J. Kacprowski, W. Mikiel

REALIZACJA PROCESU SYNTEZY MOWY ZA POMOCĄ SYNTEZATORA SYNFOR II

25/1968

WARSZAWA

N a p r a w a c h r ę k o p i s u

D o u ż y t k u w e w n ę t r z n e g o

Zakład Badania Drgań I P P T P A N .

Nakład 180 egz. Ark. wyd. 1,6. Ark. druk.2,25
Oddano do drukarni w październiku 1968 r.
Wydrukowano w listopadzie 1968 r. Nr zam. 832/0/68

Warszawska Drukarnia Naukowa , Warszawa ,
ul.Śniadeckich 8

J. Kacprowski, W. Mikiel
Pracownia Elektroakustyki
Zakład Badania Drgań IPPT

REALIZACJA PROCESU SYNTEZY MOWY ZA POMOCĄ SYNTEZATORA FORMANTOWEGO SYNFOR II

Streszczenie

Eksperymentalny formantowy syntezator mowy SYNFOR II, będący przedmiotem niniejszej pracy, jest syntezatorem typu kaskadowego o stałych skupionych, złożonym z trzech współpracujących ze sobą równolegle kanałów, przeznaczonych odpowiednio do syntezy dźwięków samogłoskowych, spółgłosek nosowych oraz pozostałych dźwięków spółgłoskowych. Funkcja transmitancji każdego kanału jest realizowana za pomocą układów biegunów sprzężonych, określających częstotliwości środkowe i szerokości pasm dystynktywnych obszarów formantowych dźwięków mowy podlegających syntezie. Funkcje źródła realizowane są odpowiednio w generatorze tonu krtaniowego w układzie fantastronu oraz w generatorze szumów z diodą półprzewodnikową. Syntezator jako całość jest sterowany jedenastoma sygnałami parametrycznymi kontrolującymi odpowiednio: częstotliwość /F_0/ i amplitudę tonu krtaniowego w kanale samogłoskowym /A_V/ i nosowym/A_N/, amplitudę szumów w kanale samogłoskowym /A_H/ i spółgłoskowym /A_C/ oraz częstotliwości biegunów funkcji transmitancji kanału samogłoskowego /F1, F2, F3/, nosowego /N2/ i spółgłoskowego /K1, K2/.

W pracy omówione są, w oparciu o poprzednie publikacje autorów z dziedziny syntezy mowy, podstawy teoretyczne projektowania syntezatora, obejmujące metody odwzorowania funk-

cji źródła $U(s)$, funkcji transmitancji $H(s)$ i funkcji promieniowania $Z_p(s)$ naturalnego organu mowy w elektronicznych układach zastępczych o stałych skupionych. W zakończeniu podane są najbardziej istotne parametry elektroakustyczne syntezatora.

1. Wstęp

Głównym celem prac badawczych w dziedzinie syntezy mowy jest, ogólnie biorąc, uzyskanie odpowiedzi na następujące pytania:

a. jakie parametry fonetyczne i akustyczne sygnału mowy są najbardziej istotne z punktu widzenia percepcji zawartych w nim informacji fonetycznych, lingwistycznych i osobniczych,

b. czy i w jaki sposób parametry te mogą być wydzielone z naturalnego sygnału mowy przy zastosowaniu dostępnych obecnie środków technicznych,

c. jak dalece można uprościć sygnały parametryczne kontrolujące proces syntezy, na przykład przez ich dyskretyzację, tj. kwantyzację amplitudową i czasową, przy zachowaniu dostatecznej zrozumiałości mowy syntetycznej.

Przynajmniej częściowe odpowiedzi na te pytania stanowią podstawę do realizacji technicznej systemów i układów do przekształcania, kodowania i transmisji sygnału mowy zarówno pod kątem zwiększania efektywności jego przesyłania kanałem telekomunikacyjnym, jak i automatycznego rozpoznawania elementów segmentalnych mowy oraz identyfikacji cech osobniczych głosu nadawcy wiadomości. Wszystkie wymienione wyżej oraz inne, w chwili obecnej jeszcze perspektywiczne i nie zawsze dające się dokładnie sprecyzować zastosowania wdrożeniowe podstawowych prac badawczych w tej dziedzinie [1] [2] mają ważne znaczenie techniczne, ekonomiczne i społeczne dla gospodarki narodowej.

Techniczny proces syntezy mowy sprowadza się do rozwiązania dwóch problemów, z których jeden, o charakterze fonetyczno-akustycznym, polega na przedstawieniu elementów segmentalnych sygnału mowy, na przykład o rozciągłości fonemu, diady lub sylaby, w postaci odpowiednio wybranego zespołu sygnałów parametrycznych lub kodowych programujących proces syntezy, a drugi obejmuje zagadnienia techniczne związane z przetwarzaniem tych sygnałów w akustyczny sygnał mowy syntetycznej. Przedmiotem niniejszej pracy jest przedstawienie w wielkim skrócie drugiego z wymienionych problemów, przy ograniczeniu się do jednej z wielu możliwych metod syntezy, a mianowicie syntezy formantowej. Zagadnienia związane z programowaniem procesu syntezy, a więc dotyczące jego strategii, będą referowane sukcesywnie w kolejnych publikacjach.

Rekonstrukcja widma sygnału mowy w jego uproszczonej postaci, ograniczonej do najważniejszych z percepcyjnego punktu widzenia zakresów częstotliwości obejmujących tzw. formanty dystynktywne[1/], może być realizowana w technicznym układzie syntezy dwiema metodami. Pierwsza polega na odtworzeniu w układzie elektrycznym o stałych rozłożonych konfiguracji geometrycznej efektorów artykulacyjnych naturalnego organu mowy i stworzeniu w ten sposób modelu analogowego o charakterystyce transmitancji odpowiadającej z założoną z góry dokładnością charakterystyce transmitancji anatomicznego modelu kanału głosowego. Oparte na tej zasadzie tzw. formantowe syntezatory analogowe /patrz np. [3] / stanowią cenne narzędzie prac badawczych i dydaktycznych w dziedzinie fonetyki akustycznej, jednak ze względu na ich skomplikowaną strukturę układową nie znajdują zastosowania przy technicznej realizacji procesu syntezy do celów przekazywania i

---

1/ Formanty dystynktywne są to te formanty danego dźwięku mowy, które określają jego najbardziej istotne cechy fonetyczne i akustyczne, umożliwiające jego rozpoznawanie.

przetwarzania informacji. W drugiej metodzie syntezy rezygnuje się z odwzorowywania w układzie analogowym konfiguracji geometrycznej organu mowy, ograniczając się do odtworzenia za pomocą kilku selektywnych obwodów elektrycznych o stałych skupionych kształtu obwiedni widma w określonych pasmach częstotliwości, tych mianowicie, w których zawarte są najbardziej istotne informacje dotyczące fonetycznych cech dystynktywnych dźwięku mowy podlegającego syntezie. Działające na tej zasadzie syntezatory nazywane są syntezatorami formantowymi o stałych skupionych /termin angielski: "terminal-analog synthesizer" [4] /. Oczywiście obie metody syntezy wymagają odwzorowania w układach zastępczych charakterystyk źródeł energii wzbudzających falę głosową oraz charakterystyk nadajnika akustycznego promieniującego tę falę w otaczającą przestrzeń.

Opracowywany i badany w Pracowni Elektroakustyki Zakładu Badania Drgań IPPT - PAN doświadczalny syntezator formantowy SYNFOR II, którego najbardziej istotne indywidualne cechy charakterystyczne będą krótko omówione w drugiej części pracy, należy do klasy syntezatorów kaskadowych o stałych skupionych. Przesłanki i założenia teoretyczne stanowiące podstawę do jego projektowania i konstrukcji oparte są na wynikach prac szczegółowych wielu autorów zagranicznych i krajowych, które ze względu na ich liczebność trudno byłoby na tym miejscu cytować. Z najważniejszych publikacji zagranicznych o charakterze podstawowym, podsumowujących wyniki dotychczasowych badań, wymienić należy przede wszystkim prace G. Fanta [5], J.L. Flanagana [6] oraz kompilacyjną pracę M.A. Sapożkowa [7]. Z publikacji autorów polskich na uwagę zasługują w pierwszym rzędzie liczne prace W. Jassema dotyczące fonetyczno-akustycznej struktury mowy polskiej, których syntetycznym podsumowaniem jest napisany przez niego 3 rozdział pozycji [7] oraz przyczynkowe prace autorów, dotyczące technicznych problemów syntezy mowy. Na wyniki tych prac będziemy się powoływali w niniejszej publikacji,

nie powtarzając znanych, a przynajmniej dostępnych w literaturze rozważań szczegółowych.

