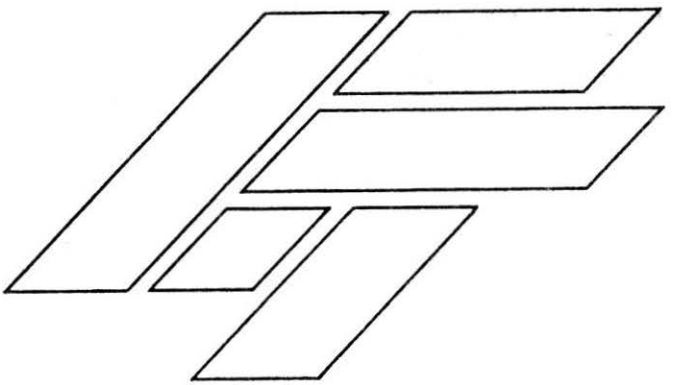


# Problemy Elektroniki i Telekomunikacji

Problemy Elektroniki
i Telekomunikacji /45



WKŁ

prof. dr hab. inż. RYSZARD TADEUSIEWICZ

# Sygnał mowy

Wydawnictwa Komunikacji i Łączności
Warszawa 1988

Opiniodawca:
dr inż. RYSZARD GUBRYNOWICZ
Redaktor:
mgr inż. IZABELA EWA MIKA
Opracowanie graficzne całości:
TADEUSZ PIETRZYK
Redaktor techniczny:
JADWIGA MAJEWSKA
Korekta:
ALICJA KALINOWSKA

*W książce omówiono metody wytwarzania mowy, w tym: budowę traktu głosowego, model procesu wytwarzania mowy, zagadnienia percepcji mowy (model systemu słuchowego, psychologiczne aspekty percepcji mowy), metody opisu sygnału mowy (w dziedzinie czasu i częstotliwości) oraz problemy związane z wytwarzaniem i rozpoznawaniem mowy w automatyce i telekomunikacji.*
*Odbiorcy: inżynierowie elektronicy i studenci.*

534.4



Tytuł dotowany przez
Ministra Nauki i Szkolnictwa Wyższego

ISBN 83-206-0705-1



Wydawnictwa Komunikacji i Łączności, Warszawa 1987
Wydanie 1. Nakład 2150+350 egz.
Ark. wyd. 20. Ark. druk. 17,5 (23,27A)
Oddano do składania we wrześniu 1986
Podpisano do druku w październiku 1987
Papier druk. sat. kl. III, 70 g, 70×100/16
Zamówienie P/87/86. K/9805
Drukarnia im. Rewolucji Październikowej w Warszawie
Zam. 4078/11/87. K-33

# Spis treści

# Od Autora

Książkę napisano opierając się na pracach naukowych prowadzonych w Zakładzie Biocybernetyki Instytutu Automatyki, Inżynierii Systemów i Telekomunikacji Akademii Górniczo-Hutniczej. Tematem tych prac było modelowanie systemu percepcyjnego człowieka, ze szczególnym uwzględnieniem analizatora słuchowego oraz z ukierunkowaniem tych prac na analizę i rozpoznawanie naturalnego sygnału mowy polskiej. Mimo więc podręcznikowego charakteru książki, wskazane wyżej zagadnienia zostały w niej potraktowane obszerniej, a inne problemy ujęto skrótowo. W ten sposób materiał książki uzupełnia dostępne w kraju piśmiennictwo na temat problematyki analizy, syntezy, rozpoznawania i transmisji sygnału mowy o te elementy, które na ogół nie były w tej postaci publikowane. Studiując książkę Czytelnik może i powinien sięgać także do innych publikacji i podręczników, wymienionych na końcu książki, choć opisany materiał, w sensie wiedzy podstawowej, jest kompletny i odpowiada współczesnym poglądom na temat sygnału mowy oraz metod jego analizy i przetwarzania.

Autor poczuwa się do miłego obowiązku podziękowania wszystkim tym Instytucjom i Osobom, które przyczyniły się do powstania książki w jej obecnej postaci. I tak większość badań, referowanych w książce, była koordynowana i finansowana (częściowo) przez Komitet Biocybernetyki Polskiej Akademii Nauk oraz Instytut Biocybernetyki i Inżynierii Biomedycznej

PAN w ramach problemu badawczego nr 06.9.01.5. Pomiary i analizy sygnału dźwiękowego prowadzone były w całości w Instytucie Mechaniki i Wibroakustyki AGH, którego Dyrektorowi, Profesorowi Zbigniewowi Englowi składam tą drogą podziękowanie za wieloletnią, bezinteresowną pomoc w realizacji licznych przedsięwzięć naukowych. Obliczenia komputerowe oraz kreślenie większości rysunków do książki odbywało się w Środowiskowym Centrum Obliczeniowym CYFRONET w Krakowie z wykorzystaniem komputera CYBER 72. Dyrekcji i personelowi tego niesłychanie sprawnie funkcjonującego, nowocześnie zorganizowanego i bardzo sumiennego ośrodka obliczeniowego należą się kolejne wyrazy wdzięczności. Więcej niż kiedykolwiek mogłem oczekiwać bezinteresownej i merytorycznie bezcennej pomocy uzyskałem od polskich uczonych, zajmujących się problematyką analizy i rozpoznawania mowy. Nie jestem w stanie wymienić wszystkich, których rady i inspirująca krytyka wzbogaciła moją wiedzę i pozwoliła mi na podjęcie próby opracowania tej książki, pozwolę sobie zatem wymienić jedynie tych, którym zawdzięczam najwięcej: Profesorów Janusza Kacprowskiego i Wiktora Jassema z Instytutu Podstawowych Problemów Techniki Polskiej Akademii Nauk. Wreszcie muszę podkreślić wielki wkład, jaki w powstanie tej książki wnieśli moi współpracownicy z Zakładu Biocybernetyki AGH: doktorzy Leszek Kot, Andrzej Izworski i Zbigniew Mikrut. Bardzo wielu usterek merytorycznych i niedociągnięć językowych udało się uniknąć dzięki bardzo wnikliwej, krytycznej recenzji dra Ryszarda Gubrynowicza z IPPT PAN.

Wszystkim wymienionym, a także licznym nie wymienionym z powodu braku miejsca pragnę serdecznie podziękować. Wszystko, co jest w tej książce dobre i wartościowe, jest także poniekąd ich dziełem, natomiast pomyłki, jeśli się wkradły, stanowią moją wyłączną winę.

Kraków, czerwiec 1986

# 1

# Wprowadzenie

Istnieją zjawiska, których złożoność przekracza wszelkie wyobrażenie, a które subiektywnie oceniamy jako pospolite i banalne. Dopiero bliższe zbadanie tych zjawisk, a w szczególności próba wykorzystania ich na gruncie techniki, uświadamiają, z jak bardzo skomplikowanym obiektem mamy do czynienia. Do zjawisk omawianej klasy należy m o w a. Doskonałość naturalnego systemu artykulacyjnego, jakim dysponują niemal wszyscy ludzie, powoduje powszechne wrażenie, że proces artykulacji jest łatwy, prosty, naturalny. Tymczasem w rzeczywistości język, wargi i struny głosowe wykonują tysiące ruchów precyzyjniejszych od manipulacji zegarmistrza i szybszych niż ewolucje akrobaty na trapezie. Powstający przy tym zespół dźwięków zawiera mnóstwo różnorodnych informacji. Są wśród nich s e m a n t y c z n e — związane z treścią wypowiedzi, o s o b n i c z e — pozwalające rozpoznać osobę mówiącą, e m o c j o n a l n e — dzięki którym można stwierdzić, że osoba mówiąca jest wzruszona, zdenerwowana lub rozbawiona, a także inne, pozwalające rozpoznać (niekiedy), skąd mówiący pochodzi, jaki jest jego status społeczny, wykształcenie, a także stan zdrowia. Wszystkie wymienione rodzaje informacji można z dźwięku mowy „wyłowić" odpowiednio wprawnym uchem, przy czym proces ten wydaje się subiektywnie jeszcze łatwiejszy niż artykulacja. W rzeczywistości analiza dźwięków mowy, pozwalająca na ich rozpoznawanie i interpretowanie, jest

bardzo złożona. Zakres dynamiki, rozdzielczość częstotliwościowa, szybkość analizy, czułość ucha, wreszcie możliwości uczenia się i dopasowywania do zmiennych warunków — wszystkie te parametry biologicznego analizatora dźwiękowego przewyższają odpowiednie charakterystyki dostępnej obecnie aparatury. Tak więc mowa — zarówno na etapie artykulacji, jak i percepcji i rozpoznawania — jest obiektem bardzo złożonym i trudnym, nasze zaś subiektywne wrażenie prostoty i naturalności procesu komunikacji głosowej jest wynikiem faktu, że w analizę i generację mowy przyroda zaangażowała ogromne fragmenty mózgu, w których zachodzą — bez udziału świadomości — tysiące procesów informacyjnych i regulacyjnych, angażowana jest pamięć, umiejętność uczenia, wreszcie — inteligencja człowieka. Przeniesienie tych czynności na grunt techniki napotyka więc ogromne trudności.

Tymczasem z punktu widzenia techniki mowa, a dokładniej — sygnał mowy, stanowi nader ważny i interesujący obiekt. Jak stwierdzono wyżej, subiektywnie mowa jest najwygodniejszym i najbardziej naturalnym sposobem komunikowania się ludzi. W technice dokłada się więc starań, aby ten najdogodniejszy sygnał optymalnie wykorzystać w systemach komunikacji człowiek — człowiek i człowiek — maszyna. W pierwszym przypadku mamy do czynienia z systemem telekomunikacji, w którym warto sygnał mowy przetworzyć i odpowiednio spreparować, aby przesyłanie wiadomości pomiędzy ludźmi mogło odbywać się bez przeszkód, a równocześnie — możliwie najtaniej. W drugim przypadku interesujące problemy mieszczą się na styku automatyki i informatyki. Sygnał mowy trzeba możliwie najefektywniej kodować i wytwarzać w systemach wykorzystujących komunikację głosową do przekazywania wiadomości od maszyny do człowieka, względnie sygnał mowy trzeba wszechstronnie i precyzyjnie analizować i rozpoznawać w systemach stosujących mowę do przekazywania poleceń człowieka wykonywanych przez maszynę.

