J. Imiołczyk, I. Nowak, G. Demenko

IMPLEMENTACJA SYSTEMU SYNTEZY CIĄGŁEJ MOWY POLSKIEJ Z TEKSTU ORTOGRAFICZNEGO WPROWADZONEGO Z KLAWIATURY KOMPUTERA TYPU PC

11/1993

WARSZAWA 1993

ISSN 0208-5658

Praca wpłynęła do Redakcji dnia 15 grudnia 1992 r.

Na prawach rękopisu

Instytut Podstawowych Problemów Techniki PAN
Nakład 100 egz. Ark.wyd. 1,5 Ark.druk. 2,0
Oddano do drukarni w marcu 1993 r.

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

Janusz Imiołczyk
Ignacy Nowak
Grażyna Demenko
Zakład Fonetyki Akustycznej IPPT PAN
Poznań

IMPLEMENTACJA SYSTEMU SYNTEZY CIĄGŁEJ MOWY POLSKIEJ Z TEKSTU ORTOGRAFICZNEGO WPROWADZANEGO Z KLAWIATURY KOMPUTERA TYPU PC - Z UWZGLĘDNIENIEM AKCENTU INTONACYJNEGO[1)]

Streszczenie

W pracy przedstawiony jest system automatycznej syntezy mowy polskiej z tekstu, opracowany na przestrzeni trzech ostatnich lat w Pracowni Analizy i Syntezy Mowy. Elementem generującym sygnał akustyczny jest wyspecjalizowany układ scalony sterowany z komputera PC AT przy użyciu oryginalnego oprogramowania, na które składają się moduły edytora, transkrypcji fonematycznej i syntezy parametrów cyfrowych. Mowa wytwarzana jest w czasie rzeczywistym i charakteryzuje się wysoką zrozumiałością, co otwiera przed układem perspektywę konkretnych zastosowań.

1. Wstęp

W okresie szybkiego postępu wiedzy o mowie, jaki dokonał się zwłaszcza na przestrzeni trzech ostatnich dziesięcioleci, zagadnienia syntezy mowy skupiły na sobie uwagę licznych ośrodków badawczych w wielu krajach, w tym głównie Stanach Zjednoczonych, Japonii, Szwecji, Francji, Wielkiej Brytanii i Niemczech Zachodnich. Jednym z najbardziej ambitnych celów w tym zakresie było opracowanie układów typu tekst-mowa (ang. *text-to-speech*, w skrócie: TTS), które dokonywałyby automatycznej konwersji tekstu ortograficznego (np. wprowadzanego z klawiatury komputera) na odpowiedni sygnał akustyczny. Od późnych lat sześćdziesiątych powstało - głównie dla odmiany amerykańskiej języka angielskiego - wiele mniej lub bardziej udanych układów tego rodzaju. Do tej drugiej grupy

---

1) Praca wykonana w ramach zlecenia IPPT nr 412

można z pewnością zaliczyć MITalk, stworzony w końcu lat siedemdziesiątych przez grupę badawczą z Massachusetts Institute of Technology [1]. Działanie tego układu oparte jest na wykorzystaniu zbioru ogromnej liczby złożonych reguł analizy tekstu i syntezy mowy, których opracowanie zajęło ponad 15 lat.

Jeden z pierwszych układów syntezy typu MZT[1)] dla języka polskiego powstał w Instytucie Informatyki Uniwersytetu Warszawskiego pod koniec lat siedemdziesiątych [9]. Ze względu na fakt, że czas realizacji syntezy był w nim jednak zbyt odległy od rzeczywistego, jego praktyczne wykorzystanie okazało się niemożliwe. Zaproponowana przez autora tzw. metoda mikrofonemowa została w późniejszym okresie zastosowana przez zespół badawczy z Instytutu Biocybernetyki i Inżynierii Biomedycznej PAN, który - w latach osiemdziesiątych - opracował dla potrzeb osób niewidomych "mówiącą" maszynę do pisania i "mówiący" kalkulator oraz komputerowy system syntezy MZT [11]. Omawiając jakość uzyskiwanej mowy syntetycznej autorzy uznali, że "nie jest jeszcze [ona] zadowalająca" i wskazali na konieczność jej polepszenia (str. 77).

Przystępując na przełomie lat 80-tych i 90-tych do pracy nad stworzeniem opisanego poniżej systemu automatycznej syntezy MZT dla języka polskiego postawiono sobie za cel główny uzyskanie mowy **w pełni zrozumiałej**. Biorąc pod uwagę konieczność działania systemu w czasie rzeczywistym, jak i brak stosownych opracowań teoretycznych, zrezygnowano z włączenia doń modułów dokonujących analizy tekstu, co byłoby niezbędne - choć zapewne nie wystarczające - do osiągnięcia wysokiej naturalności generowanej mowy.

W prezentowanym systemie wykorzystano tzw. metodę formantową, opartą na modelu typu "źródło-filtry". Sygnał mowy jest w nim generowany przez wyspecjalizowany układ scalony ze składowych częstotliwościowych stanowiących maksima energetyczne (formanty). Ze względu na rodzaj przyjętego minimalnego elementu mowy syntezę tę można określić mianem alofonicznej: każdemu fonemowi języka polskiego odpowiada od jednego do kilku uwarunkowanych kontekstowo alofonów

---

1) Proponujemy skrót MZT (**m**owa **z** **t**ekstu) jako odpowiednik TTS

syntetycznych (wzorców), z których na podstawie odpowiednich reguł konkatenacyjnych tworzone są dłuższe wypowiedzi.

Działanie systemu, zaprogramowanego w języku TurboPascal na komputer IBM PC AT, obejmuje następujące etapy:

1. Wprowadzenie tekstu ortograficznego z klawiatury
2. Transkrypcja fonematyczna
3. Przetwarzanie wzorców (alofonów) na podstawie reguł segmentalnych (modelujących tranzjenty) oraz suprasegmentalnych (tj. dotyczących iloczasu, poziomu i przebiegu częstotliwości podstawowej) - z uwzględnieniem pozycji, zajmowanych przez odpowiadające tym wzorcom głoski w obrębie wyrazu, frazy i zdania
4. Nałożenie konturów intonacyjnych na ciągi parametrów
5. Synteza sygnału akustycznego.

W przyjętym rozwiązaniu programowym synteza realizowana jest w cyklu zdaniowym. Po zakończeniu transkrypcji fonematycznej całego tekstu wykonywane są - dla jego pierwszego zdania - operacje opisane w punktach 3 i 4. Po "wypowiedzeniu" tego zdania przez syntetyzator (pkt. 5) operacjom tym jest poddawane zdanie drugie. Cykl ten powtarza się aż do wygenerowania ostatniego zdania tekstu.

Szczegółowe informacje na temat działania systemu zostaną przedstawione w kolejnych rozdziałach pracy.

2. Podstawowe dane o syntetyzatorze

Elementem generującym akustyczny sygnał mowy jest w systemie syntezy wyspecjalizowany układ scalony, stanowiący zespół pięciu połączonych szeregowo filtrów cyfrowych o programowalnych częstotliwościach środkowych i pasmach przepustowych. Filtry mogą być pobudzane z jednego z dwu dostępnych w układzie źródeł, a mianowicie źródła impulsów periodycznych lub generatora szumu. W tym pierwszym przypadku konieczne jest określenie częstotliwości (podstawowej) drgań. Przed przesłaniem do filtrów sygnał pobudzający poddawany jest wzmocnieniu w zakresie od 0 do 22 jednostek umownych.

Sygnał wyjściowy przetwarzany jest na postać analogową przy użyciu przetwornika 11-bitowego, a jego pasmo ograniczane przez filtr dolno-przepustowy do 5 kHz (wewnętrzna częstotliwość

próbkowania wynosi 10 kHz).