2. Właściwości transmisyjne efektorów artykulacyjnych organu mowy.

2.1. Założenia ogólne.

Ogólne wyrażenie na ciśnienie akustyczne p(t) fali głosowej odpowiadającej dźwiękom mowy ma w ujęciu teorii układów linearnych postać operatorową

$$P(s) = \mathcal{L}\{p(t)\} = U(s) \cdot H(s) \cdot Z_p(s) \qquad /1/$$

gdzie:

- $s = \sigma + j\omega$ - jest częstotliwością zespoloną,
- $U(s)$ - jest funkcją wzbudzającą źródła,
- $H(s)$ - jest funkcją transmitancji kanału głosowego,
- $Z_p(s)$ - jest funkcją promieniowania otworu ust lub/i nosa.

Pierwszym etapem technicznego procesu syntezy jest zatem określenie analityczne, a następnie odwzorowanie w elektrycznych układach zastępczych postaci funkcji $U(s)$, $H(s)$ i $Z_p(s)$, których charakter zależy od warunków artykulacji poszczególnych dźwięków mowy i od odpowiadającej im zmiennej konfiguracji geometrycznej organu mowy, określającej jego chwilową strukturę akustyczną. Czynnikiem podstawowym we wzorze /1/ jest funkcja transmitancji kanału głosowego $H(s)$, której postać ogólna jest różna dla różniących się pod względem sposobu artykulacji klas głosek. Jako podstawowe kryterium podziału można przyjąć lokalizację źródła wzbudzającego w kanale głosowym.

2.2. Wzbudzanie krtaniowe.

Wzbudzanie krtaniowe dotyczy najprostszego i najłatwiejszego do analitycznego ujęcia przypadku, kiedy jedynym źródłem energii są drgające wiązadła głosowe zamykające od strony wlotu, tj. w punkcie x = 0, kanał głosowy, który można aproksymować do postaci rury akustycznej o długości l i niejednostajnym przekroju A(x), zakończonej u wylotu, tj. w punkcie $x = l$, otworem ust lub nosa. Takie warunki artykulacji odpowiadają dźwiękom mowy, które z akustycznego punktu widzenia można traktować jako przebiegi quasi-periodyczne o strukturze wyłącznie harmonicznej, czyli:

a. samogłoski sylabiczne /ustne/ - [a] [o] [u] [i] [ɨ] [e],
b. samogłoski niesylabiczne - [j] [w],
c. spółgłoski nosowe - [m] [n] [ɲ] [ŋ]
d. spółgłoski boczne - [l] oraz rzadko występujący we współczesnej polszczyźnie fonem [ɫ] 1/.

Stosując odpowiednie metody aproksymacji /patrz np. [5] [6] [8]/ można wykazać, że funkcja transmitancji H(s) takiego modelu akustycznego, wyrażona stosunkiem operatorowym prędkości akustycznej U(s) w otworze wylotowym ust do prędkości akustycznej $U_0(s)$ w szczelinie wiązadeł głosowych ma postać

$$H(s) = \frac{U(s)}{U_0(s)} = \prod_{i=1}^{\infty} \frac{s_i \cdot s_i^*}{(s-s_i)\cdot(s-s_i^*)} \qquad /2/$$

i spełnia następujące warunki:

1/ W niniejszej pracy przy oznaczaniu fonemów języka polskiego stosowany jest konsekwentnie międzynarodowy system transkrypcji fonetycznej /patrz np. [7], s. 80/.

a. nie ma zer,

b. ma nieskończoną liczbę par biegunów sprzężonych s_i , $s_i^* = \sigma_i \pm j\omega_i$ odpowiadających poszczególnym obszarom formantowym rozpatrywanej głoski,

c. przy częstotliwości s = 0 przybiera wartość $|H(s)| = 1$.

Funkcję H(s) /2/ można zrealizować w kaskadowym połączeniu $n \to \infty$ niezależnych i wzajemnie od siebie odseparowanych szeregowych obwodów rezonansowych RLC o właściwościach filtrów dolnoprzepustowych /rys. 1/, z których każdy opisany jest funkcją transmitancji napięciowej typu

$$H_i(s) = \frac{E_{i+1}(s)}{E_i(s)} = \frac{\frac{1}{L_i C_i}}{s^2 + \frac{R_i}{L_i} s + \frac{1}{L_i C_i}} = \frac{s_i \cdot s_i^*}{(s - s_i)(s - s_i^*)} \qquad /3/$$

gdzie:

$$s_i , s_i^* = \sigma_i \pm j\omega_i \qquad /4/$$

a wielkości

$$\omega_i = \sqrt{\frac{1}{L_i C_i} - \frac{R_i^2}{4L_i^2}} \qquad /5a/$$

$$\sigma_i = - \frac{R_i}{2L_i} \qquad /5b/$$

związane są z częstotliwością środkową F_i i szerokością pasma ΔF_i /na poziomie -3 dB względem wierzchołka/ i-tego obszaru formantowego zależnościami:

$$F_i = \frac{\omega_i}{2\pi} \qquad /6a/$$

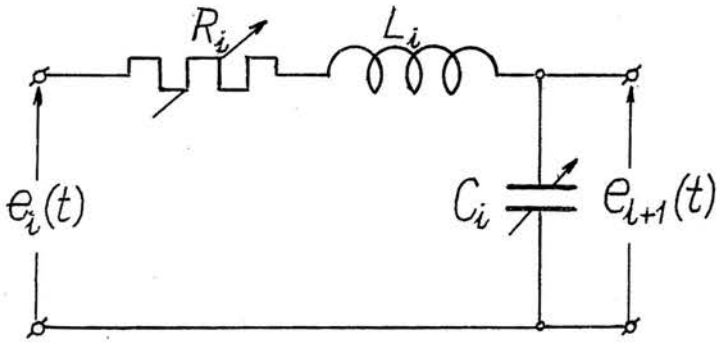
$$\Delta F_i = -\frac{\sigma_i}{\pi} \qquad /6b/$$

Rolę elementów regulacyjnych obwodu rezonansowego z rys. 1 pełni pojemność C_1 i oporność strat R_1, gdyż od ich wartości zależy, jak to wynika ze wzorów /5/ i /6/, częstotliwość F_1 i szerokość pasma ΔF_1 określonego formantu.

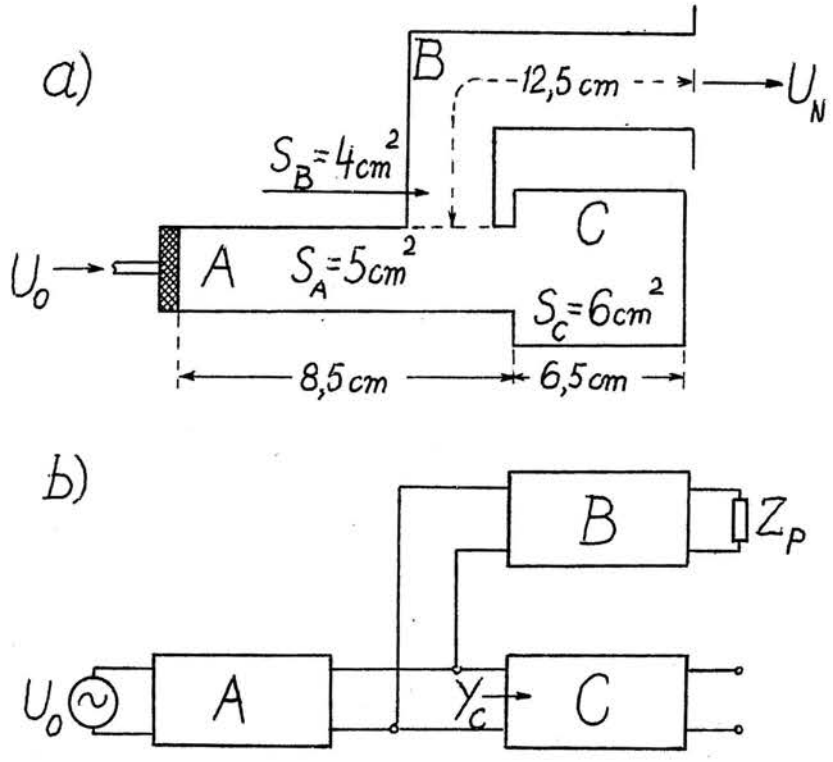
Można łatwo wykazać w sposób analityczny /patrz np. rozdz. 5 pracy [8] /, że przy kaskadowym połączeniu n obwodów rezonansowych położenie biegunów s_1, s_1^* na płaszczyźnie częstotliwości zespolonej określa w sposób jednoznaczny nie tylko częstotliwości i szerokości pasm wszystkich n formantów, lecz także względne stosunki wysokości ich wierzchołków. Fakt ten decyduje między innymi o przewadze eksploatacyjnej syntezatorów kaskadowych nad równoległymi, gdyż w przypadku tych ostatnich niezbędne jest kontrolowanie poziomów amplitud poszczególnych formantów za pomocą dodatkowych indywidualnych organów regulacyjnych lub sygnałów sterujących.