Badania nad sygnałem mowy trwają, ale wciąż jeszcze naturalne, biologiczne nadajniki i odbiorniki tego sygnału wyraźnie dominują swymi parametrami nad osiągnięciami techniki. Jest to zresztą naturalne: mowa uformowała się w toku swego rozwoju tak, aby optymalnie wykorzystać ludzkie możliwości percepcyjne i artykulacyjne. Chcąc wkraczać z systemami technicznymi w taki optymalnie dopasowany układ trzeba istotnie wielu badań i wiele pracy. W tych badaniach i pracach konstrukcyjnych liczy się każda głowa i każda para rąk — zwłaszcza że każdy język ma swoje specyficzne cechy i jeśli analizatory i syntezatory mowy polskiej nie powstaną w laboratoriach polskich badaczy i w zakładach doświadczalnych polskich fabryk — to ich nie będzie. Tymczasem odpowiedni poziom rozwiązań systemów analizy, transmisji i syntezy mowy będzie już wkrótce jednym z głównych wyróżników nowoczesności systemów automatyki, komputerów i sieci łączności. O korzyściach, jakie można uzyskać stosując procesory mowy w wymienionych systemach, będzie mowa w treści książki.

Aby jednak prowadzić prace rozwojowe, opracowywać nowe koncepcje i prototypy — trzeba najpierw zgromadzić podstawową wiedzę i poznać

już opracowane systemy. Książka ta ma za zadanie takiej podstawowej wiedzy dostarczyć. Naturalnie nie o wszystkich aspektach sygnału mowy będzie w niej mowa, nie wszystkie nowe koncepcje badawcze uda się opisać, nie wszystkie kierunki rozwoju aparatury znajdą w niej miejsce. Wiedza na temat sygnału mowy i jego przetwarzania jest bowiem dziś ogromna, a jeszcze większa jest literatura poświęcona temu tematowi. Autor wyraża nadzieję, że udało mu się zawrzeć w książce najważniejsze wyniki i najbardziej inspirujące fakty, a poszerzenie wiadomości szczegółowych może nastąpić na podstawie wykazanej na końcu książki literatury oraz nowych artykułów i monografii.

Książka składa się z sześciu w dużym stopniu niezależnych rozdziałów. Dwa pierwsze poświęcono opisom naturalnych, biologicznych procesów artykulacji i percepcji mowy. Wydaje się, że przez ten opis najbardziej płynnie i logicznie można wskazać na te własności sygnału mowy, które można uważać za najważniejsze w systemach technicznych, ponieważ człowiek starannie kształtuje je w procesie artykulacji i rejestruje przy percepcji. Centralnym rozdziałem książki jest rozdział 4, w którym opisano metody stosowane w technice analogowej i cyfrowej do analizy, syntezy i rozpoznawania mowy. Dwa rozdziały kończące książkę wskazują na wybrane zagadnienia szczegółowe, związane z problematyką sygnału mowy w telekomunikacji i w automatyce. Jak się zresztą okazuje, co podkreślono w treści książki, techniki używane przy przekazywaniu sygnału mowy w nowoczesnych systemach komunikacyjnych używanych pomiędzy ludźmi są w generalnych zarysach podobne do systemów komunikacji pomiędzy człowiekiem a urządzeniem technicznym w informatyce lub w systemach automatyki i robotyki. W istocie bowiem sygnał mowy w nowoczesnej telekomunikacji może podlegać tak daleko idącym przekształceniom w urządzeniach nadawczych i odbiorczych, że w istocie można mówić o systemach typu człowiek — maszyna — maszyna — człowiek, a nie o prostym schemacie komunikacji człowiek — człowiek. Teza ta będzie w treści szóstego rozdziału książki szeroko dyskutowana.

# 2

# Wytwarzanie mowy

## 2.1. Uwagi wstępne

Metody przetwarzania i analizy sygnału mowy muszą być oparte na znajomości jego struktury, struktura zaś sygnału w zasadniczy sposób uzależniona jest od jego wytwarzania. Do niedawna wytwarzanie sygnału mowy było domeną systemów naturalnych, to znaczy narządów artykulacyjnych człowieka. Obecnie oprócz naturalnych źródeł sygnału mowy rozważać trzeba także jego wytwarzanie przez systemy techniczne: syntezatory mowy i generatory sygnałów mowopodobnych. Często sztuczne systemy generujące sygnały mowopodobne naśladują naturalny proces artykulacji, zachodzący w trakcie głosowym człowieka. Bywa jednak również często tak, że oszczędniej można uzyskać w systemie technicznym potrzebny sygnał z wykorzystaniem technik opierających się na odtwarzaniu przebiegów czasowych wybranych z naturalnego sygnału mowy i zarejestrowanych — często w sposób bardzo wymyślny — w pamięci systemu generującego.

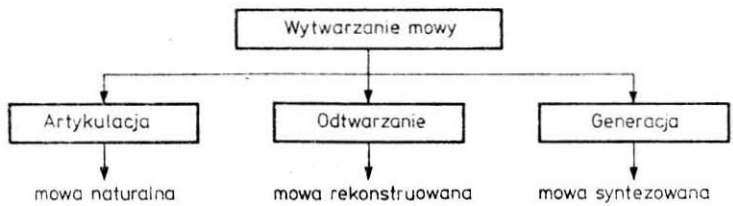
Można więc łącznie wskazać na trzy źródła rozważanego dalej sygnału (rys. 2-1):

— trakt głosowy człowieka, dokonujący artykulacji mowy;

— systemy techniczne o prostej strukturze, dokonujące odtwarzania mowy;

— syntezatory mowy, dokonujące generacji mowy na drodze modelowania procesu artykulacji.

Wzorcem sygnału o cechach stanowiących punkt wyjścia we wszystkich procesach analizy lub sztucznego wytwarzania mowy jest sygnał powstający w wyniku naturalnej artykulacji. Niezbędne jest więc poznanie, dogłębne zbadanie i wszechstronne opisanie traktu głosowego człowieka i zachodzą-



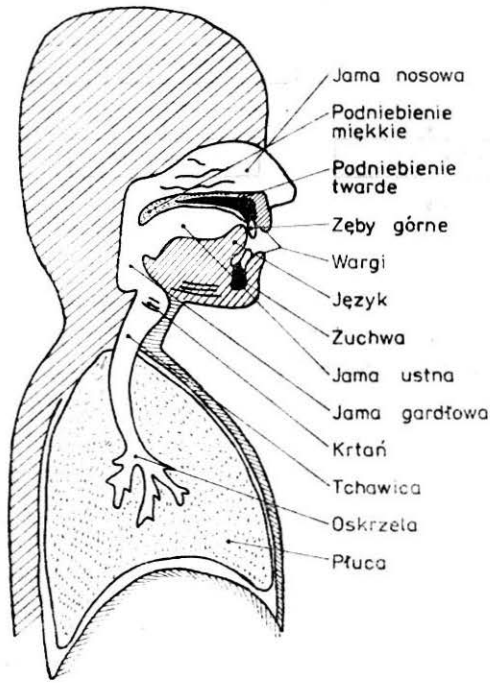

2-1. Metody wytwarzania mowy

cych w nim procesów, aby przy sztucznej syntezie w sposób świadomy i celowy nawiązywać do tych wiadomości, a w procesie analizy poszukiwać skutków poszczególnych operacji towarzyszących naturalnemu wytwarzaniu sygnału mowy. Opis, który jest potrzebny i który będzie omówiony w tym rozdziale, nie będzie takim opisem struktury i czynności traktu głosowego człowieka, jakiego używają anatomowie, fizjolodzy lub lekarze foniatrzy, gdyż inne są cele, dla których będzie on w tej książce wykorzystywany. W wykazie literatury znajdującym się na końcu książki podano pozycje, w których Czytelnik może znaleźć zarówno anatomiczny opis narządów wchodzących w skład traktu głosowego, jak i biologiczny opis ich prawidłowego funkcjonowania oraz typowych patologii. Prezentowany materiał natomiast będzie zawierać próbę opisu struktury i funkcji systemu artykulacji w kategoriach najbliższych Czytelnikom książki, to znaczy w ujęciu matematycznym. Można zatem przyjąć, że w istocie będzie prezentowany pewien model systemu głosotwórczego, uproszczony w stosunku do rzeczywistych zjawisk, lecz eksponujący te struktury i procesy, które decydują o kształcie rozważanego sygnału. Warto ponadto dodać, że jest to model wybrany spośród wielu możliwych, wyselekcjonowany z punktu widzenia maksymalnej zwartości i czytelności opisu, a nie w oparciu o kryterium najdokładniejszego osiągalnego odwzorowania, gdyż taki najdokładniejszy, najwierniejszy, najbardziej rozbudowany model jest wciąż jeszcze przedmiotem sporów naukowych i badań.

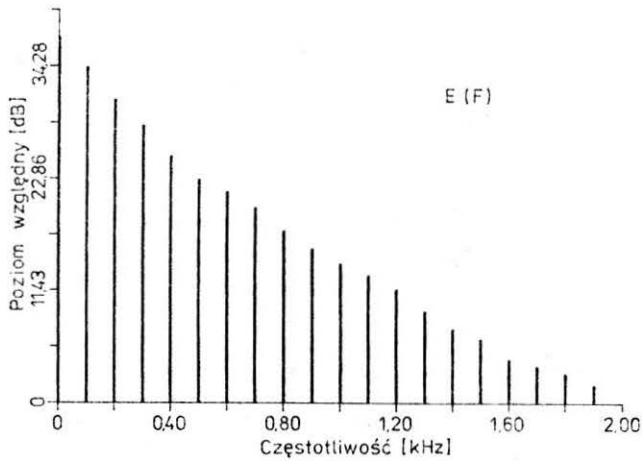
## 2.2. Struktura i czynności traktu głosowego

Ogólna struktura traktu głosowego jest przedstawiona schematycznie na rysunku 2-2. W jego skład wchodzą płuca, dostarczające powietrza do procesu artykulacji, oskrzela i tchawica prowadzące strumień powietrza do krtani, w której drgające struny głosowe są źródłem dźwięku dla dźwięcznych fragmentów mowy. Dźwięk ten jest następnie modulowany we wnękach rezonansowych tworzonych przez język, podniebienie, zęby i wargi. Przy formowaniu tych

wnęk istotną rolę odgrywają ruchy żuchwy i policzków. W przypadku głosek nosowych zamknięta jama ustna pełni rolę bocznika akustycznego, fala dźwiękowa zaś emitowana jest — dzięki odpowiedniemu ustawieniu języczka podniebienia miękkiego — przez jamę nosową i nozdrza.