W syntetyzatorze przewidziano zmienny okres aktualizacji parametrów sterujących, czyli czas trwania jednej framy (standardowej). Dostępne wartości wynoszą: 8,8 ms, 10,4 ms, 12,8 ms oraz 17,6 ms. O ostatecznym czasie trwania framy decyduje wybór jednej z czterech wartości (1, 2, 3 lub 5) dodatkowego parametru, przez które mnożona jest zadeklarowana długość framy. Dopuszczalne są wszystkie kombinacje wartości obu tych parametrów, poczynając od 8,8 ms x 1 aż do 17,6 ms x 5. Ostateczna długość framy określa tempo interpolacji liniowej pomiędzy kolejnymi zadeklarowanymi wartościami amplitudy oraz częstotliwości i szerokości pasm formantów. W przypadku framy krótkiej zmiany te zachodzą raptownie, natomiast w długiej są "rozciągnięte" w czasie, co jest przydatne na przykład w syntezie dyftongów. Bez względu na długość framy, wartości wymienionych parametrów sterujących w jej obrębie interpolowane są co 1/8 jej długości, co ma na celu wygładzenie przejść między kolejnymi framami.

Inna zasada interpolacji obowiązuje w przypadku parametru sterującego zmianami częstotliwości podstawowej: wartości tego parametru są aktualizowane co 1/8 framy standardowej. Należy zaznaczyć, że jedna z dostępnych wartości nie zawiera informacji o F0, lecz służy jako przełącznik źródła pobudzania.

3. Transkrypcja fonematyczna.

Algorytm transkrypcji fonematycznej zawarty w przedstawianym tutaj systemie oparto na pracy M. Steffen-Batogowej [18]. Weryfikacja zawartego w tej pracy zbioru reguł, przeprowadzona na znaczącym objętościowo materiale przy użyciu samodzielnego programu transkrypcji (zob. [13]), umożliwiła wykrycie i korektę pewnej, zresztą niewielkiej, liczby usterek. Istniejący zbiór reguł został w [13] uzupełniony o wersję transkrypcji dla północno-wschodniej odmiany wymowy (różnice między tymi odmianami polegają głównie na odmiennych zasadach: (a) udźwięczniania / ubezdźwięczniania głosek na granicy wyrazów oraz (b) realizacji spółgłoski nosowej [n] rozpoczynającej zbitkę spółgłoskową.

Najogólniej rzecz biorąc, transkrypcja fonematyczna polega

na przedstawieniu tekstu ortograficznego w postaci takiego umownego ciągu znaków (symboli fonemów), w którym każdemu kolejnemu znakowi odpowiada w mowie jedna głoska, względnie odpowiednia pauza.

Z formalnego punktu widzenia, transkrypcja fonematyczna jest szczególnym przypadkiem translacji. Podstawowe pojęcia dotyczące translacji zawarte są np. w [3], str. 9-14, a teoretyczne podstawy transkrypcji fonematycznej - we wspomnianej już pracy [18] (rozdz IV). W niniejszym opracowaniu pojęcia te stosowane są bez dodatkowych objaśnień.

Dla potrzeb omawianego systemu syntezy przyjęto alfabety ortograficzny (X) i fonematyczny (Y) w następującej postaci:

X : a A ą b B c C ć Ć d D e E ę f F g G h H
X : i I j J k K l L ł Ł m M n N ń o O ó Ó p
X : P r R s S ś Ś t T u U w W y z Z ź Ź ż Ż
X : - ( ) " , ; : . ! ? / <spacja> <myślnik>

Uwagi: znak myślnika reprezentowany jest przez sekwencję: spacja, łącznik, spacja, czym różni się od zwykłego łącznika. Natomiast znak "/" jest elementem sztucznym, używanym wyjątkowo - dla odróżnienia słów o jednakowej postaci ortograficznej, ale odmiennym brzmieniu (np. "zamarzać" głodem, oraz "zamar/zać" na mrozie).

Y: a b ʦ ʧ ʨ d ʣ ʤ ʥ e ɨ g ɟ x i j k c
Y: l w m n ŋ ɲ o p r s ɕ ʃ t u v ɨ ʒ ʑ
Y: - ( ) , ; : . ! ? <spacja> <myślnik>

Uwagi: zazwyczaj w tekście fonematycznym stosuje się tylko znak przerwy międzywyrazowej i znak pauzy. Znaki wymienione w trzecim wierszu alfabetu Y wnoszą jednak istotne informacje dla algorytmu syntezy mowy, dlatego prawie w komplecie zostały przeniesione z alfabetu ortograficznego. Pozostałe znaki są literami "normalnego" alfabetu fonematycznego.

Alfabet fonematyczny jest dla użytkownika "ukryty". Wewnątrz programu obydwa alfabety reprezentowane są przez umowne podzbiory znaków ASCII. Ogólnie przyjętą postać tekst fonematyczny uzyskuje po wyprowadzeniu na 9-igłową drukarkę, jeśli daje ona możliwość definiowania własnego kroju czcionek w trybie znakowym.

Operacja transkrypcji fonematycznej wykonywana jest sekwencyjnie dla wszystkich kolejnych znaków tekstu, który ma być "wypowiedziany" przez syntetyzator, tak że w momencie uruchomienia modułu syntezy w pamięci komputera znajduje się już cały tekst fonematyczny potrzebny dla kompletnej wypowiedzi.

Dla zachowania przejrzystości działania algorytmu (np. w razie konieczności dodania lub modyfikacji reguły, bądź wyróżnienia dotąd nie uwzględnionego wyjątku), zdecydowano się na "tabelaryczny" układ reguł w module transkrypcji, a nie np. drzewo kontekstów, gdzie każda wprowadzona zmiana wymaga przebudowy zbioru bazowego kontekstów.

W obydwu odmianach wymowy (północno-wschodniej i południowo-zachodniej), najdłuższy lewo- i prawostronny kontekst podstawowy mają po 9 liter, w związku z czym taką właśnie długość nadano tablicom aktualnych kontekstów: lewo- i prawostronnego, dla kolejnych liter tekstu ortograficznego, poddawanych transkrypcji. Przed transkrypcją pierwszego znaku tekstu ortograficznego lewy kontekst zawiera ciąg spacji, a prawy - następujący po pierwszym znaku fragment tego tekstu (ewentualnie uzupełniony spacjami do 9 znaków, gdy cały tekst ortograficzny jest krótszy niż 10 znaków). Po każdym kroku pętli następuje odpowiednie przesunięcie kontekstów w lewo i uzupełnienie ostatniego znaku kontekstu prawostronnego kolejną literą z przetwarzanego tekstu bądź spacją (na końcowym odcinku tekstu).

Wszystkie reguły dotyczące transkrypcji konkretnej litery według danego typu wymowy, umieszczone są razem w jednej procedurze; niektóre procedury, np. dla samogłosek i znaków przestankowych, są wspólne dla obydwu typów.

Analiza kontekstów dla danej litery polega na przeglądaniu zbioru reguł, czyli wyrażeń logicznych z właściwej procedury, do momentu natrafienia na wyrażenie rozpoznające aktualne tablice kontekstowe jako rozszerzenia lewo- i prawostronego kontekstu minimalnego odzwierciedlanego przez dane wyrażenie (często jedna reguła identyfikuje kilka minimalnych kontekstów podstawowych o podobnej strukturze). Oczywiste jest, że reguła nie odwołująca się do znaków kontekstu lewostronnego (prawostronnego), nie narzuca żadnych ograniczeń na jego postać, tzn. "z punktu widzenia" tej reguły - dany kontest może

być dowolny. Tak więc, reguły pełnią po prostu rolę filtrów.

Rozwiązanie takie umożliwia łatwe wprowadzanie wyjątków od reguł; aby para: "wyjątek - reguła ogólna" działała poprawnie, wystarczy w takiej właśnie kolejności umieścić je w procedurze, zamiast budować jedno wyrażenie znacznie bardziej złożone, które zawierałoby w sobie również opis wyjątku.

Znaleziona w ten sposób wartość funkcji transkrypcji jest dopisywana do tablicy tekstu fonematycznego.

Jeśli po przejrzeniu wszystkich reguł w danej procedurze, kontekst nie zostanie zidentyfikowany, następuje sygnalizacja błędu, po czym program przystępuje do transkrypcji następnego znaku. Jest to równoważne "wpisaniu" jako wartości transkrypcji danego znaku - słowa pustego.