2.3. Korekcja wyższych formantów.

Akustyczne i fonetyczne cechy dystynktywne głosek o pobudzaniu krtaniowym, a w szczególności samogłosek, określone są kształtem obwiedni ich widma w zakresie trzech lub co najwyżej czterech pierwszych formantów, od F1 do F4 włącznie. Z tego względu w technicznych układach syntezy kanał samogłoskowy realizowany jest zazwyczaj w postaci kaskadowego połączenia czterech niezależnych obwodów rezonansowych RLC, z których każdy opisany jest funkcją typu /3/ i wprowadza parę biegunów sprzężonych s_i, s_i^* odpowiadających i-temu formantowi, gdzie $i = 1, 2, 3$ lub 4. Uwzględnienie w układzie syntezy tylko k niższych formantów, a pominięcie formantów wyższych rzędów, od $(k + 1)$ do ∞, jest pozornie nieistotne z fonetycznego i percepcyjnego punktu widzenia,



Rys.1. Obwód rezonansowy RLC o właściwościach filtru dolnoprzepustowego reprezentujący parę biegunów sprzężonych s_i, s_i^* funkcji transmitancji H(s)



Rys.2. Uproszczony model akustyczny organu mowy przy wymawianiu spółgłosek nosowych /a/ i jego elektryczny układ zastępczy /b/. Rozmiary dotyczą głoski [m] [6]

jednak zmienia w sposób zasadniczy kształt obwiedni widma w zakresie formantów od $\underline{i} = 1$ do $\underline{i} = \underline{k}$, co wpływa ujemnie na naturalność i rozpoznawalność głoski syntetycznej. Z tego względu niezbędne jest wprowadzenie dodatkowego układu korekcyjnego, który kompensowałby to niepożądane zjawisko.

Przedstawiając funkcję transmitancji H(s) wyrażoną wzorem /2/ w postaci iloczynu dwóch czynników:

$$H(s) = \prod_{i=1}^{\infty} \frac{s_i \cdot s_i^*}{(s-s_i)(s-s_i^*)} = \prod_{i=1}^{k} \frac{s_i \cdot s_i^*}{(s-s_i)(s-s_i^*)} \cdot \prod_{i=k+1}^{\infty} \frac{s_i \cdot s_i^*}{(s-s_i)(s-s_i^*)} =$$

$$= H_k(s) \cdot Q_k(s) \qquad /7/$$

można wykazać [6], że amplitudowa charakterystyka układu korekcyjnego, kompensującego wpływ pominięcia formantów /biegunów/ wyższych rzędów, od $(k + 1)$ do ∞, ma postać

$$\ln |Q_k(\omega)| \approx \left(\frac{\omega}{\omega_1}\right)^2 \left[\frac{\pi^2}{8} - \sum_{i=1}^{k} \frac{1}{(2i-1)^2}\right] \qquad /8/$$

gdzie: $\omega_1 = 2\pi F_1$. Wartości numeryczne funkcji korekcyjnej /8/ dla $\underline{k} = 1, 2, 3, 4$ można znaleźć w pracach [5] i [8].

2.4. Nazalizacja. Bocznikujący wpływ wnęki jamy ustnej.

Omówiony poprzednio uproszczony model artykulacyjny kanału głosowego przy wytwarzaniu dźwięków mowy o pobudzaniu wyłącznie krtaniowym komplikuje się znacznie w przypadku głosek nosowych i nazalizowanych, do których w polskim systemie fonemów należą przede wszystkim spółgłoski nosowe [m] [n] [ɲ] [ŋ]. Uproszczony model anatomiczny kanału głosowego przy wymawianiu spółgłosek nosowych podany jest na rys.2a, a jego czwórnikowy układ zastępczy na rys. 2b. Szczegółową

analizę akustyczną układu zastępczego przeprowadzono w pracach [10] i [11] poświęconych specjalnie temu zagadnieniu[1/]. Nie powtarzając przytoczonych tam rozważań można stwierdzić, że o strukturze formantowej spółgłosek nosowych decyduje głównie konfiguracja geometryczna połączonych kaskadowo kanałów: gardłowego A i nosowego B, które stanowią główny tor akustyczny przy ich artykulacji i których ukształtowanie nie zmienia się w sposób istotny przy ich wymawianiu. Zasadniczą rolę pełni natomiast w tym przypadku zamknięta od strony wylotu wnęka jamy ustnej C, która bocznikuje tor główny w miejscu połączenia obu kanałów gardłowego A i nosowego B i powoduje występowanie zer funkcji transmitancji przy tych częstotliwościach, przy których wartość akustycznej admitancji wejściowej Y_C przybiera wartości $Y_C = \pm \infty$. Funkcja transmitancji $H_N(s)$ kanału głosowego przy wymawianiu spółgłosek nosowych ma zatem w zakresie częstotliwości do około 3000 Hz postać ogólną

$$H_N(s) = H_k(s) \cdot H_o(s) \qquad /9/$$

gdzie zgodnie ze wzorem ogólnym /7/

$$H_R(s) = \prod_{i=1}^{k} \frac{s_i \cdot s_i^*}{(s-s_i)(s-s_i^*)} \qquad /10/$$

odtwarza formantową strukturę kanału gardłowo-nosowego,

1/ W nowszych pracach z dziedziny fonetyki akustycznej języka polskiego wykazano [9], że samogłoski nosowe [ą] i [ę] w mowie potocznej nie mają charakteru odrębnych fonemów, lecz w zależności od kontekstu sprowadzają się do segmentu złożonego z samogłoski sylabicznej /ustnej/ i następującej po niej spółgłoski nosowej, jak np. w wyrazach: kąt → [k o n t] lub pięć → [p j̇ e ɲ t͡ɕ].

określoną rozkładem k par biegunów sprzężonych s_i, s_i^* $i = 1, 2, \ldots k$ na płaszczyźnie częstotliwości zespolonej, a funkcja

$$H_o(s) = \frac{(s-s_o)(s-s_o^*)}{s_o \cdot s_o^*} \qquad /11/$$

wyraża bocznikujący wpływ wnęki jamy ustnej. Zero sprzężone s_o, s_o^* odpowiada antyformantowi FO, którego położenie w skali częstotliwości zależne jest od rozmiarów wnęki jamy ustnej C przy wymawianiu poszczególnych spółgłosek nosowych i które, jak wykazuje analiza teoretyczna potwierdzona wynikami badań spektrograficznych [9], występuje odpowiednio przy częstotliwościach 900 - 1100 Hz w przypadku spółgłoski [m], 1700 - 1750 Hz [n] i 2300 - 2500 Hz [ɲ], a więc przy częstotliwościach tym wyższych, im mniejsza jest długość wnęki jamy ustnej ograniczonej odpowiednio zwarciem dwuwargowym [m], językowo-zębowym [n] i językowo-przedniopodniebiennym [ɲ]. Jedynym wyjątkiem w tej regule jest spółgłoska nosowa językowo-tylnopodniebienna [ŋ], przy wymawianiu której wnęka jamy ustnej praktycznie nie bierze udziału, co objawia się niewystępowaniem zera funkcji transmitancji $H_N(s)$.

W cytowanych poprzednio pracach wykazano [10] [11], że do celów technicznej syntezy polskich spółgłosek nosowych charakterystyczny dla spółgłosek [m] [n] [ɲ] zespół: formant-antyformant-formant można zastąpić formantem pozornym N2 o częstotliwości N_2 w przybliżeniu równej częstotliwości antyformantu FO i o odpowiednio dużej szerokości pasma ΔN_2 = =200 - 300 Hz. W wyniku tych uproszczeń funkcja transmitancji kanału nosowego syntezatora SYNFOR II mogła być zrealizowana w postaci kaskadowego połączenia trzech obwodów rezonansowych RLC jak na ryc. 1 i wyraża się wzorem

$$H'_N = \prod_{i=1}^{3} \frac{s_i \cdot s_i^*}{(s-s_i)(s-s_i^*)} \qquad /12/$$

przy czym częstotliwości i szerokości pasm formantów N1 i N3, odpowiadających parom biegunów sprzężonych s_1, s_1^* i s_3, s_3^*, są niezmiennikami: N_1 = 200 - 300 Hz, N_3 = 2500 - 2800 Hz, $\Delta N_1 = \Delta N_3$ = 200 - 300 Hz. Jedyną cechą dystynktywną spółgłosek nosowych jest przy tych uproszczeniach formant pozorny N2.

