdzi#T5S2D7S1D7#e się p#T2S1D-3S1D-8#odzi#T4S1D3S2D4#ały t#T3S0D-2S0D-2#e ł#T3S0D0S0D-1#ob#T1S0D1S0D1#uzy#S2D-10S1D1#@S1W1@?

Procedura ADAPTACJA jest w tym przykładzie włączona, co daje również efekt "spłaszczenia" przebiegu intonacji, szczególnie wyraźny w porównaniu do górnego okna Ryc.9, ukazującego pełne możliwości systemu komend przy ADAPTACJI wyłączonej (ten sam tekst i te same komendy sterujące, jak na Ryc.8, okno dolne).

Dolne okno Ryc.9 pokazuje, jakie niespodzianki może dać współdziałanie reguł automatycznych i komend sterujących przy wyłączonej procedurze ADAPTACJA. Przebieg ten zrealizowany został na podstawie tekstu z komendami sterującymi o postaci:

Gdzi#T5S2D7S1D7#e się p#T2S1D-3S1D-8#odzi#T4S1D3S2D4#ały t#T3S0D-2S0D-2#e ł#T3S0D0S0D-1#ob#T1S0D1S0D10#uzy@S1W1@?

Widać, że w porównaniu do przykładu z Ryc.8, wpisano inną komendę akcentową związaną z akcentem intonacyjnym w ostatnim wyrazie, a ukształtowanie konturu na ostatnim odcinku zlecono regułom automatycznym. Reguły te jednak nie przewidują (na razie) tak dużej skali spadku F0 na końcu pytania, wobec czego spadek nie został wykonany w ogóle. Należy dodać, że zdefiniowane w ostatniej komendzie akcentowej ruchy S0D1S0D10 dają w efekcie jeszcze jedno poważne zniekształcenie przebiegu w porównaniu do parametrów: wzrosty kolejno o 6 i 60 Hz nie dadzą się przeprowadzić w tempie *Slow* na danym odcinku wypowiedzi, jest on dla tego celu zbyt krótki.



Ryc.9 Realizacje konturu intonacyjnego przykładowego zdania. ADAPTACJA wyłączona. Górne okno - reguły automatyczne - przebieg normalny. Dolne okno - "przesterowanie" parametrów.

Wniosek - przy wyłączonej ADAPTACJI należy:

a) kontrolować za pomocą komend cały odcinek konturu od miejsca zadziałania pierwszej komendy akcentowej do końca frazy, chyba że mamy pewność, iż mimo zmodyfikowania przebiegu F0 pozostaliśmy w przedziale zmienności akceptowanym przez reguły automatyczne;

b) unikać definiowania takich zmian F0, które ze względu na tempo lub skalę nie mogą być w pełni wykonane z powodu zbyt krótkiego odcinka kontrolnego;

c) porównywać uzyskane efekty z zaplanowanymi - kilkuhercowe odchylenia oznaczają, że parametry zostały zadane niewłaściwie ("przesterowane").

W powyższych wnioskach pominięto efekty związane z błędną lokalizacją komend w ogóle, lub ich składnią niezgodną z definicjami. Natomiast bez względu na status ADAPTACJI, należy sprawdzać, czy wpisanie do tekstu nowej komendy daje dodatkowy efekt; gdy tak nie jest, oznacza to, że popełniono błąd - znalezienie go program pozostawia użytkownikowi. Założono, że krótki okres ćwiczeń z programem pozwoli - ze względu na prostotę jego obsługi - na opanowanie struktury i lokalizacji komend oraz identyfikowanie pomyłek i błędów związanych z ich stosowaniem.