Poniższy przykład, podany w wersji języka źródłowego Turbo Pascal 5.0, ilustruje strukturę krótkiej i prostej procedury transkrypcji (tu: litery ortograficznej "ą"), wspólnej dla obydwu odmian wymowy.

```
procedure tla1; { transkrypcja litery "ą" }
begin
br:=true; tr1[1]:='o'; lz1:=1;
if kp[1] in ['l',#80,'m'] then exit;

lz1:=2; tr1[2]:='m';
if kp[1] in ['p','b'] then exit;

tr1[2]:='n';
if (kp[1]='t')or((kp[1]='c')and(not(kp[2] in ['i','h'])))or
   ((kp[1]='d')and ((not(kp[2] in ['z',#123]))or((kp[2]='z')
   and(kp[3]<>'i')))) then exit;
tr1[2]:=#92;
if (kp[1]=#37)or(kp[1]+kp[2]='ci')or((kp[1]='d')and
   ((kp[2]=#123)or(kp[2]+kp[3]='zi'))) then exit;

if (kl[1]='i')and((kp[1] in [#95,#123])or
   ((kp[1] in ['s','z'])and(kp[2]='i'))) then exit;

tr1[2]:=#91;
if (kp[1] in (qo+['-','f','g','k','w',#125]))or
   (kp[1]+kp[2]='ch')or((kp[1] in ['s','z'])and(kp[2]<>'i'))
   then exit;

if (kl[1]<>'i')and((kp[1] in [#95,#123])or
   ((kp[1] in ['s','z'])and(kp[2]='i'))) then exit;

br:=false
end; { procedure tla1 }
```

Uwagi:

a) Zbiór "qo" zdefiniowany jest w programie następująco: const qo=[' ','q','.',',',':',';','!','?','(',')'], gdzie ' ' oznacza spację, a 'q' - myślnik.

b) Znaki #37,#60,#95,#123 oraz #91,#92 reprezentują w pamięci

maszyny odpowiednio ć,ł,ś,ź - z alfabetu X, oraz ŋ,ɲ - z alfabetu Y.

c) nazwy zmiennych nielokalnych:
kl,kp : array[1..9] of char (tablice kontekstów);
tr1 : array[1..2] of char (robocza tablica znakowa, do której wpisywany jest rezultat transkrypcji danej litery);
br : boolean (wartość TRUE, gdy którakolwiek reguła procedury zidentyfikowała aktualny kontekst, oraz FALSE - w przeciwnym razie);
lz1 : byte (długość słowa będącego wynikiem transkrypcji - przyjmuje w programie wartości 0,1,2).

4. Reguły segmentalne

Jak wspomniano we wstępie, każdy polski fonem reprezentowany jest w systemie przez pewną liczbę kontekstowych alofonów syntetycznych, zapewniającą uzyskanie zadowalającej jakości wytwarzanej mowy. W zależności od fonemu liczba ta wynosi od 1 do 8. Przykładowo, dla samogłosek oraz bezdźwięcznych spółgłosek trących (z wyjątkiem /χ/) wystarczające okazało się wyznaczenie jednego tylko zestawu docelowych wartości poszczególnych parametrów, natomiast dla innych fonemów, takich jak np. /d/, /r/ czy /p/ konieczne było wyróżnienie więcej niż sześciu takich zestawów (tj. więcej niż sześciu alofonów).

Ostateczny zbiór zawiera 81 alofonów syntetycznych. W przeważającej liczbie przypadków, w których fonem reprezentowany jest przez więcej niż jeden alofon, o konieczności takiego wyróżnienia decydował wpływ kontekstu następującego. Niezbędne okazało się uwzględnienie następujących typów tego kontekstu:

1. samogłoska przednia lub [j]
2. samogłoska tylna lub [w]
3. [i] lub [j]
4. [i]
5. [e]
6. [a]
7. [o]
8. [u] lub [w]
9. spółgłoska trąca lub zwarto-trąca
10. spółgłoska bezdźwięczna
11. pauza
12. wszystkie pozostałe konteksty (z wyjątkiem tych określonych *explicite*)

Zaledwie w przypadku kilku fonemów zaszła konieczność wyróżnienia alofonów ze względu na kontekst poprzedzający. Uwzględniono następujące typy tego kontekstu:

1. pauza
2. spółgłoska bezdźwięczna - w środku wypowiedzi, nie przed spółgłoską bezdźwięczną
3. spółgłoska bezdźwięczna - w środku wypowiedzi, przed spółgłoską bezdźwięczną
4. jakakolwiek głoska w środku wypowiedzi - z wyjątkiem spółgłoski bezdźwięcznej
5. samogłoska tylna lub [w] - przed pauzą
6. samogłoska przednia lub [j] - przed pauzą
7. spółgłoska bezdźwięczna - przed pauzą
8. jakakolwiek inna głoska (z wyjątkiem tych podanych *explicite*) - przed pauzą.

Obok 81 wzorców głosek do ostatecznego zbioru włączono także jeden element nie będący alofonem żadnego fonemu. Jest to krótki segment wokaliczny używany w niektórych 2-elementowych zbitkach spółgłoskowych dla uwyraźnienia głoski poprzedzającej.

Wyodrębnienie 81 alofonów kontekstowych tylko częściowo rozwiązało problem odpowiedniego wymodelowania tranzjentów. Aby usunąć wszelkie niepożądane nieciągłości parametrów na granicy sąsiadujących ze sobą głosek konieczne było opracowanie obszernego zbioru szczegółowych reguł, określających dla różnego typu połączeń głoskowych długość obszaru przejściowego oraz wartości parametrów sterujących w jego obrębie. Modelowaniu za pomocą reguł podlegały przebiegi amplitudy i częstotliwości F1, F2 i F3 oraz szerokość wstęgi F1 i parametry iloczasowe.

Jak wspomniano w opisie syntetyzatora, długość framy określa tempo interpolacji pomiędzy kolejnymi wartościami poszczególnych parametrów. Manipulując nią można więc uzyskać - w zależności od konkretnych potrzeb - tranzjenty raptowne lub, przeciwnie, powolne i płynne. Mimo iż bez wątpienia iloczas jest przede wszystkim cechą suprasegmentalną, ma on również pewien wpływ na wyrazistość głosek, zwłaszcza gdy idzie o długość obszarów przejściowych między nimi. Przyjęte w tym zakresie wartości zawierają się w granicach od 8,8 do 64 ms, przy typowej wartości równej 17,6 ms.

5. Reguły suprasegmentalne

5.1. Reguły dotyczące iloczasu i poziomu

Publikacje z zakresu iloczasu głoskowego w języku polskim są niestety nader nieliczne i obejmują jedynie zagadnienia najbardziej podstawowe. W tej sytuacji opracowanie odpowiednich reguł było samo w sobie poważnym wyzwaniem.

Jako punkt wyjścia przyjęto dane zawarte w opracowaniach Richter, dotyczące iloczasu samogłosek [5] i spółgłosek [15] w jedno- i dwusylabowych logatomach oraz struktury rytmicznej wypowiedzi w mowie polskiej [16], [17]. Każdemu z wyselekcjonowanych uprzednio syntetycznych alofonów fonemów polskich nadano pewną standardową, typową dla danej głoski długość, której odpowiednia modyfikacja na podstawie dostatecznie szczegółowych reguł miała zapewnić właściwy czas trwania tej głoski we wszystkich fonotaktycznie dopuszczalnych w języku polskim kontekstach i w dowolnych wypowiedziach. Standardowe długości głosek dobierano w taki sposób, aby były one właściwe - bez poprawek na podstawie reguł - dla możliwie największej liczby kontekstów i pozycji w obrębie wypowiedzi[1)].

Ze względu na wspomnianą już fragmentaryczność danych literaturowych na temat iloczasu, w toku pracy posiłkowano się wielokrotnie urządzeniem DSP Sonagraph firmy Kay Elemetrics, umożliwiającym analizę sygnału mowy w czasie rzeczywistym i dokonanie dostatecznie dokładnych pomiarów czasów trwania głosek. Te właśnie pomiary posłużyły w głównej mierze jako podstawa do uogólnień na temat czynników decydujących o czasowym kształcie wypowiedzi, a w dalszej kolejności - do sformułowania odpowiednich reguł.

W toku prac wyodrębniono i uwzględniono w regułach następujące czynniki określające strukturę czasową wypowiedzi:

I. Typ głoski
   1. samogłoska
   2. spółgłoska

---

1) Bezwzględne wartości iloczasów głoskowych zależą oczywiście w istotny sposób od tempa mowy. Dla celów syntezy przyjęto tempo średnie - niezbyt szybkie, niezbyt wolne.