Ogólnie biorąc zatem, kanał nosowy syntezatora ma strukturę układową podobną do kanału samogłoskowego, a różnice są jedynie natury ilościowej i polegają na zlokalizowaniu wszystkich biegunów /formantów/ w zakresie częstotliwości poniżej 3000 Hz i nadaniu współrzędnym σ_1 odpowiednio dużych wartości w celu odwzorowania w układzie zastępczym stosunkowo dużych tłumień występujących w modelu anatomicznym kanału nosowego B /rys. 2/. Te różnice ilościowe zdecydowały między innymi w przypadku syntezatora SYNFOR II o zastosowaniu dwóch niezależnych kanałów syntezy: samogłoskowego i nosowego o wspólnym /równoległym/ pobudzaniu krtaniowym.

2.5. Wzbudzanie ponadkrtaniowe.

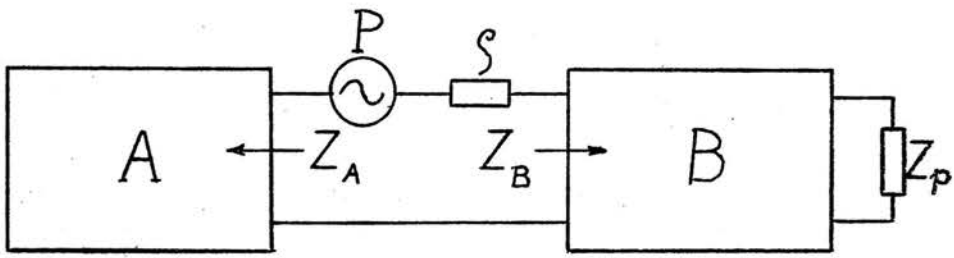
Podstawową cechą artykulacyjną spółgłosek bezdźwięcznych trących, które wytwarzane są bez udziału generatora krtaniowego /wiązadeł głosowych/ jest to, że źródło wzbudzające ma w tym przypadku charakter turbulencyjny /szumowy/ i utworzone jest przez przewężenie kanału głosowego, zlokalizowane odpowiednio w strefie zębowo-wargowej [f], zazębowej [s], dziąsłowej [ʃ], dziąsłowo-środkowojęzykowej [ɕ], lub tylnojęzykowej [x]. Z akustycznego punktu widzenia głoski te mają charakter przebiegów nieperiodycznych /szumowych/.

Zarówno skomplikowana i niedostatecznie jeszcze znana konfiguracja akustyczna kanału głosowego przy wymawianiu głosek trących, jak i zmienna lokalizacja przestrzenna źródła wzbudzającego w kanale uniemożliwia stworzenie ich dokładnego i dostępnego do analitycznego ujęcia modelu artykulacyj-

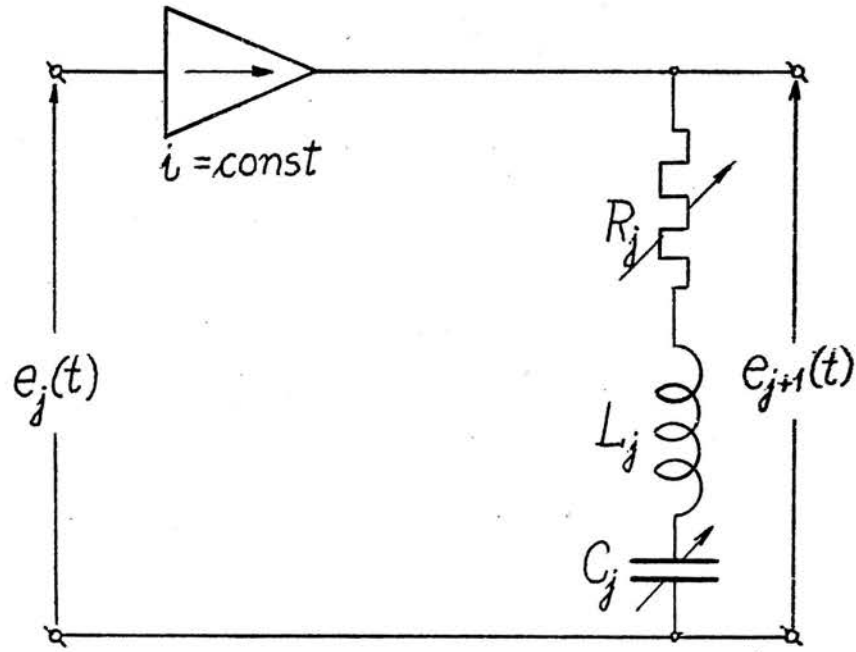
nego. Rozważania teoretyczne mogą mieć w tym przypadku raczej charakter jakościowy, a dokładne informacje można uzyskać przede wszystkim w oparciu o wyniki szczegółowej analizy spektrograficznej [12].

Spółgłoski bezdźwięczne zwarte powstają w wyniku nagłego otwarcia początkowo zamkniętej drogi przepływu powietrza w kanale głosowym i mają charakter przebiegów akustycznych złożonych, quasi-impulsowo szumowych, przy czym pod względem lokalizacji miejsca artykulacji dzielą się na dwuwargowe [p], zębowe [t], środkowojęzykowe [c] i tylnojęzykowe [k]. Przebieg szumowy, występujący po przebiegu impulsowym, jest krótki i trwa, w zależności od kontekstu, od 50 do 100 milisekund. Fonetyczno-akustyczne cechy dystynktywne spółgłosek zwartych określone są zatem nie tylko ich strukturą widmową zależną od konfiguracji geometrycznej kanału głosowego i miejsca artykulacji, lecz także charakterem stanów nieustalonych warunkujących powstawanie tranzjentów /ugięć/ formantowych. Spółgłoski zwarto-trące: zazębowe [t͡s] , dziąsłowe [t͡ʃ] i dziąsłowo-środkowojęzykowe [t͡ɕ] nie wykazują odrębnych cech akustycznych w stosunku do równoważnych im pod względem miejsca artykulacji spółgłosek trących [ʂ], [ʃ] i [ɕ], a istotne różnice sprowadzają się do innych wartości stosunków amplitud i czasów trwania segmentów impulsowych i szumowych.

Na rys. 3 przedstawiono uproszczony elektryczny schemat zastępczy kanału głosowego przy pobudzaniu ponadkrtaniowym. Symbol P oznacza źródło ciśnieniowe /napięciowe/ o rezystancji wewnętrznej ζ odpowiadającej akustycznej rezystancji przewężenia, czwórniki A i B o impedancjach wejściowych Z_A i Z_B reprezentują strefy kanału głosowego leżące odpowiednio poniżej /A/ i powyżej /B/ źródła, a Z_p jest akustyczną impedancją promieniowania otworu ust. Funkcja transmitancji $H_C(s)$ kanału głosowego przy pobudzaniu ponadkrtaniowym ma postać złożoną:



Rys.3. Uproszczony schemat zastępczy kanału głosowego przy pobudzaniu ponadkrtaniowym



Rys.4. Obwód antyrezonansowy RLC o właściwościach filtru środkowo-zaporowego reprezentujący parę zer sprzężonych s_j, s_j^* funkcji transmitancji $H_C(s)$

$$H_c(s) = H(s) \cdot H_o(s) =$$

$$= \prod_{i=1}^{\infty} \frac{s_i \cdot s_i^*}{(s-s_i)(s-s_i^*)} \cdot \prod_{j=1}^{\infty} \frac{(s-s_j)(s-s_j^*)}{s_j \cdot s_j^*} \qquad /13/$$

gdzie $H(s)$ jest funkcją określającą rozkład biegunów s_i, s_i^* $(i = 1, 2, \ldots \infty)$ odpowiadających rezonansom kanału głosowego złożonego z dwóch odcinków A i B i podobnie jak w przypadku pobudzania krtaniowego wyrażona jest wzorem /2/, natomiast $H_o(s)$ określa rozkład sprzężonych zer s_j, s_j^* $(j = 1, 2, \ldots \infty)$ funkcji $H_c(s)$, które odpowiadają antyrezonansom występujących przy tych częstotliwościach, przy których impedancja wejściowa Z_A czwórnika A reprezentującego strefę kanału głosowego leżącą poniżej przewężenia przybiera wartości $Z_A = \pm \infty$.