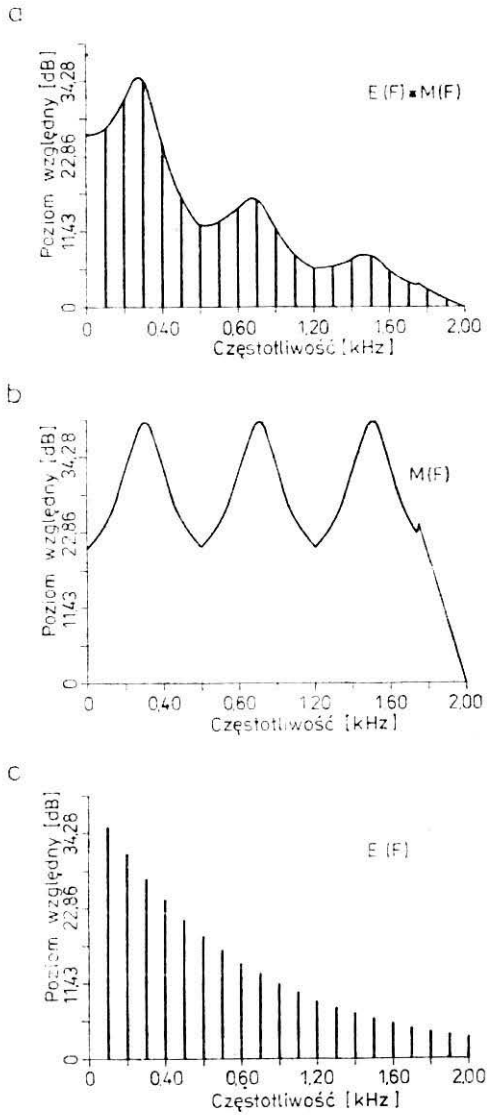

2-2. Uproszczony schemat traktu głosowego (w przekroju)




2-3. Widmo tonu krtaniowego. Malejąca amplituda w zakresie wyższych częstotliwości wymaga na ogół korekty („preemfazy") przy analizie sygnału

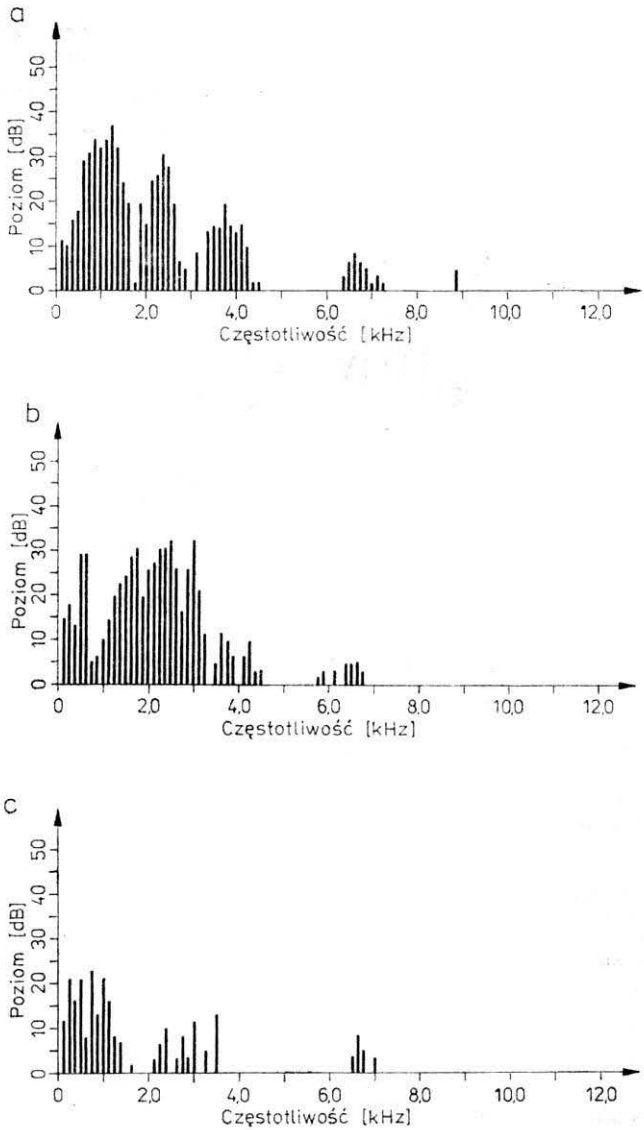
Kształtujące dźwięk rezonanse powstają zarówno w wymienionych wnękach, głównie w jamie ustnej, ale także (chociaż ma to mały wpływ na postać mowy) w klatce piersiowej, w tchawicy i w krtani (szczególnie w tzw. kieszonce Morgagniego pomiędzy strunami głosowymi rzeczywistymi a strunami głosowymi rzekomymi). Wszystkie wskazane rezonatory formują

widmo sygnału krtaniowego, powstającego podczas przetłaczania powietrza między strunami głosowymi. Przepływ powietrza, pobudzając do drgań struny głosowe, powoduje powstanie dźwięku nazywanego t o n e m p o d s t a w o w y m lub k r t a n i o w y m. Ton podstawowy odznacza się




**2-4. Widmo większości głosek powstaje w wyniku modulacji tonu lub szumu generowanego w narządach mowy przez charakterystykę amplitudowo-częstotliwościową traktu głosowego. Na rysunku pokazano (kolejno od góry): widmo samogłoski, charakterystykę amplitudowo-częstotliwościową traktu głosowego i widmo tonu krtaniowego. Niekiedy w procesie tym uczestniczy dodatkowo jama nosowa, a dla spółgłosek szumowych źródłem dźwięku jest szum, a nie ton krtaniowy**

bogatym widmem, w którym wyższe harmoniczne są wprawdzie tłumione z nachyleniem około 12 dB/oktawę, ale mimo to wyraźnie widoczne są nawet harmoniczne o częstotliwości trzydziestokrotnie wyższej od częstotliwości podstawowej (rys. 2-3). Wynikowe widmo określonej głoski dźwięcznej powstaje jako nałożenie charakterystyki traktu głosowego (rys. 2-4), w której poszczególne rezonanse zaznaczone są w postaci maksimów charakterystyki częstotliwościowej, na widmo tonu krtaniowego, w rezulta-

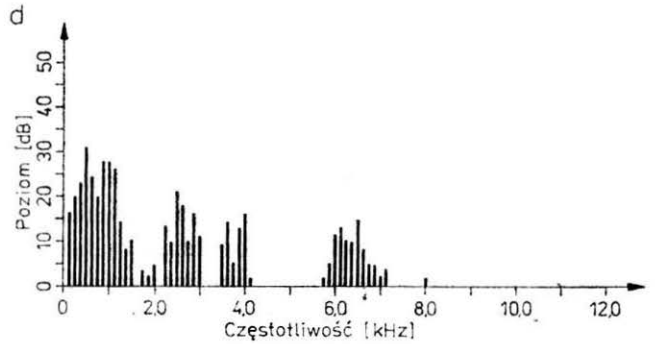


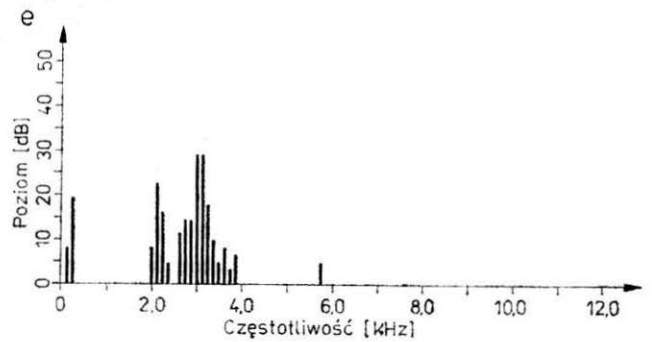

2-5. Widma samogłosek
Rysunek ten, podobnie jak wiele dalszych uzyskano za pomocą komputera Cyber 72, do którego wprowadzono wypowiedzi testowe i rysowano wybrane ich charakterystyki za pomocą autokreślarki Calcomp

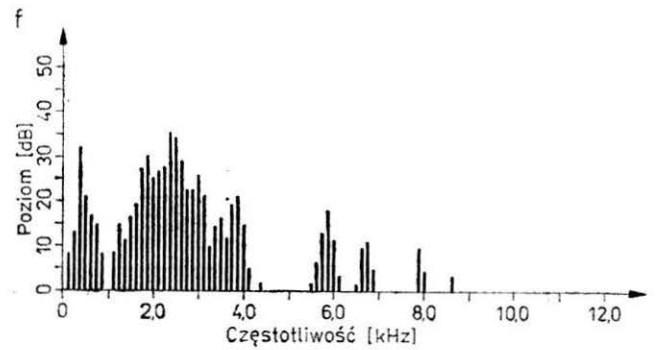
a — widmo głoski *a*,
b — widmo głoski *e*,
c — widmo głoski *u*,
d — widmo głoski *o*,
e — widmo głoski *i*,
f — widmo głoski *y*;
w transkrypcji fonematycznej głoskę (fonem) odpowiadającą polskiej literze *y* zapisuje się zwykle jako przekreślone *i*, na przykład tak ɨ

cie powstaje widmo o kształcie zależnym od konfiguracji narządów mowy w chwili artykulacji danej głoski, odmienne dla każdej głoski i umożliwiające jej identyfikację. Na rys. 2-5 pokazano przykładowo widma samogłosek języka polskiego.

Ton krtaniowy zmienia swą częstotliwość, co jest podstawowym czynnikiem kształtującym intonację wypowiedzi i formuje melodykę głosu — zwłaszcza w śpiewie. Przybliżony zakres tych zmian jest zależny od płci (głosy kobiece mają z reguły dwukrotnie większą częstotliwość tonu krtaniowego niż głosy męskie), wieku (głosy dziecięce są wyższe niż głosy osób dorosłych) i od cech osobniczych (częstotliwość tonu krtaniowego i jej modulacja jest jedną z najważniejszych cech branych pod uwagę przy identyfikacji




osoby mówiącej). Przykładowo można podać zakresy częstotliwości tonu podstawowego dla głosów śpiewaczych:

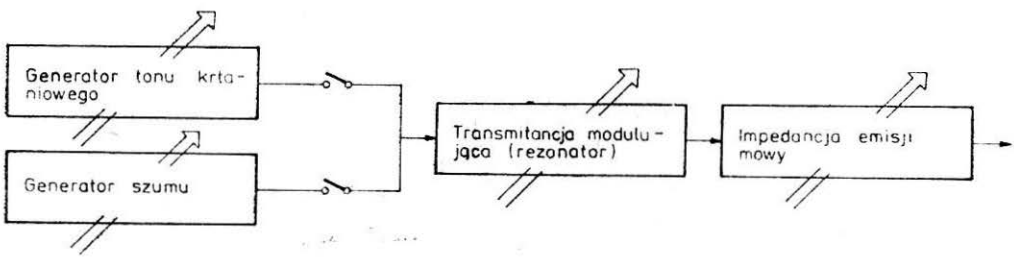
— bas 80÷320 Hz,
— baryton 100÷400 Hz,
— tenor 120÷480 Hz,
— alt 160÷640 Hz,
— mezzosopran 200÷800 Hz,
— sopran 240÷960 Hz.