Długości standardowe oraz minimalne poszczególnych głosek były zróżnicowane. Wśród samogłosek ustnych najdłuższe było [a], najkrótsze - [i], wśród spółgłosek najdłuższe były trące bezdźwięczne, najkrótsze zaś - [r].

II. Typ połączenia głoskowego

1.1. w obrębie wyrazu / 1.2. na granicy wyrazów

a) samogłoska + inna samogłoska

b) samogłoska + identyczna samogłoska

c) [j], [w] oraz spółgłoski miękkie ([ɕ], [ʑ], [tɕ], [dʑ], [c], [ɟ], [ɲ]) + samogłoska

d) samogłoska + [w], [j], [ɲ] lub [ŋ]

e) dźwięczne spółgłoski zwarte oraz bezdźwięczne spółgłoski trące (z wyj. miękkich) i bezdźwięczne spółgłoski zwarto-trące (z wyj. miękkich) + samogłoska

f) spółgłoska + [j]

g) spółgłoska + [w]

Od typu połączenia głoskowego zależna jest w znacznej mierze długość obszaru przejściowego między stadiami ustalonymi sąsiadujących ze sobą głosek (tzw. tranzjentu). Początkową część takiego obszaru można zaliczyć do głoski poprzedzającej, część końcową zaś - do głoski następującej. Całkowita długość obu głosek równa jest więc sumie długości stadium ustalonego i przejścia. Na dłuższe tranzjenty występują w połączeniach głosek [j] oraz [ɲ] z samogłoskami, zwłaszcza samogłoskami tylnymi [a], [o], [u], najkrótsze natomiast - w połączeniach głosek wargowych z [j] (w tym przypadku nie tylko sam obszar przejściowy, ale i całkowity czas trwania [j] jest znacznie krótszy niż w innych kontekstach).

Wpływ typu połączenia głoskowego na czasy trwania obu elementów zależny jest w pewnym stopniu od tego, czy połączenie to występuje w obrębie wyrazu (np. "gEOgrafia"), czy też na granicy wyrazów (np. "żE Ograbia").

III. Typ kontekstu spółgłoskowego występującego po samogłosce

1. dźwięczność/bezdźwięczność
2. sposób artykulacji
3. pojedyncza spółgłoska/ zbitka spółgłoskowa

Dźwięczność spółgłoski następującej po samogłosce jest czynnikiem istotnie wydłużającym czas trwania samogłoski (np.

[o] w wypowiedzi "koza" będzie przy tym samym tempie mowy dłuższe niż w wypowiedzi "kosa"). Przeciwny efekt wywiera występowanie po samogłosce zbitki spółgłoskowej (przykładowo, [a] w wypowiedzi "masz" będzie dłuższe niż w wypowiedzi "maszt"). Wpływ sposobu artykulacji spółgłoski na długość poprzedzającej ją samogłoski jest bardziej złożony: średnio, samogłoski są najkrótsze przed zwartymi, najdłuższe zaś przed [r].

IV. Typ zbitki spółgłoskowej

1. rodzaj spółgłosek występujących w zbitce
2. długość (liczba elementów) zbitki

Najogólniej można stwierdzić, że stopień redukcji (skrótu) spółgłosek w zbitce zależy od ich "naturalnej" (ang. *intrinsic*) długości (mającej swoje odzwierciedlenie w przyjętej dla każdej spółgłoski długości standardowej). Najdłuższe ze wszystkich spółgłosek spółgłoski bezdźwięczne trące (por. pkt I.2) podlegają w zbitce największemu skróceniu, natomiast najkrótsze ze spółgłosek [r] nie jest skracane w ogóle.

Pewien wpływ na stopień redukcji iloczasowej w zbitce ma również pokrewieństwo sąsiadujących ze sobą spółgłosek pod względem miejsca artykulacji i dźwięczności. Redukcja jest największa, jeśli sąsiednie spółgłoski mają to samo miejsce artykulacji i obie są dźwięczne (bezdźwięczne). Przykładowo, skrót segmentu zwarcia po [m] jest większy w przypadku [b] (obie głoski są dwuwargowe i dźwięczne; np. "uzębienie") niż w przypadku [p] (obie głoski są dwuwargowe, ale różnią się pod względem dźwięczności; np. "potępienie"). Przy braku podobieństwa w zakresie miejsca artykulacji sąsiadujących ze sobą w zbitce spółgłosek skrót tego typu nie jest realizowany (np. "uwielbienie").

Czynnikiem decydującym o stopniu redukcji iloczasowej spółgłosek w zbitce jest również długość zbitki, wyrażona liczbą składających się na nią elementów. Ogólnie można powiedzieć, że im dłuższa jest zbitka, tym krótsze są współtworzące ją spółgłoski. Przykładowo, spółgłoska [s] jest krótsza w wypowiedzi "wstrętny" niż w wypowiedzi "stąpać".

V. Pozycja/status sylaby w obrębie wypowiedzi

1. nagłos

2. anakruza
3. sylaba o akcencie wyrazowym
4. sylaba o akcencie zdaniowym
5. sylaba w wygłosie zamkniętym
6. sylaba w wygłosie otwartym
7. sylaba nie akcentowana - poza anakruzą oraz wygłosem otwartym i zamkniętym

Czas trwania głosek w dowolnej wypowiedzi zależny jest od statusu sylaby, w skład której głoski te wchodzą (np. akcent/brak akcentu) oraz od pozycji, jaką sylaba ta zajmuje w obrębie wypowiedzi. Dotyczy to w głównej mierze samogłosek, w przypadku których zróżnicowanie czasu trwania w zależności od wymienionych czynników może zbliżać się do 100 ms. Samogłoski są najkrótsze w anakruzie i sylabie nie akcentowanej (pkt 7 powyżej), najdłuższe zaś - w sylabie o akcencie zdaniowym i w wygłosie otwartym. W przypadku spółgłosek największe iloczasowe "odstępstwo" od długości standardowej zaznacza się w wygłosie zamkniętym - bezpośrednio przed pauzą.

VI. Liczba sylab w stopie rytmicznej

Czas trwania głosek zależny jest od liczby sylab współwystępujących w obrębie stopy rytmicznej. Wraz ze wzrostem liczby sylab następuje redukcja iloczasów poszczególnych głosek.

VII. Typ pauzy

Długość pauzy uzależniona jest od rodzaju "generującego" ją znaku przestankowego: np. pauza po przecinku jest krótsza niż po kropce.

Wyodrębnienie wymienionych wyżej czynników oraz określenie siły ich oddziaływania stanowiło podstawę do sformułowania szczegółowych reguł iloczasowych w formie systemu wydłużeń i skrótów. Zadanie to było niełatwe ze względu na sumowanie się wpływów poszczególnych czynników. Co prawda nie wszystkie z nich mogą się ze sobą łączyć (np. anakruza i akcent), ale w wielu przypadkach o ostatecznym czasie trwania głoski decyduje suma oddziaływań dwóch, trzech, a nawet większej liczby czynników. Dla przykładu: czas trwania samogłoski [e] w wypowiedzi "on tu jest" współokreślony jest aż przez 8 czynników: I.1, II.1.1.c, III.1, III.2, III.3, V.4, V.5, VI.

Wpływ każdego z nich powoduje odpowiednie wydłużenie lub skrót samogłoski. Zestaw reguł iloczasowych w swojej ostatecznej wersji obejmuje ok. 300 pojedynczych instrukcji modyfikujących czasy trwania głosek.

Ze względu na relatywną nieistotność zróżnicowań intensywności w percepcji mowy, reguły modyfikujące kształt obwiedni amplitudowej są bardzo nieliczne i odnoszą się - na płaszczyźnie suprasegmentalnej - wyłącznie do zjawisk zachodzących w wygłosie wypowiedzi.

5.2. Reguły kształtowania konturów intonacyjnych.

5.2.1. Modelowanie intonacji w syntezie mowy.

Mimo licznych prób podejmowanych w ostatnich latach (np. [1], [7], [10]), problem sterowania częstotliwością podstawową w syntezie mowy nie doczekał się jeszcze zadowalającego rozwiązania. Brak praktycznych rozwiązań w tej dziedzinie dla języka polskiego istotnie przyczyniał się do złej jakości dotychczasowych rozwiązań syntezy mowy. Istniejące opracowania, obejmujące swym zakresem głównie wypowiedzi izolowane, dostarczają tylko fragmentarycznych informacji o cechach intonacji mowy polskiej. W tej sytuacji, dla sformułowania reguł sterowania parametrem F0 w układzie syntezy konieczne okazało się wykorzystanie wyników prac dotyczących innych języków ([8], [12], [14]), jak również własnych doświadczeń.