Z teoretycznego punktu widzenia, kanał spółgłoskowy syntezatora można zatem zrealizować w postaci kaskadowego połączenia k obwodów rezonansowych o postaci jak na rys. 1, reprezentujących bieguny s_i, s_i^* $(i = 1, 2, \ldots k)$ funkcji transmitancji, oraz l obwodów antyrezonansowych np. o postaci jak na rys. 4, reprezentujących jej zera s_j, s_j^* $(j = 1, 2, \ldots l)$. Wpływ pominiętych wyższych biegunów /formantów/ od $(R + 1)$ do ∞ oraz zer od $(l + 1)$ do ∞ można skompensować za pomocą odpowiednich układów korekcyjnych $Q_k(s)$ i $Q_l(s)$, jak to wyjaśniono w rozdziale 2.3.

Ponieważ jednak fonetyczno-akustyczne cechy dystynktywne dźwięków spółgłoskowych określone są nie tyle przez ich formanty i antyformanty sensu stricto, jak to miało miejsce w przypadku dźwięków samogłoskowych i nosowych, lecz przez maksima i minima obwiedni widma w szerokich zakresach częstotliwości, przeto rekonstrukcja ich widma do celów syntezy technicznej sprowadza się zazwyczaj do odtworzenia kształtu obwiedni za pomocą dwóch par biegunów sprzężonych funkcji transmitancji, zrealizowanych odpowiednio przez obwody rezo-

nansowe RLC, i jedno zero sprzężone zrealizowane przez obwód antyrezonansowy. System taki zastosowano między innymi w syntezatorze OVE II [13]. W syntezatorze SYNFOR II uproszczenie struktury kanału spółgłoskowego posunięto jeszcze dalej, stosując tylko dwie pary biegunów sprzężonych, realizowanych przez dwa obwody rezonansowe RLC odpowiednio o właściwościach filtrów górnoprzepustowego /K1/ i dolnoprzepustowego /K2/. Układ biegunów K1 i K2 odtwarza obwiednię widma spółgłosek w górnym zakresie częstotliwości, powyżej 2800 Hz. Dolna część widma odtwarzana jest za pomocą biegunów kanału samogłoskowego pobudzanego w tym przypadku ze źródła szumowego, a pozorne zero funkcji transmitancji powstaje w wyniku sumowania się obwiedni na granicy obu obszarów widma: dolnego i górnego.

2.6. Wzbudzanie złożone, krtaniowo-ponadkrtaniowe.

Dźwięczne odpowiedniki omówionych poprzednio spółgłosek bezdźwięcznych trących [v] [z] [ʒ] [ʑ], zwartych [b] [d] [ɟ] [g] i zwartotrących [d͡z] [d͡ʒ] [d͡ʑ], wytwarzane przy współudziale źródła krtaniowego, mają takie same lub prawie takie same parametry widmowe określone konfiguracją geometryczną kanału głosowego i lokalizacją źródła szumowego, z tą jedynie różnicą, że w dolnej części ich widma występują monotonicznie malejące składowe harmoniczne częstotliwości tonu krtaniowego. W układzie syntezatora SYNFOR II synteza widma spółgłosek dźwięcznych odbywa się zatym przy równoległej współpracy kanałów: samogłoskowego i spółgłoskowego, zasilanych jednocześnie z obu źródeł: generatora tonu krtaniowego i generatora szumów.

3. Funkcja wzbudzająca źródła krtaniowego i szumowego.

3.1. Założenia ogólne.

Poprzedni rozdział poświęcony był zagadnieniu odtwarzania w układach elektrycznych o stałych skupionych uproszczonej postaci funkcji transmitancji biernych efektorów artykulacyjnych $H(s)$ we wzorze ogólnym /1/. Należy z kolei omówić zagadnienie aproksymowania w układach elektrycznych funkcji źródła $U(s)$ wzbudzającego drgania w kanale głosowym. Zagadnienie to wymaga niezależnego rozpatrzenia dwóch przypadków: wzbudzania krtaniowego i wzbudzania szumowego.

3.2. Funkcja źródła tonu krtaniowego.

Model akustyczny generatora krtaniowego utworzonego przez szczelinę drgających wiązadeł głosowych można przedstawić w postaci źródła prądowego /tj. źródła o stałej wydajności prędkości akustycznej/ o nieskończenie dużej impedancji wewnętrznej. Z tego względu chwilowa struktura akustyczna kanału głosowego i jego zmienna impedancja wejściowa nie wpływa w sposób istotny na postać funkcji wzbudzającej $U(s)$, którą można uważać za czynnik w przybliżeniu stały we wzorze /1/.

Funkcję prędkości akustycznej $f(t)$ w szczelinie wiązadeł głosowych można aproksymować w postaci ciągu impulsów trójkątnych o amplitudzie A i częstotliwości powtarzania $F_0 = \frac{1}{T}$ równej częstotliwości tonu krtaniowego /rys. 5/. Transformata Laplace'a funkcji $f(t)$ wyraża się wzorem [8]:

$$F(s)=\frac{1}{1-e^{-sT}}\cdot\frac{A}{t_1}\cdot\frac{\frac{t_1}{t_2-t_1}\cdot e^{-st_2}-\frac{t_2}{t_2-t_1}\cdot e^{-st_1}+1}{s^2} \qquad /14/$$

i ma jeden biegun podwójny s = 0 w początku układu oraz nieskończony szereg zer sprzężonych położonych na osi urojonej płaszczyzny częstotliwości zespolonej, spełniających warunek

$$s_k, s_k^* = \pm j k \frac{2\pi}{t_2} \qquad /15/$$

gdzie: k = 0, 1, 2, ... ∞ .

Występowanie zer funkcji źródła F(s) jest zjawiskiem drugorzędnym, gdyż wyraża jedynie zależną od przyjętego kształtu impulsu krtaniowego okresowość pofalowań obwiedni widma amplitudowego. W zastosowaniach technicznych syntezy ważny jest przede wszystkim średni kształt obwiedni widma, który określony jest biegunem funkcji /14/ i który odpowiada charakterystyce amplitudowej opadającej ze stromością - 12 dB na oktawę. Funkcję źródła krtaniowego można zatem wyrazić w postaci przybliżonej określonej wzorem

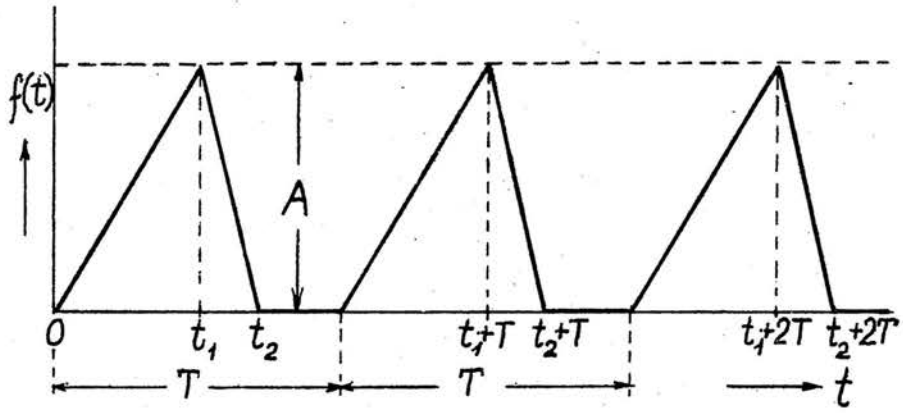
$$U(s) = \frac{K}{(1-e^{-sT})s^2} \qquad /16/$$

w którym: K = U(0) - jest rzeczywistym współczynnikiem stałym proporcjonalnym do amplitudy prędkości akustycznej w szczelinie wiązadeł głosowych,

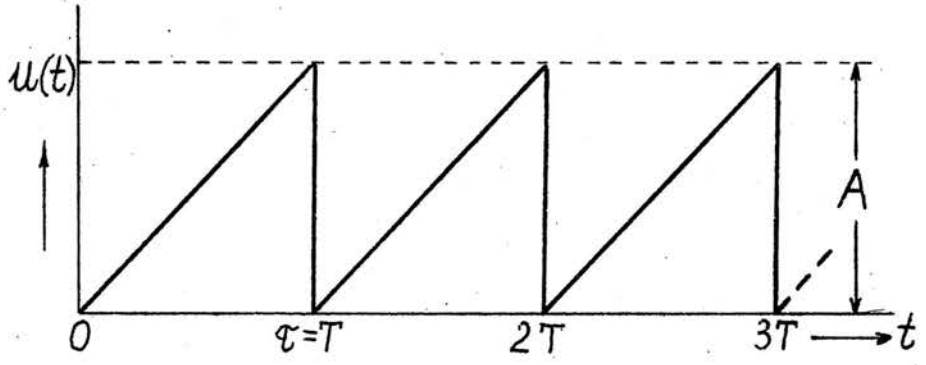
$s = \sigma + j\omega$ - jest częstotliwością zespoloną,

$\frac{1}{T} = F_0$ - jest częstotliwością powtarzania impulsów krtaniowych, równą częstotliwości podstawowej tonu krtaniowego.