Są to oczywiście dane uśrednione, indywidualne zakresy głosów śpiewaków mogą nawet dość istotnie odbiegać od podanych granic.

Drgania strun głosowych, będące źródłem omawianego tonu krtaniowego, są

drganiami biernymi. Oznacza to, że powietrze przetłaczane przez szparę głośni, czyli szczelinę między fałdami błony śluzowej, nazywanymi fałdami lub (częściej i mniej dokładnie) strunami głosowymi, wprawia je w drgania na skutek dynamicznego oddziaływania strumienia powietrza i elastycznych fałdów. Odbywa się to bez dodatkowego angażowania mięśni i bez udziału systemu nerwowego. Drgania strun głosowych nie są więc ruchami tego samego rodzaju, jak ruchy warg czy języka; o ich przebiegu bowiem decydują siły aerodynamiczne. System nerwowy ma natomiast możliwość wpływania na parametry układu dynamicznego, w którym drgania zachodzą. Mięśnie i więzadła (opisane dalej bardziej szczegółowo) wchodzące w skład samych strun głosowych, a także ustawiające ruchome sprężyste rusztowanie krtani mięśnie powierzchowne i mięśnie głębokie krtani pozwalają jednak precyzyjnie „stroić" ten drgający układ, zmieniając dowolnie rozwarcie i długość szpary głośni oraz napięcie i grubość (masę) strun głosowych. W ten sposób bierny z fizycznego punktu widzenia proces generacji drgań głosowych w krtani staje się aktywnie sterowanym i precyzyjnie kontrolowanym procesem formowania dźwięków, a intonacja i modulacja głosu, zależna od pracy tych mięśni, jest głównym parametrem pozwalającym na identyfikację osoby mówiącej — zarówno przy kontaktach międzyludzkich, jak i w automatycznych systemach rozpoznających.

Ruchy języka, żuchwy, warg, podniebienia i (w mniejszym stopniu) gardła, formujące wspomniane rezonatory i kształtujące definitywny obraz widma sygnału mowy zachodzą w sposób precyzyjnie sterowany przez odpowiednie elementy systemu nerwowego i są w całości ruchami czynnymi, niekiedy bardzo szybkimi, a niekiedy powolnymi, z płynnym przechodzeniem od stanu do stanu i z doskonałą koordynacją pracy wszystkich zaangażo-




2-6. Schemat zastępczy traktu głosowego
W procesie artykulacji włączane są i wyłączane generatory tonu krtaniowego i szumu (na przemian lub obydwa łącznie), modulowane są charakterystyki generatorów, zmieniany jest kształt toru głosowego, co zmienia transmitancję modulującą sygnał i położenie rezonansów (por. rys. 2-4) a także zmieniana jest impedancja promieniowania ust. Z zewnątrz widoczny jest głównie ten ostatni składnik, to znaczy modulujące impedancję ruchy warg, tymczasem dla formowania sygnału istotniejsze znaczenie mają pozostałe ruchy

wanych mięśni. W schemacie zastępczym traktu głosowego (rys. 2-6) ta część narządów mowy pełni dwojakiego rodzaju funkcje: głównie jest biernym układem filtrów o zmiennych parametrach, formującym transmitancję modulującą sygnał ze źródła dźwięku — na przykład w omawianym wyżej przypadku głosek dźwięcznych tonu podstawowego drgających strun głoso-

wych, jednak obok tej funkcji może być rozpatrywana także jako źródło dźwięku dla głosek szumowych. W tym ostatnim przypadku zamiast (w głoskach bezdźwięcznych, np. s) lub obok tonu krtaniowego (w głoskach dźwięcznych, np. z), źródłem podlegającego formowaniu sygnału dźwiękowego jest szum turbulentnego przepływu powietrza poprzez przewężenia wytworzone przez wymienione narządy.

Ostatnim elementem traktu głosowego jest otwór ust lub/i nozdrza, stanowiący obciążenie omówionego wyżej schematu zastępczego traktu głosowego. Impedancja tego obciążenia jest regulowana przez ruchy artykulacyjne — głównie otwieranie i zamykanie warg, co wpływa dość istotnie na obraz emitowanego sygnału dźwiękowego. Podsumowując, należy stwierdzić, że:

1. Świadoma artykulacja sygnału mowy polega głównie na kształtowaniu parametrów rezonatora, w którym formowany jest sygnał pochodzący ze źródła dźwięcznego lub szumowego.
2. Formowanie, o którym mowa, dotyczy głównie charakterystyk amplitudowo-częstotliwościowych sygnału, gdyż stosunki fazowe kształtowane są między innymi przez drgające biernie struny głosowe, których sterowanie dokonywane jest jedynie przez zmianę parametrów (naprężenia, sztywności, stopnia rozwarcia szpary głośni itp.) lub przez czysto przypadkowy proces generacji szumu w przewężeniach.
3. Model zastępczy systemu artykulacji mowy może być stosunkowo prosty, gdyż składa się jedynie z generatora tonu lub/i szumu o regulowanych parametrach, układu rezonansowego o swobodnie kształtowanej charakterystyce i zmiennej impedancji promieniowania ust lub/i nosa.
4. Artykulacja głosek nosowych polega na propagacji fali dźwiękowej przez kanał nosowy przy bocznikującym wpływie jamy ustnej o zamkniętym wylocie i aktywnie formowanym kształcie.
5. W procesie artykulacji uczestniczą oczywiście również płuca, tchawica, drzewo oskrzelowe i część krtani poniżej strun głosowych, ponieważ jednak nie biorą one udziału bezpośrednio w kształtowaniu wytwarzanego sygnału, przeto w dalszych rozważaniach ich wpływ na brzmienie dźwięku będzie pomijany i ich rola będzie sprowadzana do funkcji źródła energii.

## 2.3. Wybrane szczegóły budowy traktu głosowego i problemy jego sterowania

Omówiona wyżej generalna koncepcja struktury i funkcji traktu głosowego pomijała wiele interesujących szczegółów anatomicznych i fizjologicznych, których poznanie może lepiej zorientować Czytelnika w stopniu złożoności systemu głosotwórczego człowieka i uzmysłowić przybliżony charakter opisanych dalej prób przytoczenia matematycznego modelu tego systemu i przebiegających w nim procesów, a także stopień uproszczenia i zubożenia tego procesu w technicznych syntezatorach mowy.

Mało kto zdaje sobie sprawę, jak wiele mięśni zaangażowanych jest bezpośrednio w proces artykulacji mowy. Pomijając mięśnie oddechowe, których udział w procesie wytwarzania dźwięków jest konieczny, lecz których funkcja biologiczna jest zasadniczo inna, naliczyć można aż 43 mięśnie (w tym znaczna część jest parzysta) bezpośrednio biorące udział w procesie wytwarzania mowy. Są to kolejno:

— mięśnie krtani: pierścienno-tarczowy, pierścienno-nalewkowy tylny, pierścienno-nalewkowy boczny, tarczowo-nalewkowy, głosowy, przedsionkowy, nalewkowy poprzeczny;

— mięśnie gardła: rylcowo-gardłowy, podniebienno-gardłowy, zwieracze gardła — górny, środkowy i dolny;

— mięśnie podniebienia: dźwigacz podniebienia miękkiego, napinacz podniebienia miękkiego, podniebienno-językowy, języczka;

— mięśnie języka: bródkowo-językowy, gnykowo-językowy, rylcowo-językowy, podłużny górny, poprzeczny języka, pionowy języka;

— mięśnie poruszające żuchwę: dwubrzuścowy, żuchwowo-gnykowy, bródkowo-gnykowy, skroniowy, żwacz, skrzydłowy boczny, skrzydłowy przyśrodkowy;

— mięśnie poruszające wargi: okrężny ust, przysieczny górny i dolny, jarzmowy większy, miechowy, dźwigacz wargi górnej, jarzmowy mniejszy, dźwigacz kąta ust, obniżacz kąta ust, obniżacz wargi dolnej, bródkowy, policzkowy;

— mięśnie poruszające nozdrza: nosowy poprzeczny i nosowy skrzydłowy.

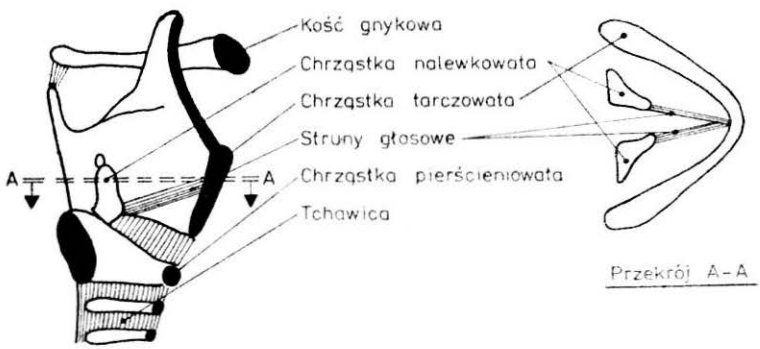
W zestawionym wykazie pominięto mięśnie zaangażowane w proces artykulacji mowy pośrednio (nieodzowne jednak przy jego realizacji), a więc obok wspomnianych już mięśni oddechowych także mięśnie poruszające kość gnykową, stanowiącą nieodzowny punkt zaczepienia dla krtani i mięśni poruszających żuchwę.

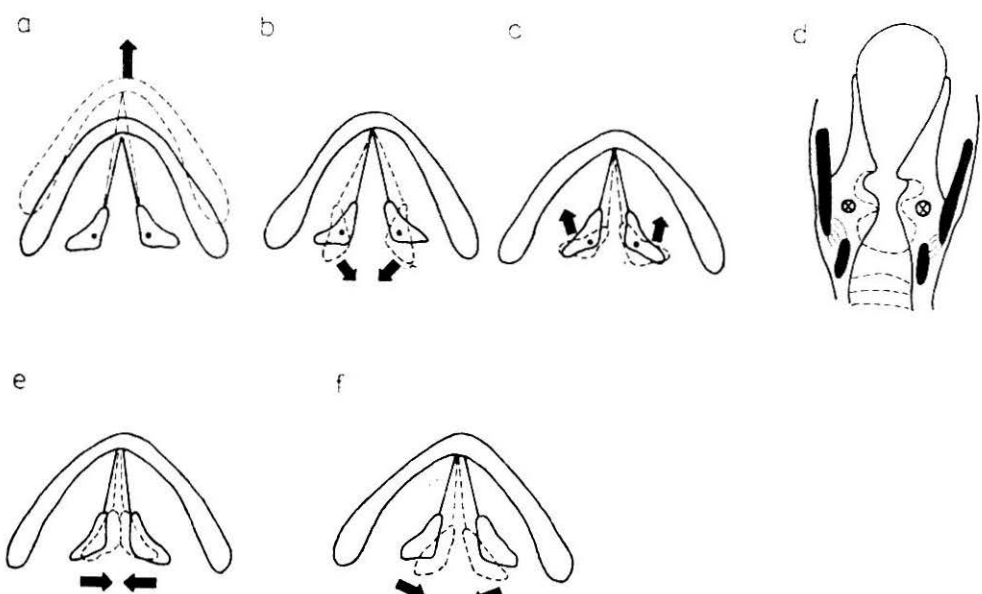
Dokładna dyskusja działania wszystkich wymienionych mięśni jest zbyt obszerna, aby ją tu przytaczać. Warto jedynie — zgodnie z wcześniejszą zapowiedzią — zwrócić uwagę na rolę mięśni krtani w procesie formowania tonu krtaniowego.