Rezultaty badań własnych, mających na celu określenie zasad sterowania intonacją, wykazały możliwość implementacji funkcji aproksymujących przebiegi parametru F0 oraz zasady ich superpozycji opracowanej przez Fujisaki [6]. Model ten zakłada superpozycję składowej frazowej (określającej linię podstawową, deklinację) i składowych akcentowych (wyznaczonych dla poszczególnych grup akcentowych) oraz wartość Fmin, charakterystyczną dla danego mówcy. Grupę akcentową zdefiniowano jako prozodyczną jedność składającą się z sylaby akcentowanej oraz sąsiednich sylab nieakcentowanych [19].

W celu określenia rodzaju oraz zakresu modyfikacji modelu dla języka polskiego przeprowadzono analizę częstotliwości podstawowej w tekście gazetowym, czytanym 3-krotnie przez 6 osób. Szczegółowej ocenie poddano materiał składający się z

wypowiedzi fonetyka realizującego najbardziej konsekwentnie poszczególne 3 replikacje tekstu. Modyfikacji funkcji sterującej składową frazową oraz akcentową dokonano głównie w zakresie przedziału wartości parametrów funkcji (współczynnika wzmocnienia oraz tłumienia), jak również w sposobie ich sterowania. Wstępne wartości parametrów sterujących przyjęto na podstawie danych eksperymentalnych. Ich optymalizację przeprowadzono na podstawie doświadczeń odsłuchowych. Techniczne rozwiązanie generowania wartości parametru F0 na podstawie stabelaryzowanych danych funkcji frazowej oraz akcentowej w zasadniczy sposób ułatwiło sformułowanie reguł sterowania intonacją, jak również znacznie skróciło czas realizacji programu.

5.2.2. Zasady przygotowania informacji koniecznej do sterowania parametrem F0 w układzie syntezy.

Przed sformułowaniem reguł modelowania częstotliwości podstawowej określono informację konieczną do sterowania tym parametrem. Uznano, że program realizujący kształtowanie konturów intonacyjnych powinien uwzględniać następujące rodzaje danych:

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 sposób sterowania składową frazową.

b) Struktura frazy.

Frazy mogą posiadać odmienne struktury, wynikające z liczby oraz rozkładu sylab akcentowanych. Struktura frazy ma bezpośredni wpływ na sposób sterowania składową frazową oraz akcentową.

c) Pozycja frazy.

Frazy pierwsze i końcowe zdania odgrywają szczególnie istotną rolę w modelowaniu przebiegu częstotliwości podstawowej, decydują o dynamice zmian oraz określają typ wypowiedzi (zdanie oznajmujące, pytające o rozstrzygnięcie, o uzupełnienie).

d) Zakończenie frazy.

Przyjęto, że frazy mogą się kończyć następującymi znakami

interpunkcyjnymi: ( ) . , : ; ? ! - <myślnik>.

e) Długość frazy.

Przyjęto 7 następujących kategorii, zależnych od długości frazy wyrażonej w sekundach: 0-1.5, 1.5-2.5, 2.5-3.5, 3.5-4.5, 4.5-5.5, 5.5-6.5, powyżej 6.5.

2. Dane opisujące frazę.

a) Liczba grup akcentowych.

Liczba grup akcentowych wchodzących w skład frazy wpływa na sposób modelowania przebiegu parametru F0 w założonym przedziale zmienności Fmax - Fmin.

b) Pozycja grup akcentowych.

Szczególnie istotne znaczenie ma modelowanie przebiegu częstotliwości podstawowej w pierwszej i w ostatniej grupie akcentowej, ponieważ w sposób decydujący wpływa na percepcję intonacji całej frazy.

c) Struktura grup akcentowych.

Przeanalizowano dla języka polskiego 16 grup akcentowych, zróżnicowanych pod względem liczby oraz pozycji sylab nieakcentowanych. W szczególności dla pierwszej grupy akcentowej otrzymano następujące struktury (literą A oznaczono pozycję sylaby akcentowanej, literą N - nieakcentowanej): A, AN, ANN, ANN, NA,NAN, NANN, NANNN, NNA, NNAN,NNANN,NNANNN, NNNA, NNNAN, NNNANN, NNNANNN, dla grupy śródkowej i ostatniej: A, AN, ANN, ANNN. Na tej podstawie sformułowano ogólniejszą regułę dobierania kształtu i skali ruchu intonacyjnego ze zbioru stabelaryzowanych wykresów funkcji akcentowych.

3. Dane opisujące sylabę.

Rozróżnienie sylaby akcentowanej oraz nieakcentowanej stanowi podstawę do wyznaczenia odległości między maksimami osiąganymi na kolejnych sylabach akcentowanych.

5.2.3. Reguły sterowania składową frazową oraz składowymi akcentowymi.

1) Sterowanie składową frazową.

Funkcja sterująca frazą G_{pi} opisana jest następującą zależnością:

$$G_{pi}(t) = A_{pi}\,\alpha_i\,t\,\exp(-\alpha_i\,t)$$

gdzie A_{pi} oznacza współczynnik wzmocnienia, α_i - współczynnik

tłumienia, i- numer kolejnej frazy, t- czas.

Zależnie od długości frazy przyjęto współczynniki wzmocnienia A_{pi} w zakresie 0.018-0.633 oraz tłumienia α_i w przedziale 1.14-8.00. Ustalono zbiór maksymalnych wartości funkcji frazowych w zakresie 100-124 Hz. Wyznaczono 3 typy linii deklinacyjnej (niski, średni oraz wysoki) i w każdym z nich rozróżniono 7 konfiguracji parametrów A_{pi} oraz α_i zależnie od długości frazy. Przykładowo pierwszej frazie długiego zdania, rozpoczynającej się od sylaby akcentowanej przypisano maksymalną wartość współczynnika wzmocnienia A_{pi}.

2) Sterowanie składowymi akcentowymi.

Funkcję sterującą składową akcentową G_{aj} opisano zależnością:

$$G_{aj}(t) = A_{aj} \{ 1 - (1 + \beta_j t) \exp (-\beta_j t) \}$$

gdzie A_{aj} oznacza współczynnik wzmocnienia j-tego akcentu, β_j - współczynnik tłumienia, j- numer kolejnego akcentu, t- czas.

Rozróżniono następujący podział grup akcentowych w zależności od ich pozycji: grupa początkowa, grupy środkowe oraz grupa końcowa.

Dla pierwszej grupy akcentowej przyjęto następujące zasady:

a) Dla struktur rozpoczynających się sylabą akcentowaną:

W przypadku sylaby rozpoczynającej się spółgłoską/spółgłoskami dźwięcznymi przyjęto przebieg rosnący z maksimum przypadającym pod koniec samogłoski akcentowanej i podwyższoną wartością początkową. W przypadku samogłoski akcentowanej poprzedzonej segmentem bezdźwięcznym przyjęto przebieg lekko rosnący z wysoką wartością początkową.

b) Dla struktur rozpoczynajacych się sylabą/sylabami nieakcentowanymi:

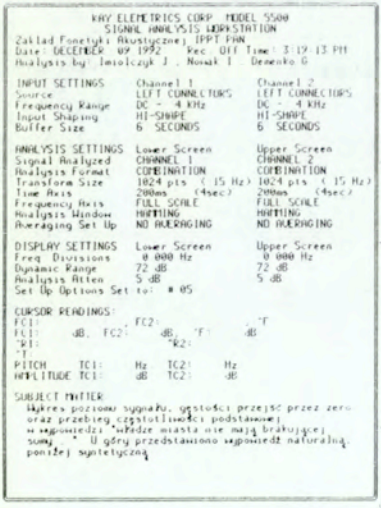
W obrębie sylaby/sylab nieakcentowanych założono małą zmienność parametru F0 (rzędu 5-10 Hz). Wzrost częstotliwości podstawowej następuje w obrębie sylaby akcentowanej i zależnie od liczby poprzedzających ją sylab nieakcentowanych ma mniejszy lub większy zakres (zwykle 40-50 Hz).