W postaci funkcji /16/ źródła krtaniowego można uzwględnić jednocześnie funkcję promieniowania $Z_p(s)$, która występuje jako trzeci czynnik we wzorze ogólnym /1/ i która jest definiowana jako operatorowy stosunek ciśnienia akustycznego P(s) fali głosowej w określonym punkcie przestrzeni, zazwyczaj na osi ust w odległości l, do prędkości akustycznej U(s) w otworze wylotowym ust lub nosa:



Rys.5. Aproksymacja funkcji f(t) tonu krtaniowego w postaci ciągu asymetrycznych impulsów trójkątnych



Rys.6. Funkcja prędkości akustycznej u(t) źródła tonu krtaniowego przedstawiona w postaci ciągu impulsów piłozębatych

$$Z_p(s) = \frac{P(s)}{U(s)} \qquad /17/$$

Podobnie jak funkcja źródła krtaniowego U(s), funkcja promieniowania $Z_p(s)$ w niewielkim tylko stopniu zależy od sposobu artykulacji i może być traktowana jako czynnik w przybliżeniu stały. Przyjmując jako uproszczony model akustyczny organu mowy promieniującego falę głosową tłok drgający o promieniu r równoważnym promieniowi otworu wylotowego ust, umieszczony w obudowie kulistej o promieniu R równoważnym zastępczemu promieniowi głowy ludzkiej, otrzymuje się wyrażenie [8]

$$Z_p(s) = K(s)\frac{s}{s-s_p} \qquad /18/$$

w którym współczynnik K(s) uwzględnia kierunkowe właściwości rozpatrywanego modelu nadajnika akustycznego. Funkcja /18/ określona jest jednym zerem w punkcie s = 0 i jednym biegunem na osi rzeczywistej punkcie $s_p = \text{const} \frac{R_p}{M_p} \approx -2\pi \cdot 4000\ [\text{sek}^{-1}]$, gdzie R_p i M_p oznaczają odpowiednio akustyczną rezystancję promieniowania i masę współdrgającego ośrodka w otworze ust. W interesującym z punktu widzenia technicznych zastosowań syntezy mowy zakresie częstotliwości $f \leq 4000$ Hz amplitudowa charakterystyka funkcji promieniowania $|Z_p(\omega)|$ wzrasta zatem ze stromością + 6 dB na oktawę i może być aproksymowana w prostym elektrycznym układzie korekcyjnym np. typu RC. Biorąc pod uwagę wyniki uproszczonej analizy funkcji źródła U(s) i funkcji promieniowania $Z_p(s)$, z których pierwsza maleje ze stromością - 12 dB na oktawę, a druga rośnie ze stromością + 6dB na oktawę, można w technicznym procesie syntezy połączyć oba czynniki stałe we wzorze /1/ w jeden, uwzględniając charakter funkcji promieniowania $Z_p(s)$ w postaci analitycznej funkcji źródła U(s), której amplitudowa charakterystyka widma powinna zatem opadać ze stromoś-

cią - 6 dB na oktawę. Warunek ten można łatwo zrealizować w układzie elektronicznym, aproksymując funkcję źródła krtaniowego w postaci ciągu impulsów piłozębatych o amplitudzie A i czasie narastania τ równym pełnemu okresowi powtarzania T

$$\tau = T = \frac{1}{F_0} \qquad /19/$$

jak to przedstawiono na rys. 6. Transformata Laplace'a nieskończonego ciągu takich impulsów ma postać [14]

$$U(s) = \frac{A}{T}\left[\frac{1}{s^2} - \frac{T \cdot e^{-sT}}{s(1-e^{-sT})}\right] \qquad /20/$$

a jej amplitudowa charakterystyka widma maleje ze stromością - 6 dB na oktawę. W syntezatorze SYNFOR II zastosowano jednak bardziej skomplikowany kształt obwiedni widma funkcji źródła krtaniowego, nadając jej postać

$$U(s) = \frac{U(0)}{\prod_{i=1}^{2}(s-s_i)} \cdot \frac{1}{(1-e^{-sT})} \qquad /21/$$

gdzie $s_{i=1}$ oraz $s_{i=2}$ są biegunami położonymi na osi urojonej w ujemnej półpłaszczyźnie częstotliwości zespolonej. Jako kryterium wyboru kształtu obwiedni przyjęto wyrazistość i naturalność brzmienia polskich samogłosek i spółgłosek nosowych w oparciu o wyniki badań odsłuchowych.

3.3. Funkcja źródła szumowego.

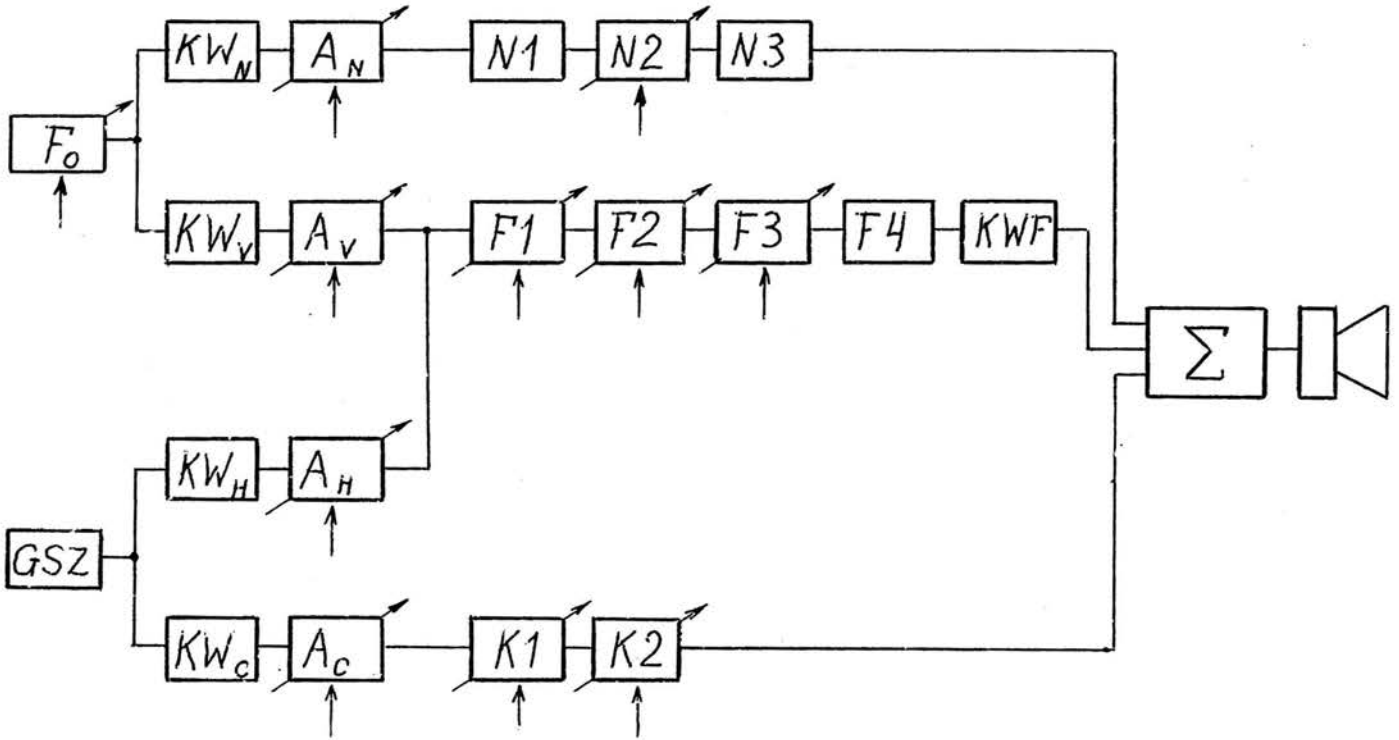
Struktura akustyczna źródła szumowego przy pobudzaniu ponadkrtaniowym jest jeszcze obecnie stosunkowo mało znana. Badania doświadczalne wykazały, że najlepsze wyniki przy

syntezie spółgłosek trących uzyskuje się stosując generator szumów o widmie ciągłym, którego obwiednia jest odpowiednio kształtowana przez funkcję transmitancji H(s) kanału głosowego. W przypadku spółgłosek zwartych źródło szumowe ma charakter impulsowy w celu odtworzenia stanów nieustalonych odpowiadających plozji, tj. nagłemu otwarciu chwilowo zamkniętej drogi przepływu powietrza w kanale głosowym.