Struny głosowe rozpięte są między wewnętrzną powierzchnią kąta chrząstki tarczowatej a wyrostkami głosowymi chrząstek nalewkowatych (rys. 2-7). Działanie mięśni krtani prowadzi do ruchów zarówno chrząstki tarczowatej jak i przemieszcza, obraca, zbliża i oddala chrząstki nalewkowate. W rezultacie szpara głośni jest powiększana i zwężana, a struny głosowe są napinane lub zwalniane, przy czym obecność w samych strunach głosowych dodatkowego mięśnia głosowego powoduje, że mogą one w sposób regulowany zwiększać lub zmniejszać swoją grubość i sztywność.

Skrótowo powyższe procesy opisać można w następujący sposób. Mięsień pierścienno-tarczowy kurcząc się napina cały mechanizm strun głosowych, gdyż oddala chrząstkę tarczowatą od łuku chrząstki pierścieniowatej, na której zamocowane są chrząstki nalewkowate stanowiące punkt zaczepienia strun głosowych (rys. 2-8a). Powoduje to zwiększenie częstotliwości genero-

2-7. Uproszczony schemat przekroju krtani, wskazujący na lokalizację strun głosowych i wzajemne stosunki pozostałych elementów krtani. Zarówno na pionowym przekroju z lewej strony rysunku (przekrój w płaszczyźnie symetrii ciała, widok od prawej strony), jak i na przekroju poziomym z lewej strony (przekrój poziomą płaszczyzną, widok z góry) uwidoczniono fakt, że struny głosowe rozpięte są pomiędzy ruchomo osadzonymi chrząstkami. Zmiana położenia chrząstek napina i zmienia położenie strun głosowych, modulując generowany dźwięk



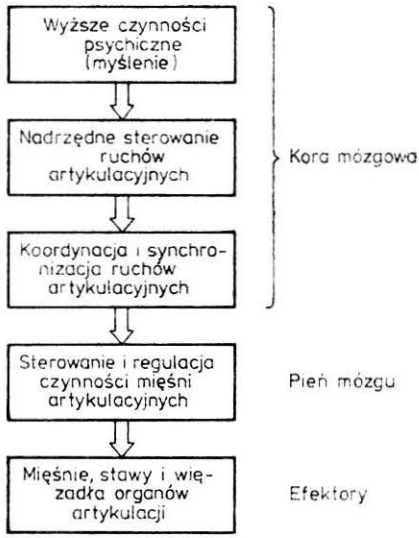




2-8. Procesy zachodzące w trakcie głosowym podczas artykulacji. Działanie mięśni:
**a** — pierścienno-tarczowego, b — pierścienno-nalewkowatego tylnego poszerzającego szparę głośni, c — pierścienno-**-nalewkowatych** bocznych zwierających szparę głośni, d — w wyniku napięcia mięśnia głosowego (co zaznaczono na rysunku **symbolem** ⊗) **struny głosowe** stają się cieńsze i bardziej sztywne; e — nalewkowatego poprzecznego; f — przedsionkowego

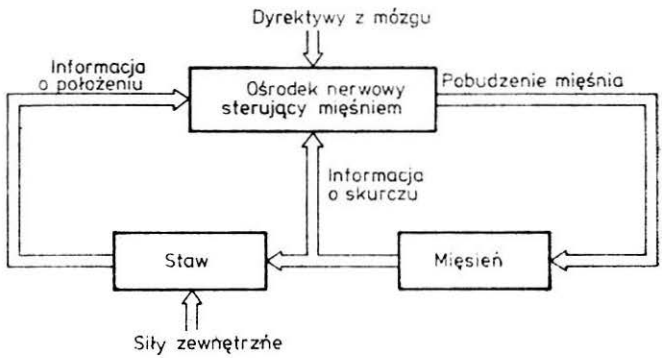
wanego tonu. Mięsień pierścienno-nalewkowy tylny powoduje obracanie chrząstek nalewkowatych i poszerzanie szpary głośni (rys. 2-8b), zaś mięsień pierścienno-nalewkowy boczny powoduje obrót chrząstek nalewkowatych w przeciwną stronę i zwiera szparę głośni (rys. 2-8c). Mięsień tarczowo-nalewkowy zwiera szparę głośni obracając chrząstki nalewkowate do wewnątrz. Część jego włókien przebiegająca bezpośrednio w strunach głosowych, nazywana z tego powodu mięśniem głosowym, napina fałd głosowy po nadaniu wargom głosowym odpowiedniej długości przez mięsień pierścienno-tarczowy, kurcząc się skurczem izometrycznym (bez zmiany długości) i wpływając na sztywność i masę drgających elementów (rys. 2-8c). Na koniec mięsień nalewkowy poprzeczny zbliża do siebie obydwie chrząstki nalewkowate zamykając szparę głośni (rys. 2-8e) a mięsień przedsionkowy zwęża szparę przedsionka, przez co głos staje się przytłumiony (rys. 2-8f). Łatwo zauważyć, że na geometrię i parametry dynamiczne głośni mają wpływ wszystkie wskazane mięśnie, a ich współdziałanie i precyzyjne sterowanie pozwala na sterowanie procesem generacji tonu krtaniowego. Często spotykany pogląd, że modulacja głosu zachodzi pod wpływem działania mięśnia pierścienno-tarczowego, pełniącego funkcję napinacza strun głosowych, musi być w świetle przytoczonej dyskusji oceniony jako bardzo uproszczony.

Naturalnie w podobny sposób można omawiać działanie dalszych mięśni zaangażowanych w proces wytwarzania mowy, utwierdzając się w przekonaniu, że są to procesy nadzwyczaj złożone (szczególnie dotyczy to ruchów języka i warg), wymagające doskonałej koordynacji i dokładnego sterowania. Istotnie, dyskusja dotycząca efektorów mięśniowych realizujących proces artykulacji, omija najistotniejszy i najciekawszy problem — sterowania tego procesu ze strony systemu nerwowego. Zagadnieniem tym zajmiemy się teraz nieco dokładniej.

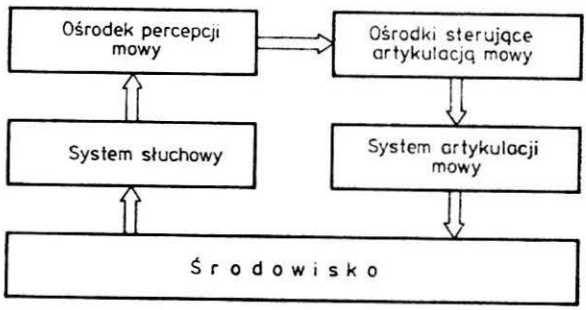
System sterowania procesem wytwarzania mowy jest rozmieszczony we wszystkich tradycyjnie wyróżnianych częściach systemu nerwowego, a więc włącza określone nerwy należące do obwodowego systemu nerwowego, wykorzystuje liczne ośrodki w centralnym systemie nerwowym, na różnych jego piętrach z obszernym fragmentem kory mózgowej włącznie, wreszcie ma liczne wielokierunkowe powiązania z systemem autonomicznym — sympatycznym i parasympatycznym (rys. 2-9). Sterowanie procesem wytwarzania mowy jest zadaniem złożonym i opiera się silnie na działaniu licznych pętli sprzężeń zwrotnych, poczynając od lokalnych układów regulacji stabilizujących pracę pojedynczych mięśni lub kontrolujących położenie poszczególnych stawów (rys. 2-10), a kończąc na globalnym sprzężeniu zwrotnym (rys. 2-11), wykorzystującym analizator słuchowy i sterującym precyzyjnie jakość wytwarzanych dźwięków na drodze bezpośredniej oceny ostatecznego efektu procesu artykulacji. Jest przy tym oczywiste, że to ostatnie sprzężenie zwrotne odgrywa pierwszoplanową rolę w procesie formowania mowy; ogólnie znane trudności z mową ludzi głuchych oraz nowsze doświadczenia, związane z zaburzeniami mowy ludzi znajdujących się w warunkach utrudniających odsłuchową kontrolę własnego głosu (np.




2-9. Schemat czynności poszczególnych pięter ośrodkowego układu nerwowego przy sterowaniu procesem artykulacji mowy




2-10. Ogólna struktura pętli sprzężenia zwrotnego, wiążącej element wykonawczy (mięsień i staw) z elementem sterującym (odpowiedni ośrodek nerwowy). Według tego schematu funkcjonują układy regulacyjne sterujące pracą mięśni zaangażowanych w proces artykulacji mowy




2-11. Struktura globalnego sprzężenia zwrotnego, odgrywającego zasadniczą rolę przy artykulacji sygnału mowy

problem mowy nurka na dużych głębokościach) dostarczają w tym zakresie aż nadto przekonywających dowodów. Warto przy tym zwrócić uwagę na fakt, że dla poprawnego funkcjonowania rozważanego sprzężenia zwrotnego równie ważne jest połączenie akustyczne między narządem głosu a uchem, jak i połączenie nerwowe między analizatorem słuchowym a ośrodkiem sterowania procesem artykulacji mowy. Połączenie takie w mózgu czło-

wieka istnieje, natomiast anatomowie nie mogą odnaleźć jego odpowiednika w mózgu innych zwierząt, w tym także u najbliższych nam naczelnych (małp człekokształtnych). Z tego powodu — prawdopodobnie — zwierzęta wykształciły rozmaite systemy komunikacyjne: „języki" ruchowe, dotykowe, węchowe — ale nie głosowe. Biorąc pod uwagę niewątpliwy wpływ języka na rozwój cywilizacji ludzkiej, możemy nieco fantazjując powiedzieć, że właśnie ten pęczek włókien nerwowych stworzył homo sapiens...