Przebieg częstotliwości podstawowej w środkowych grupach akcentowych uwarunkowany jest pozycją grupy (im bliżej początku wypowiedzi znajduje się grupa, tym większą ma dynamikę zmian) oraz ilością sylab nieakcentowanych pomiędzy sąsiadującymi

akcentowanymi (w przypadku tylko 1 sylaby nieakcentowanej lub ich braku następna grupa akcentowa nie jest wyraźnie intonacyjnie zaznaczona). Jako ogólną zasadę przyjęto, że poziom na sylabach akcentowanych zawsze musi być wyższy niż na sąsiadujących sylabach nieakcentowanych. Zakres zmienności parametru F0 nie może być większy niż w grupie akcentowej pierwszej lub ostatniej.

Reguły dotyczące końcowej grupy akcentowej zależą w sposób istotny od znaku interpunkcyjnego umieszczonego na końcu frazy. Dla fraz stanowiących koniec zdania i zakończonych kropką przyjęto najczęstszy dla języka polskiego schemat spadku częstotliwości podstawowej, rozpoczynający się od ostatniej sylaby akcentowanej i - zależnie od liczby sylab nieakcentowanych występujących między przedostatnią i ostatnią sylabą akcentowaną - bardziej lub mniej stromy. Założony zakres zmienności wynosi 35-45 Hz. Dla zdań pytających założono odrębne schematy sterowania parametrem F0; zależnie od rodzaju pytania przyjęto wzrost częstotliwości w zakresie 50-80 Hz lub niewielki wzrost kontynuacyjny (10-20 Hz.).

Optymalizacji wartości sterujących dokonano na podstawie doświadczeń odsłuchowych oraz dodatkowych analiz akustycznych.

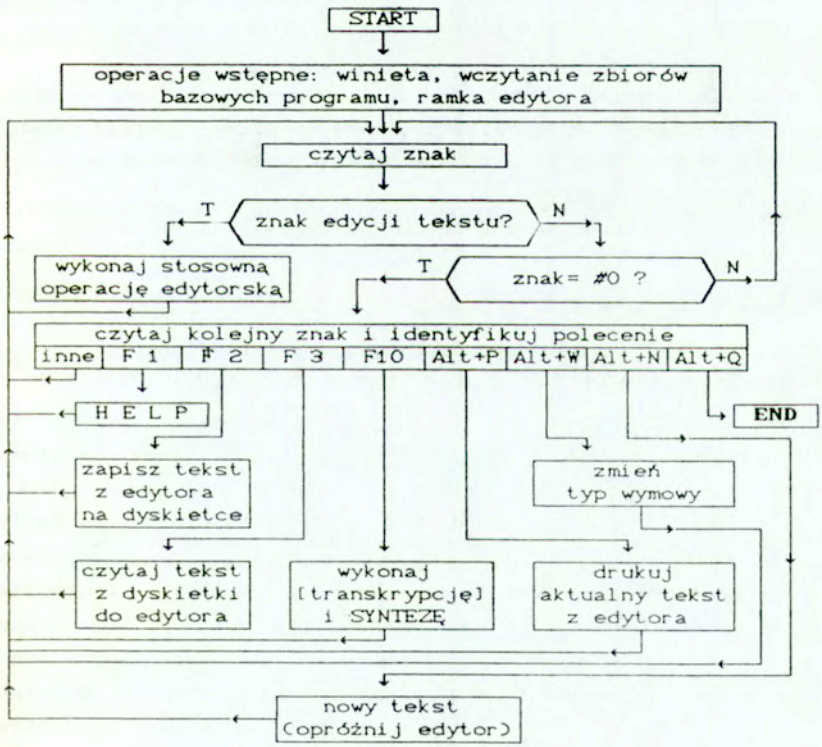


Zamieszczony na stronie 20 rysunek przedstawia postać czasową naturalnej (a) i syntetycznej (b) wypowiedzi: "Władze miasta nie mają brakującej sumy,.." oraz przebiegi częstotliwości podstawowej w ich obrębie (linia ciągła).

6. Oprogramowanie.

Program główny i wszystkie jego moduły zostały napisane w języku Turbo Pascal, wersja 5.0. Jak już podano we wstępie, program korzysta z modułów edytora, transkrypcji oraz syntezy, tworząc zintegrowane mini-środowisko przetwarzania polskich tekstów ortograficznych na mowę. Program działa w trybie nakładkowania modułów, co umożliwia jego łatwą modyfikację w środowisku języka źródłowego, śledzenie przebiegu dołączanych w toku prac nad systemem nowych części algorytmu i eliminację ewentualnych błędów czynnościowych systemu.

Poniżej przedstawiono ogólny schemat działania systemu.



Uwagi:

a) Polecenia: <F1>, <F2>, <F3>, <Alt>+N, <Alt>+W, <Alt>+Q realizowane są typowo. Po podaniu polecenia <Alt>+P program oczekuje wyboru jednej z opcji: "o" - tylko tekst ortograficzny; "f" - tylko tekst fonematyczny; "x" - obydwa teksty.

b) Wszystkie operacje transkrypcji, zapisu na dyskietce i wydruku dotyczą:

- całego tekstu, gdy nie zaznaczono bloku;
- bloku, w przeciwnym przypadku.

c) Transkrypcja wykonywana jest tylko wtedy, gdy tekst lub blok tekstu, który ma zostać "wygłoszony", nie był jeszcze poddawany syntezie, lub został zmodyfikowany po ostatniej operacji <F10> (sygnalizacja: parametr "synt" na ramce edytora).

6.1. Wykaz operacji edytora.

Wpisywanie tekstu: litery "polskie" poprzez sekwencję: '? , gdzie "?" - znak bazowy dla danej litery (np. 'a -> ą); inne znaki - w normalnym trybie.

Klawisze edycyjne:

<Left>, <Right>, <Up>, <Dn>, <PgUp>, <PgDn>, <Del>, <Bsp>, <Home>,<End>, <Ctrl>+<Left>, <Ctrl>+<Right>, <Ctrl>-<End>.

Zaznaczenie początku i końca bloku, zniesienie bloku:

<Ctrl>+B, <Ctrl>+K, <Ctrl>+D;

Justacja całego tekstu (znosi zaznaczenie bloku):

<Ctrl>-J.

6.2. Dane wejściowe programu.

1° Zbiory danych wykorzystywane przez moduł syntezy:

a) ALL_PCF.BAZ - parametryczne wzorce głosek.

Liczebność zbioru wzorców jest kompromisem między taką wersją algorytmu syntezy alofonicznej, w której nie stosuje się w jawnej postaci reguł segmentalnych (tzn. nie ma procedur tworzenia tranzjentów, gdyż tablica wzorców zawiera elementy odpowiadające wszystkim możliwym w danym języku sekwencjom : wówczas jednak liczba potrzebnych wzorców staje się bardzo duża), oraz wersją, w której każdemu znakowi fonemu odpowiada dokładnie jeden wzorzec, a procedury oparte na regułach segmentalnych realizują wszystkie potrzebne modyfikacje (wtedy typowe i często używane w syntezie

modyfikacje trzeba wciąż powtarzać od nowa, co czasowo jest nieekonomiczne).

b) TAB_DEK.BAZ - linie deklinacyjne dla konturu intonacyjnego oraz

c) TAB_AKC.BAZ - wzorce zmian F0 dla akcentów intonacyjnych. Dane te są stablicowanymi wykresami rodzin linii deklinacyjnych oraz ruchów akcentowych, utworzonych wg wzorów Fujisaki [6], (zob. rozdz. 5). Krok tablicowania wynosi 10 ms.

2° ORTA.BAZ, FONA.BAZ - wzorce: polskich znaków ortograficznych oraz znaków fonematycznych dla trybu tekstowego 9-igłowej drukarki.

3° Wzorce "polskich" znaków dla kart EGA lub VGA - zawarte jako stała struktura w danych globalnych programu.

4° Teksty ortograficzne i fonematyczne - powstające w trakcie sesji z programem (mogą być wczytywane z dyskietki).

Dane 1°- 3° są jednokrotnie wczytywane/inicjowane na początku sesji, 4° są właściwymi danymi użytkowymi programu.