4. Opis techniczny syntezatora.

Syntezator formantowa SYNFOR II powstał w wyniku sukcesywnej rozbudowy i ulepszeń poprzedniego modelu SYNFOR I, którego założenia teoretyczne oraz najważniejsze dane techniczne i konstrukcyjne referowane były w oddzielnych pracach /por. np. [15] [16] /. Pod względem strukturalnym SYNFOR II przypomina w pewnym stopniu syntezator formantowy OVE II G.Fanta [13] i złożony jest z trzech współpracujących ze sobą równolegle torów, przeznaczonych odpowiednio do syntezy dźwięków samogłoskowych, spółgłosek nosowych oraz pozostałych spółgłosek, tj. trących, zwartych i zwartotrących. Najważniejsze parametry elektroakustyczne syntezatora SYNFOR II, przede wszystkim te, które różnią go od jego pierwowzoru SYNFOR I oraz innych syntezatorów, będą krótko omówione w oparciu o schemat blokowy podany na rys. 7.

Generator tyratronowy, stosowany w poprzednim modelu jako źródło tonu krtaniowego, został obecnie zastąpiony generatorem fantastronowym o dużej stabilności, wytwarzającym ciąg impulsów piłozębatych o częstotliwości powtarzania F_0. Amplitudowa obwiednia widma tonu krtaniowego, opadająca ze stromością - 6 dB na oktawę, może być odpowiednio kształtowana w dwóch niezależnych układach korekcyjnych umieszczonych w torze samogłoskowym /KW_V/ i nosowym /KW_N/. Korekcja polega na kolejnym różniczkowaniu i całkowaniu pierwotnej funkcji źródła U(s) w prostych układach RC. W wyniku tej



Rys.7. Funkcjonalny schemat blokowy syntezatora SYNFOR II

korekcji wypadkowa obwiednia funkcji źródła jest płaska w zakresie częstotliwości poniżej pierwszego bieguna $s_1 = (R_1 \cdot C_1)^{-1} = 2\pi \cdot 60 \left[sek^{-1}\right]$, a następnie opada ze stromością - 6 dB na oktawę do częstotliwości drugiego bieguna $s_2 = (R_2 \cdot C_2)^{-1}$ ze stromością - 12 dB na oktawę przy wyższych częstotliwościach. Najlepsze wyniki z punktu widzenia naturalności i rozpoznawalności głosek dźwięcznych osiągnięto przy wartościach $s_2 = 2\pi \cdot 2000 \left[sek^{-1}\right]$ w torze samogłoskowym i $s_2 = 2\pi \cdot 600 \left[sek^{-1}\right]$ w torze nosowym.

Generator szumów GSZ stanowi wstecznie spolaryzowana dioda krzemowa DKS-50 współpracująca z dwustopniowym wzmacniaczem 25 dB. Widmo szumów ma charakter ciągły, a jego obwiednia jest kształtowana w układzie korekcyjnym KW_C toru spółgłoskowego, złożonym z dwóch filtrów RC, górnoprzepustowego i dolnoprzepustowego, o stromościach zboczy - 6 dB na oktawę i niezależnie regulowanych częstotliwościach granicznych. Przy syntezie większości spółgłosek szumy są wprowadzane również do toru samogłoskowego po uprzedniej korekcji ich widma w korektorze KW_H, który jest filtrem górnoprzepustowym o stromości zbocza + 12 dB na oktawę i ma na celu wyrównanie poziomów formantów spółgłoskowych w dolnej części widma.

Regulacja poziomów w poszczególnych torach syntezatora realizowana jest za pomocą czterech niezależnych modulatorów amplitudy A_V, A_N, A_H i A_C, zbudowanych na heksodach regulacyjnych ECH 81. Wzmocnienie może być regulowane w granicach 60 dB przez zmianę napięcia sterującego od 0 do - 25 V. W celu zwiększenia dokładności regulacji w zakresie małych tłumień modulatory zostały tak zaprojektowane, że zmiana napięcia regulacyjnego od 0 do - 12,5 V powoduje zmianę poziomu od 0 do - 15 dB, podczas gdy pozostałe 45 dB tłumienia wymagają zmiany napięcia regulacyjnego od - 12,5 V do - 25 V. W każdym modulatorze można wprowadzić dodatkowe tłumienie 35 dB, regulowane skokowo, w celu wyrównania względnych po-

ziomów energii w poszczególnych torach syntezatora.

Cechą charakterystyczną syntezatora SYNFOR II, różniącą go od syntezatora OVE II, jest niestosowanie układów realizujących zera funkcji transmitancji w torach spółgłoskowych. Pominięcie zera w torze nosowym wynikło z omówionych poprzednio doświadczeń nad syntezą polskich spółgłosek nosowych [10][11]. Natomiast tor spółgłoskowy ma tylko dwa układy biegunowe K1 i K2, odpowiednio o właściwościach filtrów górno- i dolnoprzepustowego. Uproszczenie to, uzasadnione poprzednio w rozdziale 2.5. niniejszej pracy, jest szczególnie korzystne z punktu widzenia sterowania syntezatora sygnałami parametrycznymi wydzielanymi z sygnału mowy naturalnej /wokoder formantowy/, gdyż parametr określający anty-formant odpowiadający zeru funkcji transmitancji trudno jest wydzielić z sygnału mowy za pomocą dostępnych środków technicznych. Przestrajanie częstotliwości biegunów we wszystkich torach syntezatora odbywa się przez zmianę pojemności C, zrealizowanej w układzie Millera, za pomocą napięciowych sygnałów sterujących.

Syntezator jako całość może być sterowany jedenastoma napięciowymi sygnałami parametrycznymi, kontrolującymi następujące parametry syntezy: częstotliwość /F_0/ i amplitudę tonu krtaniowego w torze samogłoskowym /A_V/ i nosowym /A_N/, poziom szumów w torze samogłoskowym /A_H/ i spółgłoskowym /A_C/ oraz częstotliwości biegunów w torze samogłoskowym /F1, F2, F3/, nosowym /N2/ i spółgłoskowym /K1, K2/. Częstotliwość formantu F4 w torze samogłoskowym jest stała i może być nastawiana ręcznie w granicach od F_4 = 2500 Hz do F_4 = 4500 Hz, a korekcja pominiętych wyższych formantów, od F5 do F_∞ zrealizowana jest w układzie korektora KWF o charakterystyce określonej wzorem /8/.

Programowanie syntezatora SYNFOR II na obecnym etapie prac odbywa się za pomocą 12-kanałowego generatora funkcji parametrycznych, których przebieg rysowany jest atramentem

przewodzącym na przesuwającej się z jednostajną prędkością taśmie z folii plastikowej i przetwarzany jest na sygnały napięciowe w specjalnym układzie potencjometrycznym [17]. Zasada działania generatora funkcji parametrycznych i sposób programowania procesu syntezy będzie tematem oddzielnych prac. Widok ogólny syntezatora SYNFOR II wraz z generatorem funkcji pokazany jest na rys. 8.

Sygnały parametryczne sterujące syntezator i przedstawiane obecnie w postaci funkcji ciągłych, mogą być kwantowane w skali amplitudy i czasu przez nadanie im kształtu schodkowego. Z konstrukcji generatora funkcji parametrycznych i przyjętej techniki zapisu wynika, że sygnały parametryczne kontrolujące parametry częstotliwościowe /F_0, F1, F2, F3, N2, K1, K2/ mogą być kwantowane teoretycznie w 32 stopniach poziomu, a sygnały kontrolujące parametry amplitudowe /A_V, A_N, A_H, A_C/ - w 16 stopniach poziomu. Wynikającą stąd sumaryczną objętość informacyjną H_{max} jedenastu sygnałów sterujących można wyznaczyć ze wzoru

$$H_{max} = 7 \cdot \log_2 32 + 4 \cdot \log_2 16 = 51 \text{ bitów} \quad /22/$$

co przy przyjętej częstotliwości kwantyzacji czasowej rzędu 40 razy na sekundę / Δt = 25 ms/ daje maksymalną przepustowość informacyjną I_{max} kanału sterującego syntezator