7. Wnioski końcowe.

Przedstawiony w niniejszym opracowaniu system MZT z wbudowaną możliwością sterowania przebiegiem częstotliwości podstawowej za pomocą komend, jest narzędziem oczekującym na zastosowanie w badaniach. Dotychczas za jego pomocą sprawdzono jedynie zakres możliwości sterowania intonacją za pomocą modelu Fujisaki - z upraszczającymi modyfikacjami opisanymi w części 3. Uzyskane kontury intonacyjne, z których jeden (Ryc.9, górne okno przy wyłączonej ADAPTACJI) zaprezentowano w pracy, pokazują bardzo naturalny przebieg częstotliwości podstawowej. Na opracowanie czeka jednak poszerzony zbiór reguł automatycznych, w których trzeba uwzględnić przede wszystkim: a) możliwość lokalizacji ruchów akcentowych o dużej skali i/lub bardzo szybkim tempie, b) stworzenie zasad "zapełniania luk" w sytuacji znacznych odległości między kolejnymi akcentami wyrazowymi. Drugi z wymienionych punktów wymagać będzie zapewne zmiany zbioru bazowego ruchów akcentowych, bądź też odstąpienia od zasady, iż na odcinku między dwoma kolejnymi akcentami realizowane są najwyżej dwa ruchy akcentowe. Dotychczasowe doświadczenia wskazują także, iż zbiór wzorców w bazie linii deklinacyjnych jest wystarczający dla potrzeb automatycznego systemu MZT; trzeba jednak dopracować się bardziej spójnego zbioru reguł określania wartości wzrostu początkowego - jest to, jak dotąd, najsłabiej zbadany element opisanego tu modelu sterowania intonacją.

Prace badawcze z wykorzystaniem prezentowanego programu jako narzędzia pomocniczego z pewnością przyczynią się także do jego

udoskonalenia, wyjawiając niedostatki zaprojektowanego systemu sterowania oraz stawiając nowe wymagania względem formy i zakresu dialogu z użytkownikiem. Można spodziewać się zatem powstania dodatniego sprzężenia zwrotnego między jakością mowy syntetycznej generowanej przez w pełni zautomatyzowany system MZT oraz jakością narzędzia, które do ulepszania tej mowy jest stosowane.

BIBLIOGRAFIA

[1] FUJISAKI,H., NAGASHIMA,S., *A model for the synthesis of pitch contours of connected speech*, Annual Bulletin, Engineering Research Institute, 28, 1969, Tokyo, str.53-60.

[2] FUJISAKI,H., HIROSE,K., TAKAHASHI,N., MORIKAWA,H., *Acoustic characteristics and the underlying rules of intonation of the common Japanese used by radio and television announcers*, 1986, Proceedings IEEE ICASSP, Tokyo, str.2039-2042.

[3] t'HART,J., COLLIER,R., COHEN,A., *A perceptual study of intonation. An experimental-phonetic approach to speech melody*, Cambridge University Press, Cambridge, 1990.

[4] IMIOŁCZYK,J., NOWAK,I., DEMENKO,G., *Implementacja systemu syntezy ciągłej mowy polskiej z tekstu ortograficznego wprowadzanego z klawiatury komputera typu PC - z uwzględnieniem akcentu intonacyjnego*, Prace IPPT 11/1993, Warszawa 1993.

[5] MARCINIAK,A., *Turbo Pascal 5.5*, PWN, Warszawa-Poznań 1990.

[6] MÖBIUS,B., DEMENKO,G., PÄTZOLD,M., *Parametrische Beschreibung von Intonations Konturen*, Beitrage zur angewandten und experimentallen Phonetik, Steiner, Stuttgart, 1990, str.109-125.

[7] NOWAK,I., *Automatyczna transkrypcja polszczyzny nieregionalnej (odmiana odmiana północno-wschodnia i południowo-zachodnia)*, Prace IPPT 31/1991, Warszawa 1991.

[8] de PIJPER, J.,R., *Modelling British English Intonation*, Foris Publications, Dordrecht, 1983.

[9] STEFFEN-BATOGOWA,M., *Automatyzacja transkrypcji fonematycznej tekstów polskich*, PWN, Warszawa 1975.