6.3. Dane wyjściowe programu.

1° Parametry mowy dla syntetyzatora - zbiory tworzone dynamicznie w pamięci komputera dla każdego kolejnego zdania, po wydaniu polecenia syntezy (<F10>);

2° Teksty ortograficzne (i fonematyczne - jeśli dla danego tekstu ortograficznego wykonano transkrypcję) - mogą być zapamiętane na dyskietce i/lub wyprowadzone na drukarkę.

6.4. Wykorzystanie pamięci przez program oraz dane

1° Pamięć dyskowa:

a) suma długości tekstów źródłowych - ok.303 Kb (Turbo Pascal v.5.0);

b) suma długości zbiorów danych wejściowych, wczytywanych na początku sesji z programem - ok. 21 Kb;

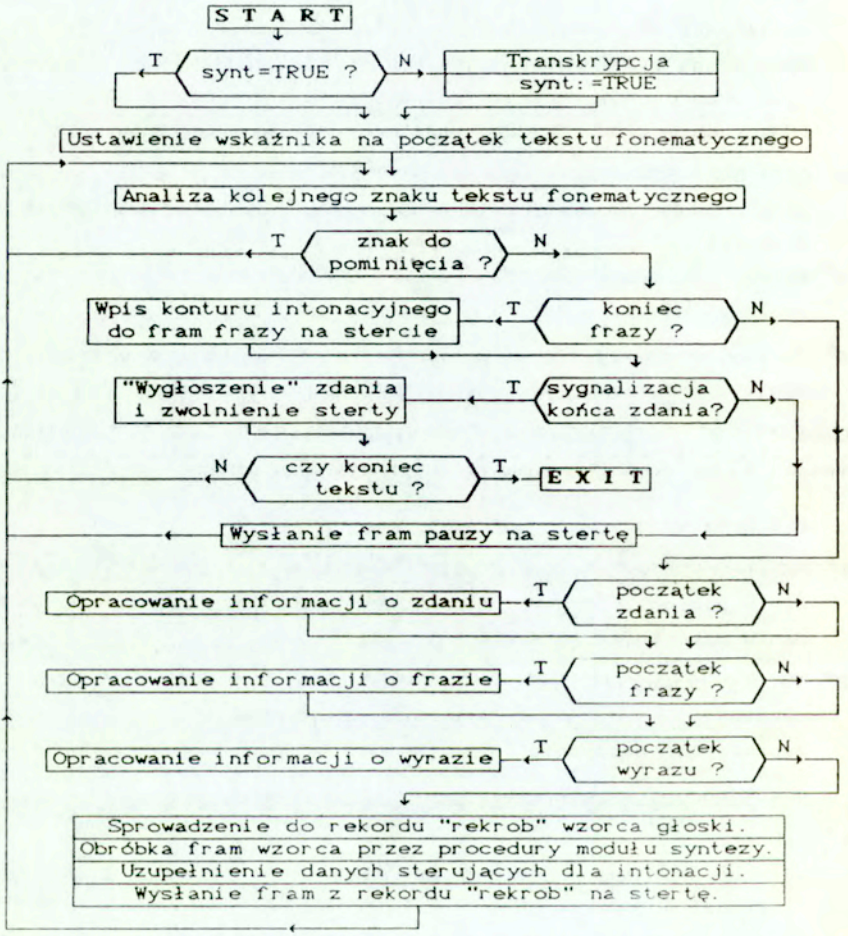
c) kod wykonywalny: program główny 36 Kb; zbiór nakładkowy 173 Kb.

2° Pamięć operacyjna:

a) pamięć przydzielana programowi razem z buforem nakładek, segmentem stosu programu i segmentem danych - ok. 155 Kb;

b) potrzebna pamięć sterty - zależy od najdłuższego zdania w syntetyzowanym tekście (np. zdanie 10-wierszowe ok.25 Kb).

6.5. Struktura algorytmu syntezy:



Na wszystkich poziomach powyższego schematu moduł syntezy opracowuje informację sterującą dla rozmaitych celów, realizowanych przez oddzielne zbiory reguł (por. poprzednie rozdziały).

W poniższym zestawieniu informacja przeznaczona do wykorzystania przez reguły segmentalne jest oznaczona literą

"S" (prawie w całości zapewnia ją tablica tekstu fonematycznego), przez reguły intonacyjne - "I", przez pozostałe reguły prozodyczne - literą "P".

Informacja o zdaniu :

a) liczba "fraz" składających się na zdanie, przy czym termin "fraza" ma tutaj jedynie znaczenie techniczne i oznacza ciąg wyrazów: od początku zdania (tekstu) do pierwszego znaku przestankowego, między dwoma kolejnymi znakami przestankowymi, bądź też od znaku przestankowego do końca zdania (tekstu) - "I";

b) typ zdania: oznajmujące, pytające (z rozróżnieniem pytań o rozstrzygnięcie i pytań o uzupełnienie), wykrzyknikowe, wykrzyknikowo-pytające (aktualnie algorytm rozróżnia tylko oznajmienia i dwa typy pytań) - "I";

c) numer aktualnie opracowywanej frazy w bieżącym zdaniu - "P","I".

Informacja o frazie ("fraza" nadal w powyższym znaczeniu) :

a) lista wyrazów wchodzących w skład frazy, ich liczba oraz rozmieszczenie sylab akcentowanych (odrębny algorytm ustala akcenty w ciągach wyrazów jednosylabowych) - "P","I";

b) lista odcinków iloczasowych, na które dzielą całą frazę sylaby akcentowane, oraz łączny iloczas frazy - "I";

c) numer aktualnie opracowywanego wyrazu - "S","P";

Informacja o wyrazie :

a) liczba sylab w wyrazie - "P";

b) występowanie i pozycja sylaby akcentowanej, stopień wyróżnienia tej sylaby (akcent "wyrazowy" lub "zdaniowy") - "P";

c) numer aktualnie opracowywanej sylaby - "S","P".

Informacja o bieżącej głosce :

a) lewe i prawe sąsiedztwo - "S";

b) występowanie w obszarze anakruzy, w obszarze wygłosu zamkniętego lub otwartego frazy - "P";

c) pozycja względem najbliższych samogłosek akcentowanych (przed- lub poakcentowa) - "P";

d) przynależność do stopy akcentowej o określonej długości (wartość w sylabach) - "P";

e) dla spółgłosek - przynależność do zbitki spółgłoskowej - "S","P".

6.6. Cykle algorytmu syntezy.

Główną pętlą modułu jest cykl zdaniowy - sekwencja operacji prowadząca do "wypowiedzenia" przez syntetyzator kolejnego zdania. Realizowany on jest przez opracowanie wszystkich kolejnych "fraz" składających się na zdanie, a następnie przesłanie syntetyzatorowi przygotowanych parametrów, w miarę gotowości urządzenia do przyjęcia następnej jednostki danych, tzw. *framy*. Frama jest 6-bajtową strukturą, kodującą w polach bitowych o różnej długości: 1.składowe formantowe, 2.szerokości wstęg formantowych, 3.częstotliwość podstawową, 4.amplitudę, 5.iloczas framy podstawowej, 6.krotność framy podstawowej. Ważną właściwością użytego w systemie syntetyzatora jest interpolacja schodkowa parametrów 1-4 metodą "od framy do framy", co umożliwia modyfikowanie przez program przejść między głoskami - za pomocą sterowania iloczasem kolejnych fram. Następny cykl rozpoczyna się znowu od przygotowania ciągu fram - parametrów zdania, co w komputerze PC/AT z zegarem 12 MHz trwa, zależnie od długości tego zdania, 0.5 - 1.5 sekundy, dając efekt symulacji przerwy międzyzdaniowej w mowie naturalnej. Ciąg fram dla zdania umieszczany jest w pamięci rezerwowanej dynamicznie - pamięć ta jest zwalniana natychmiast po wypowiedzeniu zdania przez syntetyzator.

Cykl frazowy zaczyna się od sprawdzenia numeru frazy w zdaniu. Gdy nie jest to fraza pierwsza, na stercie odkładany jest ciąg fram generujących odpowiednią pauzę. Tworzona jest lista wyrazów frazy z zaznaczonymi sylabami akcentowanymi. Następnie opracowywany jest ciąg głosek składających się na frazę, "taktowany" podwójnie: uzupełnianiem informacji o kolejnym wyrazie oraz przechodzeniem do nowego fragmentu ciągu, zawartego między kolejnymi samogłoskami akcentowanymi (zob. informacja o frazie, punkt b)). Kopia wzorca głoski - po ewentualnym zmodyfikowaniu przez procedury kształtowania tranzjentów, iloczasu i amplitudy - wysyłana jest na stertę z dodatkową informacją o łącznym iloczasie każdej framy.