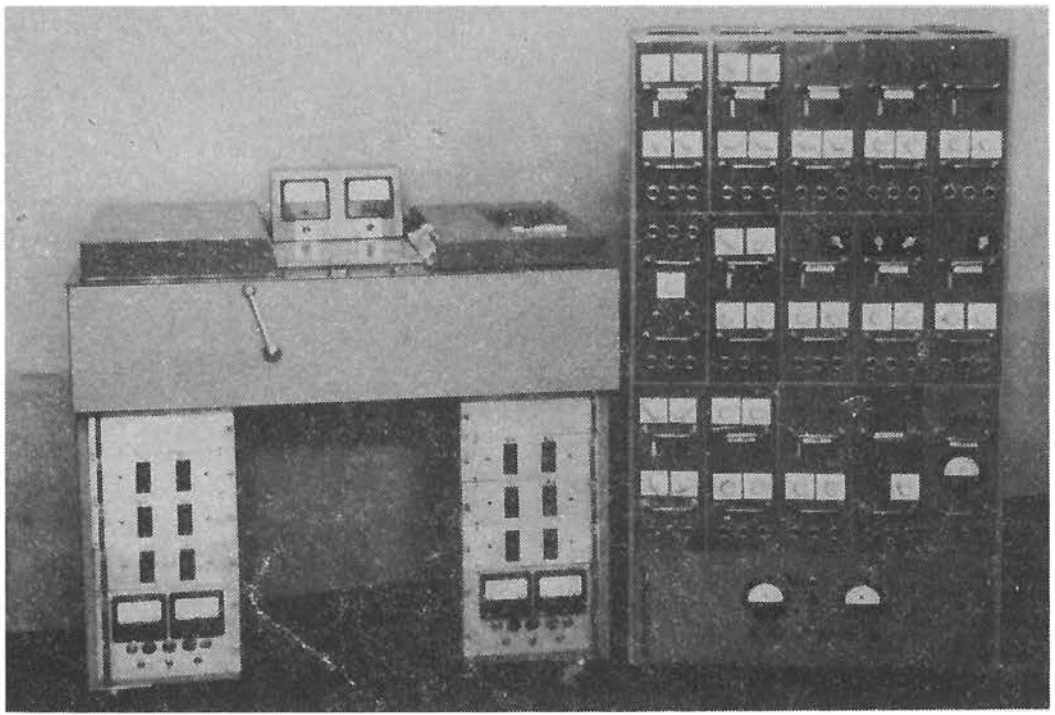
$$I_{max} = H_{max} \cdot \frac{1}{\Delta t} = 51 \cdot \frac{10^3}{25} \approx 2000 \text{ bit/s} \quad /23/$$

Obliczona wartość może być przyjęta jako orientacyjna miara korzyści, jakie z punktu widzenia zwiększenia wykorzystania przepustowości informacyjnej kanałów łączności przynieść może zastosowanie formantowej syntezy mowy. Ponieważ przepustowość informacyjna konwencjonalnych systemów telefonicznych miernej jakości, o szerokości pasma ΔF = 3000 Hz i stosunku sygnału do szumów S/N = 30 dB jest rzędu

$$I = \Delta F \cdot \log_2 (1 + S/N) \approx 30000 \text{ bit/s} \quad /24/$$

można się liczyć z około 15-krotnym zwiększeniem wykorzystania możliwości przepustowych kanałów telekomunikacyjnych.

Należy podkreślić, że wartość teoretyczna I_{max} obliczona ze wzoru /23/ jest znacznie zawyżona choćby ze względu na to, że niektóre parametry syntezy, jak np. N2, K1 oraz K2, w praktyce nie wymagają więcej niż $\underline{n} = 4$ stopni kwantyzacji. Zagadnienia związane z kwantyzacją amplitudową i czasową sygnałów parametrycznych sterujących syntezator, między innymi pod kątem zbadania możliwości jego programowania z maszyny cyfrowej /termin angielski: "synthesis by rule" [18] / będą przedmiotem dalszych prac badawczych.



Rys. 8 - Widok ogólny syntezatora SYNFOR II
wraz z generatorem funkcji parametrycznych.

WYKAZ LITERATURY

[1] Kacprowski,J.: Speech compression by means of analysis-synthesis methods. - Proc.Vibr.Probl., 5 /1964/, 3, s. 203-217.

[2] Kacprowski, J.: Zastosowania analizy i syntezy mowy w telekomunikacji i automatyce. - Rozprawy Elektrotechniczne, 11 /1965/, 2, s. 479-491.

[3] Rosen, G.: Dynamic analog speech synthesizer. - Journ. Acoust.Soc.Amer., 30 /1958/, s. 201-209.

[4] Flanagan, J.L.: Note on the design of "terminal-analog" speech synthesizers. - Journ.Acoust.Soc.Amer., 29 /1957/, 2, s.306-310.

[5] Fant, G.: Acoustic theory of speech production. - Wyd. Mouton and Co., 's-Gravenhage, 1960.

[6] Flanagan, J.L.: Speech analysis, synthesis and perception. - Wyd. Springer-Verlag, Berlin-Heidelberg, 1965.

[7] Sapożkow, M.A.: Sygnał mowy w telekomunikacji i cybernetyce. - Wyd. WNT, Warszawa, 1966.

[8] Kacprowski, J.: Theoretical bases of the synthesis of Polish vowels in resonance circuits. - Wydawnictwo zbiorowe IPPT-PAN p.t.: "Speech analysis and synthesis", tom I, wyd. PWN, Warszawa, 1968, s. 219-287.

[9] Dukiewicz, L.: Polskie spółgłoski nosowe. Analiza akustyczna. - IPPT-PAN, wyd. PWN, Warszawa, 1967.

[10] Kacprowski, J.: An approach to the synthesis of Polish nasal consonants by means of "terminal-analog" speech synthesizer. - Proc.Vibr.Probl., 4 /1963/, 3, s.235-254.

[11] Kacprowski, J.: Synteza polskich spółgłosek nosowych w rezonansowych układach formantowych. - Rozprawy Elektrotechniczne, 9 /1963/, 3, s.439-465.

[12] Jassem, W.: Formants of fricative consonants. - Language and Speech, 8 /1965/, s.1-16.

[13] Fant, G.: Speech analysis and synthesis. - Royal Inst. of Technology, Speech Transm.Lab., Report No 26 /1962/.

[14] Kacprowski, J.: Synteza formantowa dźwięków samogłoskowych i nosowych. Podstawy teoretyczne. - Archiwum Elektrotechniki, 13 /1964/, 3, s.661-676.

[15] Kacprowski, J., Mikiel, W.: Preliminary synthesis of Polish vowels by means of recurrently impulsed formant filters. - Proc.Vibr.Probl., 4 /1963/, 1, s.27-41.

[16] Kacprowski, J., Mikiel, W.: Simplified rules for parametric synthesis of nasal and stop consonants in C-V syllables by means of the "terminal-analog" speech synthesizer. - Acoustica, 16 /1965-1966/, 6, s.356-364.

[17] Mikiel, W., Kacprowski, J., Mikiel, J., Nowicki, J.: Sposób programowania przebiegów napięć i prądów zmiennych oraz urządzenie do stosowania tego sposobu. - Patent PRL nr 25216 z dnia 4.IV.1966.

[18] Holmes, J.N., Mattingly, I.G., Shearme, J.N.: Speech synthesis by rule. - Language and Speech, 7 /1964/, 3, s.127-143.

J. Kacprowski, W. Mikiel
Electroacoustics Laboratory
Department for Vibration Research
IPPT-PAN

THE PROCESS OF SPEECH SYNTHESIS BY MEANS OF THE TERMINAL-ANALOG SYNTHESIZER SYNFOR II

Summary

The experimental speech synthesizer SYNFOR II being the subject of the present paper is a terminal-analog cascade synthesizer and consists of three parallel channels used for the synthesis of vowels and vowel-like sounds, nasal consonants and remaining consonants, viz. fricatives, stops and affricates, respectively. The transfer function of each channel is formed by means of a few conjugated pole circuits which determine the formant frequencies and formant band widths of the speech sounds to be synthesized. The source functions are approximated by a larynx tone generator, viz. a saw-tooth phantastron oscillator and/or by a broad-band noise generator consisting of a semi-conductor diod. The synthesizer as a whole may be controlled by eleven parametric voltage signals varying the following synthesis parameters: the larynx tone frequency /F_0/ and amplitude in the vowel /A_V/ and nasal /A_N/ channel, the noise level in the vowel /A_H/ and in the consonant /A_C/ channel, as well as the pole frequencies in the vowel /F1, F2, F3/, nasal /N2/ and in the consonant /K1, K2/ channel, respectively.

Taking as the starting point the author's former publications in the domain of speech synthesis, the theoretical fundamentals used for the engineering design of the synthesizer, concerning the approximation of the source function

U(s), the transfer function H(s) and the radiation function $Z_p(s)$ of the human organ of speech in electronic lumped-constant circuits, are briefly discussed. In the conclusion the most important electroacoustic parameters of the synthesizer are succinctly described.