Porzucając jednak hipotezy na rzecz sprawdzonych faktów dokonamy teraz przeglądu ośrodków nerwowych sterujących narządami głosotwórczymi i podejmiemy próbę konstrukcji schematu systemu sterowania procesu artykulacji mowy. Zaczynając od dołu (w sensie hierarchii systemu nerwowego) wymienimy nerwy odpowiedzialne za sterowanie wyszczególnionych wyżej mięśni, zaangażowanych w wytwarzanie sygnału mowy. Konsekwentnie pominiemy przy tym ośrodki nerwowe stwarzające warunki do prawidłowego funkcjonowania narządów mowy, lecz nie sterujące tym funkcjonowaniem w sposób bezpośredni. Chodzi tu głównie o system sterujący procesem oddychania i dostarczający powietrza o wymaganym ciśnieniu do generacji potrzebnych do artykulacji dźwięków (szumu i tonu krtaniowego). W podobny sposób nieodzowne, ale pomijane w rozważaniach są ośrodki układu sympatycznego i parasympatycznego, regulujące wydzielanie śliny, śluzu i płynu surowiczego na powierzchniach błon wyścielających narządy mowy. Nieprawidłowe funkcjonowanie tych ośrodków może prowadzić do nadmiernego przesuszenia lub — przeciwnie, wzmożonej sekrecji, co bardzo utrudnia, a w skrajnych przypadkach może całkowicie uniemożliwić artykulację mowy. Trzeba przy tym pamiętać, że wzmożony wysiłek oddechowy i przepływ powietrza podczas mówienia prowadzą do dodatkowej (w stosunku do normalnej aktywności człowieka) utraty płynów z powierzchni narządów artykulacji mowy, sięgającej 250 ml/godzinę, która musi być kompensowana przez odpowiednie sterowanie procesów wydzielniczych.

Analizując sterowanie samych mięśni uczestniczących w procesie wytwarzania mowy możemy kolejno stwierdzić, że:

— mięśnie krtani są unerwione (sterowane) przez nerw krtaniowy dolny, a mięsień pierścienno-tarczowy przez nerw krtaniowy górny (gałąź zewnętrzną); oba od nerwu błędnego;

— mięśnie gardła są unerwione przez nerw językowo-gardłowy i nerw błędny, przy czym włókna tych nerwów tworzą splot gardłowy, wymieniany także przy omawianiu dalszych mięśni;

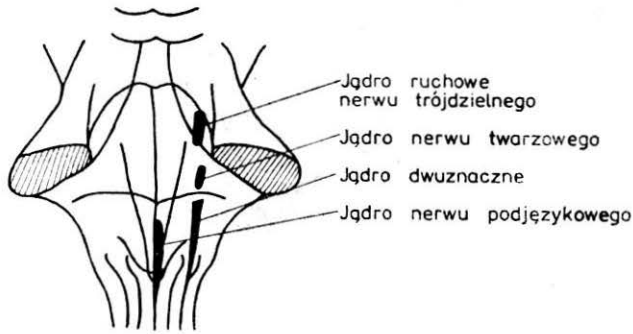
— mięśnie podniebienia są unerwione przez gałązki splotu gardłowego, a ponadto mięsień dźwigacz podniebienia miękkiego przez nerw twarzowy;

— mięśnie języka są unerwione przez nerw podjęzykowy;

— mięśnie poruszające żuchwę unerwione są przez nerw twarzowy, żuchwowo-gnykowy, podjęzykowy, trójdzielny (trzecia gałąź, tzw. nerw żuchwowy), a także — w pewnym zakresie — przez gałązki odchodzące od tzw. pętli szyjnej, tworzonej przez gałęzie brzuszne nerwów rdzeniowych $C_1 - C_3$;

— mięśnie poruszające wargi i nozdrza są unerwiane przez nerw twarzowy.

Wymienione struktury są najniższym piętrem systemu nerwowego i wchodzą oczywiście w skład obwodowego systemu nerwowego. Kolejne piętro stanowią jądra, gromadzące szarą substancję (ciała komórek nerwowych) sterujące pracą wymienionych nerwów (rys. 2-12). Jądra te mieszczą się

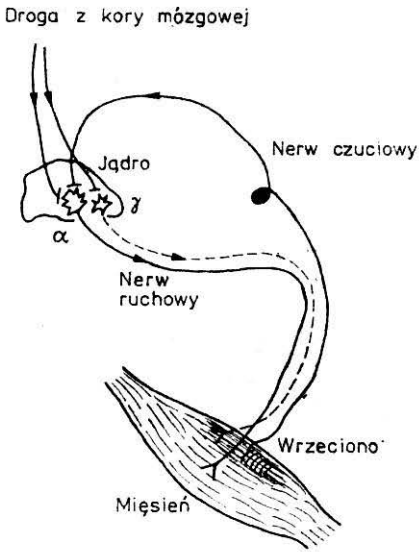



2-12. Lokalizacja w pniu mózgu jąder nerwów czaszkowych odgrywających pierwszoplanową rolę w sterowaniu procesu artykulacji mowy

głównie w pniu mózgu (w dnie komory czwartej — jądro nerwu podjęzykowego) oraz w bocznej części rdzenia przedłużonego (jądro dwuznaczne nerwu językowo-gardłowego i błędnego). W moście (w części grzbietowej dolnego odcinka) mieszczą się jądra ruchowe nerwu twarzowego oraz (w bocznej części środkowego odcinka) jądra ruchowe nerwu trójdzielnego. Nerwy rdzeniowe $C_1 - C_3$ mają odpowiadające sobie skupiska substancji szarej w rogach przednich szyjnych odcinków rdzenia kręgowego.

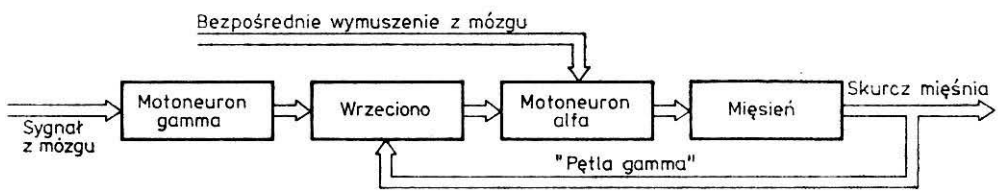
Wymienione jądra stanowią bezpośrednie źródło sygnałów sterujących pracą odpowiednich mięśni i pełnią w stosunku do tych mięśni rolę regulatorów, zapewniających poprawne funkcjonowanie wymienionych mięśni niezależnie od ewentualnego wpływu zakłóceń pochodzących od zmiennych oporów ruchu. Ogólny układ sterowania mięśni można bowiem przedstawić zgodnie ze schematem pokazanym na rys. 2-13, na którym komórki bezpośrednio wymuszające skurcz odpowiednich mięśni (tak zwane motoneurony alfa) znajdują się pod wpływem zarówno bezpośrednich sygnałów sterujących z wyższych pięter systemu nerwowego (omawianych niżej), jak i pod wpływem sygnałów pochodzących z tzw. pętli gamma. W skład pętli gamma wchodzą motoneurony gamma wymuszające skurcz włókien intrafuzalnych („wrzecion") i komórki sygnalizujące stan naprężenia włókna, powstającego w przypadku niezgodności długości włókna intrafuzalnego i całego mięśnia. Patrząc na ten układ z punktu widzenia technika widzimy tu (rys. 2-14) typowy serwomechanizm, w którym pętla gamma pełni rolę zadajnika długości mięśnia, a komórka alfa staje się regulatorem sterującym stanem napięcia mięśnia koniecznym dla uzyskania potrzebnego skrócenia.

Motoneurony alfa i gamma (by pozostać przy tym uproszczonym schemacie), znajdujące się w jądrach odpowiednich nerwów, są sterowane przez korę mózgową. Połączenia pomiędzy korą a wymienionymi jądrami




2-13. Struktura biologicznego regulatora, tak zwane pętli gamma, sterującej pracą pojedynczego mięśnia. Wskazano lokalizację neuronów α pełniących funkcję głównych regulatorów pracy mięśnia oraz neuronów γ służących do zadawania pożądanego stanu napięcia mięśnia. Strukturę połączeń czuciowych, biorących początek we wrzecionie, znacznie uproszczono

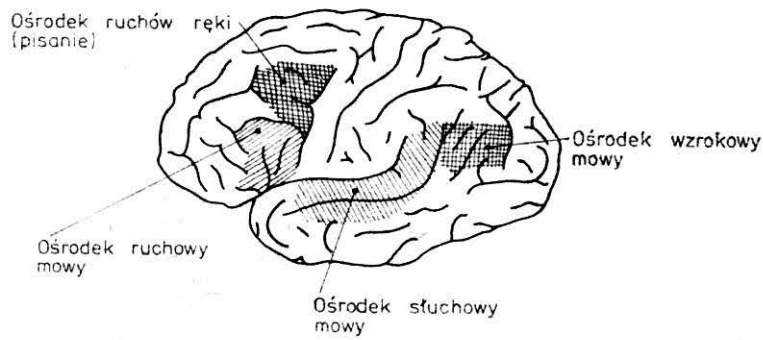
należą do tak zwanej drogi piramidowej, której odgałęzienie docierające do rozważanych jąder jest nazywane drogą korowo-jądrową. Drogi korowo-jądrowe bywają zarówno skrzyżowane, jak i nie, czyli dla większości jąder nerwów czaszkowych obie półkule mózgowe sterują równocześnie mięśniami po obu stronach ciała. Jest to struktura odmienna od występującej dla wszystkich mięśni szkieletowych, dla których istnieje reguła połączeń skrzyżowanych, czyli prawa półkula steruje pracą mięśni lewej części ciała i odwrotnie. Skrzyżowane są w drodze korowo-jądrowej jedynie nerwy prowadzące sygnały do nerwu podjęzykowego i do nerwu twarzowego, co objawia się niekiedy patologicznymi zniekształceniami mimiki twarzy i oczywiście rozważanej tu artykulacji mowy.