Informację tę wykorzystuje uruchamiany na zakończenie cyklu frazowego wpis konturu intonacyjnego, który rozpoczyna się

wyborem wzorca linii deklinacyjnej. Następnie pętla akcentowa działa na kolejnych odcinkach z listy "informacja o frazie", punkt b). Fragment konturu dla każdego odcinka tworzony jest w trzech etapach. Procedura POSTULAT proponuje dopuszczalne w danych warunkach parametry ruchów akcentowych, wynikające z reguł kształtowania intonacji. Dopasowaniem tych parametrów do konkretnego odcinka konturu zajmuje się procedura ADAPTACJA. Kontroluje ona przedział zmienności parametu F0, a także bada, czy badany odcinek wypowiedzi ma długość wystarczającą do wykonania postulowanych ruchów akcentowych - jeśli tak nie jest, to skala odpowiednich ruchów ulega zmniejszeniu. Ostateczny wpis wartości F0 realizuje procedura KONTUR i wywoływane przez nią funkcje pomocnicze, sumując wartości z tablicy deklinacyjnej z wartościami odpowiednich ruchów akcentowych.

Trójetapowy model czynnościowy kształtowania konturu intonacyjnego został przyjęty w celu uproszczenia struktur decyzyjnych, prowadzących do jego wygenerowania. Podstawą takiej koncepcji modelu jest zasada, dająca zastosować się w wielu jego punktach: "wykonaj to, co możliwe do wykonania spośród tego, co powinno być wykonane". Uzyskane rezultaty wskazują, że droga dalszego doskonalenia intonacji leży przede wszystkim w jej urozmaiceniu. Przy zastąpieniu szeregu reguł, aktywnie ustawiających szczegółowe parametry intonacji, przez powyższą "zasadę oportunizmu" można będzie w przyszłości znacznie zwiększyć liczbę wariantów konturu, mieszcząc się wciąż w czasie rzeczywistym działania algorytmu syntezy.

7. Uwagi końcowe.

Opisanego wyżej układu syntezy MZT nie można jeszcze uznać za produkt w pełni skończony. Istnieje z pewnością co najmniej kilka możliwych kierunków jego optymalizacji (np. w zakresie oprogramowania czy generowania konturów intonacyjnych). Już jednak nawet w swojej obecnej postaci układ ten mógłby znaleźć konkretne zastosowania.

W wersji prototypowej urządzenia sygnał wytworzony przez syntetyzator przechodzi poprzez wzmacniacz na kolumnę głośnikową. Możliwe są oczywiście alternatywne rozwiązania,

polegające na umieszczeniu syntetyzatora na karcie instalowanej w komputerze, bądź skonstruowaniu urządzenia zewnętrznego, wyposażonego w mikrokontroler i pamięć ROM, przetwarzającego przesyłane z komputera teksty w kodzie ASCII do ostatecznej postaci - sygnału mowy.

Synteza wykonywana jest "w czasie rzeczywistym", tzn. jedynymi przerwami w toku jej trwania są pauzy między kolejnymi frazami wypowiedzi. Długość odcinka czasu od wywołania syntezy do startu wypowiedzi zależy od długości tekstu oraz szybkości działania komputera: np. dla komputera PC AT286 z zegarem procesora 12 MHz oraz tekstu o długości 1,5 strony ekranowej (do 65 znaków w wierszu), czas ten wynosi 2-3 sekundy.

Maksymalna długość syntezowanego tekstu obejmuje obecnie 11 stron ekranowych; w razie zaistnienia takiej potrzeby - po niewielkiej przeróbce programu - może ona zostać powiększona do stu i więcej stron.

Program syntezy mowy jest programem "przyjaznym": posługiwanie się nim nie wymaga żadnych specjalnych ćwiczeń - wystarczy podstawowa znajomość klawiatury komputera i kilku standardowych poleceń edytora.

Zrozumienie wytwarzanej przez układ mowy nie nastręcza trudności, co potwierdzają opinie także tych osób, które po raz pierwszy zetknęły się z mową syntetyczną. Słuchanie nie jest męczące, nie wymaga bowiem stałej koncentracji i napięcia uwagi, które są konieczne w przypadku mowy o niskiej zrozumiałości.

Synteza mowy z tekstu byłaby z pewnością bardzo przydatna osobom niepełnosprawnym - zwłaszcza niewidomym i niemym, bądź cierpiącym na poważne zaburzenia mowy. Tej pierwszej grupie stworzyłaby możliwość zastąpienia brakującego, wzrokowego sprzężenia zwrotnego, sprzężeniem słuchowym - w przypadku pracy z komputerem, lub - po sprzężeniu z czytnikiem pisma - możliwość odczytywania tekstów "na głos". Drugiej zaś umożliwiłaby komunikowanie się z otoczeniem za pomocą głosu, co byłoby szczególnie pożyteczne w przypadku łączności telefonicznej, gdy nie istnieje możliwość nawiązania z rozmówcą bezpośredniego kontaktu wzrokowego.

Przed syntezą mowy z tekstu otwiera się oczywiście także

szereg innych potencjalnych zastosowań, np. w nauczaniu języka polskiego, w rehabilitacji zaburzeń mowy, pisania i czytania, w automatyzacji prac biurowych, w komputerowej edycji tekstów, w "mówiących" bazach danych, w systemach informacyjnych i alarmowych, w bankowości czy telefonii. **Już obecnie synteza mowy może wyposażyć maszynę w specyficznie ludzką dotychczas zdolność - przekazywania człowiekowi informacji w najbardziej naturalny dla niego sposób - za pomocą mowy.**

BIBLIOGRAFIA

[1] ALLEN, J., M.S. HUNNICUTT, D.H. KLATT, *From Text to Speech: The MITalk System*, Cambridge University Press, Cambridge 1987.

[2] BIELECKI, J.,*Turbo Pascal 5.0. Wersja profesjonalna*, Wyd. Komunikacji i Łączności, Warszawa, 1989.

[3] BOLC, L., M. MAKSYMIENKO, *Komputerowy system przetwarzania tekstów fonematycznych*, Wydawnictwa Uniwersytetu Warszawskiego, Warszawa, 1981.

[4] BUŁHAK, L., R.GOCZYŃSKI, M.TUSZYŃSKI, *DOS od środka. Techniczny opis systemu operacyjnego DOS*, Wyd. "HELP", Warszawa, 1990.

[5] FRĄCKOWIAK-RICHTER, L., *The duration of Polish vowels*, w: Speech analysis and synthesis (red. W. Jassem), vol. 3, PWN, Warszawa, 1973, str. 87-115.

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

[7] FUJISAKI, H., K. HIROSE, N. TAKAHASHI, H. MORIKAWA, *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.

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

[9] KIEŁCZEWSKI, G., *Digital synthesis of speech and its prosodic features by means of a microphonemic method*, Sprawozdania Instytutu Informatyki Uniwersytetu Warszawskiego, nr 65, 1978, Warszawa.

[10] KLATT, D.H., *Review of text-to-speech conversion for English*, 1987, JASA 82, 737-793.

[11] ŁUKASZEWICZ, K., A. RĘGOWSKI, *Reguły syntezy wzorców fonemów i transkrypcja fonetyczna tekstu polskiego w mikrofonemicznym syntezatorze mowy*, Prace IBiB PAN, nr 25, 1988, Warszawa.

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

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

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

[15] RICHTER, L., *Duration of Polish consonants*, w: Speech analysis and synthesis (red. W. Jassem), vol. 4, str. 219-238, PWN, Warszawa, 1976.

[16] RICHTER, L., *Wstępna charakterystyka izochronizmu zestrojowego w języku polskim*, Prace IPPT 4/1983, Warszawa, 1983.

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

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

[19] THORSEN, N., *Stress group patterns, sentence accents and sentence intonation in Southern Jutland - with a view to German*, Annual Report of the Institute of Phonetics, University of Copenhagen, 23, 1989, str. 1-85.