2-14. Struktura biologicznego regulatora z rys. 2-13 widziana oczami inżyniera. Taka prezentacja połączenia elementów nerwowych sterujących pracą mięśnia podkreśla ich podobieństwo do technicznych serwomechanizmów

Najważniejszą rolę w omawianym hierarchicznym systemie sterowania procesem wytwarzania mowy odgrywają najwyższe piętra, zlokalizowane w korze mózgowej. To właśnie ich działalność decyduje o możliwości komunikacji głosowej i z tego rejonu wywodzą się impulsy, które sterują pracą jąder nerwów czaszkowych, a za ich pośrednictwem — wszystkimi mięśniami. Obserwacje kliniczne, w których obserwowane u pacjentów ubytki poszczególnych funkcji (w rozważanym przypadku — zubożenie lub całkowity zanik artykulacji mowy) były wiązane z rozpoznawanymi uszkodzenia-

mi określonych rejonów kory mózgowej, pozwoliły na stosunkowo pewną lokalizację obszarów w korze mózgowej. W szczególności z generowaniem i przetwarzaniem informacji językowej wiążą się cztery obszary kory mózgowej (rys. 2-15): ośrodek ruchowy mowy, ośrodek słuchowy mowy, ośrodek dla ruchów ręki (pisanie) oraz ośrodek wzrokowy mowy (ośrodek czytania). Ośrodki te są położone w rejonach, o których od dawna wiadomo, że powiązane są z określonymi funkcjami mózgu. Ośrodek ruchowy mowy i graniczący z nim ośrodek ruchów pisarskich ręki są położone w zwoju



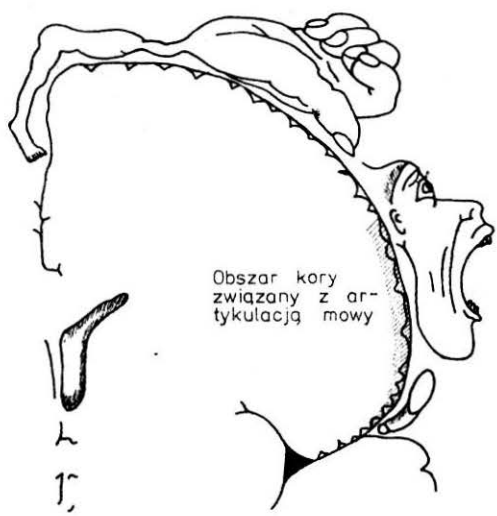

2-15. Lokalizacja w komorze mózgowej struktur, odgrywających pierwszoplanową rolę w procesie generacji mowy. Zwraca uwagę topograficzne powiązanie w korze mózgowej obszarów związanych z artykulacją mowy, jej percepcją słuchową, a także pisaniem i czytaniem

przedcentralnym (sąsiadującym z bruzdą Rolanda), który w całości zawiera korowe ośrodki sterowania ruchem. Ośrodek słuchowy mowy, ulokowany w płacie skroniowym, jest położony w rejonie projekcyjnym wrażeń słuchowych, a ośrodek wzrokowy mowy jest przesunięty w kierunku pól potylicznych, pełniących funkcję korowej reprezentacji wzroku. Wymienione ośrodki sąsiadują ponadto z polami kojarzeniowymi, to znaczy z obszarami kory mózgowej, którym przypisuje się dominującą rolę w procesach myślenia i kojarzenia.

Na szczególną uwagę zasługuje potyliczno-skroniowo-ciemieniowa okolica kojarzeniowa, uważana za nadrzędny ośrodek mowy, któremu trzy uprzednio wymienione mają być podporządkowane. Należy jednak podkreślić, przytaczając i omawiając dane biologiczne na temat lokalizacji ośrodków w korze mózgowej, że wszelkie tego typu informacje są przybliżone i nie mają tak pewnego charakteru, jak uprzednio dyskutowane informacje na temat mięśni, nerwów, czy nawet jąder w pniu mózgu. Kora mózgowa jest zbyt złożonym systemem, aby można było zrozumieć i szczegółowo opisać jej działanie przy użyciu dostępnych metod badań strukturalnych (morfologicznych), czynnościach (fizjologicznych) i obowiązujących koncepcji metodologicznych. Wystarczy wskazać na fakt, że wymienione ośrodki, tak zdefiniowane i zlokalizowane jak to uczyniono wyżej, obejmują około 450 mln komórek nerwowych, których połączeń, czynności i współdziałania niepodobna dziś prześledzić — zwłaszcza że czynności pojedynczej komórki nerwowej wydają się na tyle złożone i zarazem uporządkowane, że wielu badaczy utożsamia ten elementarny fragment systemu nerwowego z mikroprocesorem.

Przechodząc do nieco dokładniejszego omówienia strefy ośrodka ruchowe-

go mowy (ośrodka Broca), który głównie nas interesuje, należy odnotować fakt istnienia w tym obszarze związków między rozmieszczeniem poszczególnych mięśni uczestniczących w artykulacji mowy a ich reprezentacją — w sensie neuronów inicjujących i sterujących ich pracą — w korze



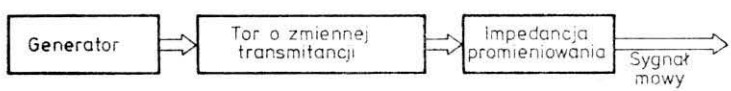

2-16. Tak zwany „humunkulus", obrazujący rozmieszczenie obszarów odpowiadających poszczególnym częściom ciała na powierzchni przedcentralnego zwoju kory mózgowej sterującej jego ruchami. Zwraca uwagę duży obszar zajmowany w korze ruchowej przez neurony sterujące ruchami ręki (zwłaszcza dłoni), a także relatywnie znaczna reprezentacja struktur związanych z czynnością artykulacji mowy. Dowodzi to, jak trudna i precyzyjna jest regulacja mięśni artykulacyjnych. Warto zauważyć przy okazji, jak niewiele miejsca w mózgu zajmuje obszar sterujący pracą mięśni korpusu i nóg, pomimo ich dominującego udziału w wadze całej muskulatury człowieka

mózgowej. Skrótowo można powiedzieć, że im wyżej znajduje się określony mięsień, tym wyżej także zlokalizowane są sterujące go komórki nerwowe. Na tym jednak analogia topograficzna się kończy. Ponadto wielkość reprezentacji poszczególnych mięśni w korze mózgowej absolutnie nie odpowiada rozmiarom samych mięśni, lecz jest raczej pewną miarą ich biologicznej ważności i stopnia szczegółowości sygnałów sterujących, generowanych dla tych mięśni przez korę mózgową. Szkic rozmieszczenia reprezentacji poszczególnych części ciała wzdłuż sterującego ruchami zwoju przedcentralnego kory mózgowej prezentuje wręcz karykaturalnie zniekształcony obraz sylwetki człowieka (rys. 2-16), w którym na uwagę Czytelników zasługują rozmiary obszaru poświęconego artykulacji mowy (zaznaczone na rysunku): uderzające, że mięśnie artykulacyjne, stanowiące wagowo najwyżej 1% masy ciała zajmują blisko 25% komórek nerwowych, sterujących pracą wszystkich mięśni całego ciała. Jest to miara stopnia złożoności ruchów wykonywanych przy wytwarzaniu mowy i wyraz znaczenia, jakie organizm i mózg człowieka przywiązują do tej funkcji.

## 2.4. Model procesu wytwarzania mowy przez człowieka

Przytaczane wyżej opisy anatomii i czynności narządów głosotwórczych człowieka miały charakter zbliżony do formy typowych opisów biologicznych. Z punktu widzenia inżyniera opis taki jest mało czytelny i mało przydatny, nawet jeśli pozbawi się go nadmiaru szczegółów i przedstawi bez odwoływania do hermetycznej łacińskiej terminologii. Aby opis narządów i procesów wytwarzających mowę wykorzystywać przy próbach naśladowania tego procesu w technice — na przykład w syntezatorach lub do optymalizacji procesów przetwarzania mowy w telekomunikacji i cybernetyce — trzeba opis ten przedstawić w formie zwartej i operatywnej zarazem. Idealną formą jest tu model matematyczny, wyróżniający poszczególne systemy i procesy w postaci równań i funkcji, pozwalający przez formalne rozważania wykryć prawidłowości w funkcjonowaniu rozważanego systemu oraz umożliwiający badanie systemu na drodze symulacji komputerowej. Tworzenie modelu i jego formalizacja jest najlepszym sprawdzianem spoistości i kompletności wiedzy biologicznej na temat rozważanego systemu.

W systemie wytwarzania mowy modelowaniu muszą podlegać kolejno: źródło tonu krtaniowego, tor głosowy — ustny i nosowy oraz impedancje promieniowania ust i nosa zamykające odpowiednie tory. Uwzględnianie w modelu płuc, oskrzeli i tchawicy, dostarczających powietrza o wymaganym ciśnieniu podgłośniowym i regulowanym natężeniu przepływu, a także wzbogacających wytwarzany sygnał o dodatkowe rezonanse, nie wydaje się niezbędne. Ich wpływ na proces artykulacji wyraża się wartościami o dwa rzędy wielkości mniejszymi od wpływu elementów wymienionych na wstępie. Tak więc opisany tu model systemu artykulacji ma strukturę jak na rys.



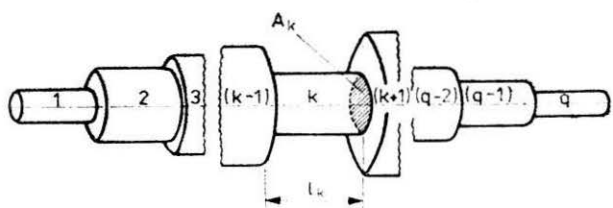

2-17. Uproszczony schemat modelu procesu wytwarzania mowy. Według tego schematu działa system naturalnej artykulacji, jest on jednak również przyjmowany dla sztucznych systemów generacji mowy dla potrzeb automatyki lub telekomunikacji

2-17, będącym odpowiednikiem wcześniej rozważanego schematu z rys. 2-6. Schemat ten tymczasowo nie uwzględnia procesów szumowych odpowiadających artykulacji głosek trących i zwartotrących, których wytwarzanie polega na tworzeniu dodatkowych źródeł szumów położonych wewnątrz traktu głosowego. W uproszczeniu można przyjąć, tak jak to przedstawiono na rys. 2-6, że źródło szumów znajduje się również na wejściu układu o zmiennej konfiguracji i zmiennych parametrach, reprezentującego trakt głosowy. Natomiast uwzględnienie faktu, że źródło szumów znajduje się dalej wzdłuż osi traktu głosowego, polega na dość prostym uzupełnieniu transmitancji układu o zmiennych parametrach, formującego wynikowe widmo sygnału. Zagadnienie to będzie dalej przedstawione bardziej szczegółowo.

Model systemu głosowego człowieka był przedmiotem wielu prac naukowych. Istnieją krytyczno-przeglądowe zestawienia, w których znaleźć także

można ważniejsze pozycje źródłowe. Na szczególną uwagę zasługują wśród nich klasyczne prace Fanta, Flanagana, Ishizaki oraz Kacprowskiego. Naturalnie różni autorzy w odmienny sposób definiują swoje modele, rozmaicie opisują ich elementy, w wyniku czego uzyskują bardziej lub mniej dokładne odwzorowanie rzeczywistych procesów mających miejsce w trakcie głosowym człowieka podczas artykulacji mowy. Podstawowa trudność, jaka przy tym występuje, polega na wyborze racjonalnie prostego modelu, którego struktura zawiera możliwie jak najmniej elementów i odwołuje się do możliwie prostych zależności matematycznych, a który mimo to jest jeszcze stosunkowo wierny. Podana niżej propozycja jest jedną z możliwych. Czytelnik w miarę potrzeb może ten model jeszcze bardziej uprościć, godząc się na mniej wierne odwzorowanie rzeczywistych zjawisk, albo poszukiwać — samodzielnie lub opierając się na cytowanej literaturze — dokładniejszego modelu, z reguły jednak znacznie bardziej złożonego, niestety.

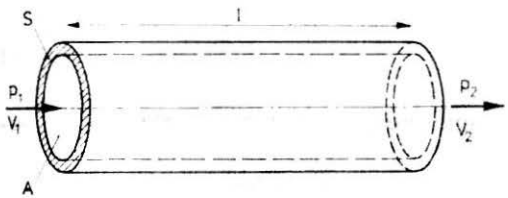
Modele traktu głosowego zazwyczaj są budowane w postaci superpozycji odcinków rur cylindrycznych o sztywnych ścianach, tak dobranych, aby powierzchnia ich przekroju i zmiany średnicy wzdłuż osi symulowanych narządów mowy były zgodne — z założoną dokładnością — z rzeczywistymi wymiarami krtani, gardła, jamy ustnej, warg itd. (rys. 2-18). Oczywiście



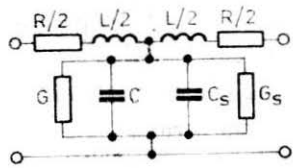

2-18. Uproszczony model traktu głosowego, modulującego sygnał mowy w naturalnym procesie artykulacji

taki model zawiera z założenia niedokładności, a ich przyczyn upatrywać można w przynajmniej trzech istotnych uproszczeniach. Po pierwsze, przekrój rzeczywistych narządów mowy odbiega niemal wszędzie od przekroju kołowego, a to ma wpływ na własności rezonansowe odpowiednich fragmentów traktu głosowego — różniące się w tym przypadku od rozważanego modelu. Po drugie, rzeczywisty kształt narządów mowy zmienia się płynnie i poszczególne przekroje przechodzą płynnie jeden w drugi, bez ostrych granic, natomiast w modelu wprowadza się te granice zniekształcając obraz zjawisk akustycznych i utrudniając wnioskowanie. Po trzecie, ściany rzeczywistych narządów mowy są elastyczne, a nie sztywne, jak to przyjęto w modelu; wymaga to wprowadzenia w modelu dodatkowych elementów stratnych dla uwzględnienia oddziaływania fali akustycznej ze ścianą. W dalszych rozważaniach będziemy więc posługiwali się modelem zbudowanym z odcinków rur cylindrycznych o długości $l$ i polu przekroju poprzecznego $A$ (rys. 2-19); na wejściu rury przyjmujemy istnienie fali akustycznej o ciśnieniu akustycznym $p_1$ i prędkości objętościowej $V_1$. Na wyjściu analogiczne wartości oznaczymy $p_2$ oraz $V_2$. Zgodnie z sugestiami wielu autorów celowe jest zastąpienie układu akustycznego, przedstawionego na rys. 2-19 zastępczym schematem elektrycznym w postaci czwórnika o strukturze podanej

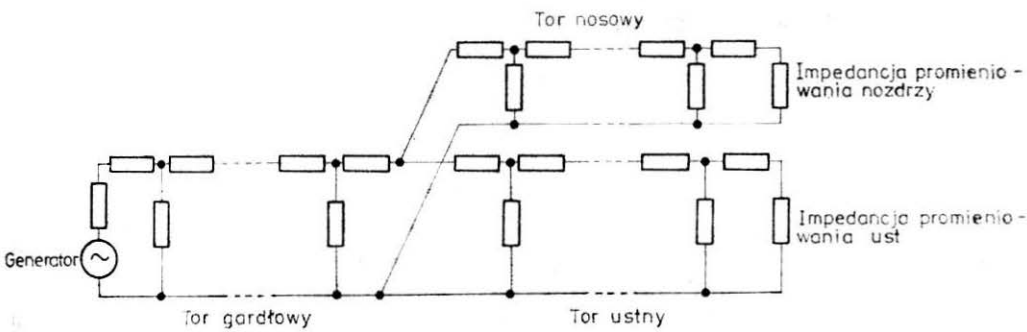
na rys. 2-20. Układ taki, wzorowany na pracach Kacprowskiego, pozwala na zastąpienie modelu akustycznego, mającego postać ciągu rur o zmiennym przekroju, układem łańcuchowo połączonych czwórników, łatwym do obliczeń i analizy (rys. 2-21). Jedynym punktem, w którym zachodzi potrzeba szczegółowszego rozważenia struktury i funkcji schematu zastępczego jest punkt rozgałęzienia toru ustnego i nosowego; zagadnienie to będzie dalej dokładniej zanalizowane.




2-19. Elementarny fragment modelu przedstawionego na rys. 2-18. Element ma zgodny z rzeczywistością przekrój $A$ i długość $l$ — wartości te odpowiadają stosownym parametrom mierzonym w wydzielonym fragmencie aproksymowanego traktu głosowego. Działanie fragmentu modelu można rozważać w kategoriach relacji pomiędzy wejściowymi ($p_1$ i $V_1$) oraz wyjściowymi ($p_2$, $V_2$) parametrami fali akustycznej




2-20. Czwórnik elektryczny stosowany jako analogia elementarnego odcinka modelowanego traktu głosowego, przedstawionego na rys. 2-19. Parametry elektryczne czwórnika mogą być jednoznacznie wyliczone na podstawie parametrów geometrycznych odcinka „rury" z rys. 2-19, przebiegi zaś elektryczne na wejściu i na wyjściu czwórnika stanowią wierną analogię parametrów fali akustycznej transmitowanej przez symulowany przewód




2-21. Ogólna struktura modelu elektrycznego, będącego analogiem układu akustycznego z rys. 2-18 (zastosowano czwórniki postaci podanej na rys. 2-20). W stosunku do wcześniej omówionych modeli traktu głosowego wprowadzono trzy uzupełnienia: uwzględniono rozgałęzienie kanału na tor ustny i nosowy, zastosowano generator wymuszenia krtaniowego oraz dodano zamykające oba łańcuchy impedancje promieniowania — odpowiednio otworu ust oraz nozdrzy. W ten sposób model tu pokazany jest kompletniejszy i bardziej wierny od uprzednio omawianych

Podstawowym elementem modelu jest układ akustyczny z rys. 2-19, którego elektrycznym odpowiednikiem jest czwórnik z rys. 2-20. Ponieważ dane antropometryczne dostarczają wystarczających wiadomości, aby określić parametry rury z rys. 2-19, to kluczową sprawą jest wyznaczenie zależności między wymiarami rury a wartościami parametrów elektrycznych czwórnika. Wprowadzając obok zdefiniowanych wyżej wymiarów, długości $l$ i powierzchni przekroju rury $A$ (wyrażanych odpowiednio w m i $m^2$), obwód otworu rury $S$ [m], gęstość powietrza $\varrho$ [kg $m^{-3}$], prędkość fali dźwiękowej $c$ [$ms^{-1}$],

współczynnik tarcia powietrza $\mu$ [N m$^{-2}$ s], współczynnik przewodności cieplnej powietrza $\lambda$ [m$^{-1}$ s$^{-1}$ K$^{-1}$], ciepło właściwe powietrza przy stałym ciśnieniu $c_p$ [kg$^{-1}$ K$^{-1}$], stałą adiabatyczną $\eta$, rezystancję akustyczną ścian toru głosowego (na jednostkę powierzchni) $r_s$ [kg m$^{-2}$ s$^{-1}$], masę ścian kanału głosowego na jednostkę powierzchni $m_s$ otrzymuje się następujące zależności:

— indukcyjność czwórnika $L$, będąca odpowiednikiem masy akustycznej powietrza zawartego w rurze

$$L = \frac{\varrho l}{A} \quad [\text{kg m}^{-4}] \tag{2.1}$$

— pojemność czwórnika $C$, odpowiadająca podatności akustycznej powietrza w rurze

$$C = \frac{Al}{\varrho c^2} \quad [\text{kg}^{-1}\,\text{m}^4\,\text{s}^2] \tag{2.2}$$

— szeregowa rezystancja czwórnika $R$, odpowiadająca rezystancji strat wskutek wiskotycznego tarcia powietrza przy ścianach rury

$$R = \frac{S}{A^2}\sqrt{\frac{\omega\varrho\mu}{2}}\,l \quad [\text{kg m}^{-2}\,\text{s}^{-1}] \tag{2.3}$$

gdzie: $\omega$ — pulsacja, $\omega = 2\pi f$ ($f$ w [Hz])

— przewodność czwórnika $G$, odpowiadająca akustycznej konduktancji strat wskutek przewodnictwa cieplnego przy ścianach rury

$$G = \frac{S(\eta-1)}{\varrho c^2}\sqrt{\frac{\lambda\cdot\omega}{2c_p\varrho}}\,l \quad [\text{kg}^{-1}\,\text{m}^3\,\text{s}] \tag{2.4}$$

— dodatkowa pojemność czwórnika $C_s$ (ujemna), odpowiadająca odwrotności akustycznej masy drgającej ścian kanału głosowego

$$C_s = -\frac{m_s S}{r_s^2+\omega^2 m_s^2}\,l \quad [\text{kg}^{-1}\,\text{m}^4\,\text{s}] \tag{2.5}$$

— przewodność czwórnika $G_s$, odpowiadająca akustycznej konduktancji strat drgających ścian kanału głosowego, opisana jest zależnością

$$G_s = \frac{r_s S}{r_s^2+\omega^2 m_s^2}\,l \quad [\text{kg}^{-1}\,\text{m}^3\,\text{s}] \tag{2.6}$$

Wstawiając do wzorów (2.1)÷(2.6) konkretne wartości otrzymujemy parametry czwórników modelujących trakt głosowy przy artykulacji określonej głoski. Jest to możliwe, gdyż — co zostanie dalej pokazane — odpowiednie wymiary są znane. Mając parametry poszczególnych czwórników możemy wstawić je do wzorów opisujących — na zasadzie analizy obwodów elektrycznych — i symulować funkcjonowanie całego systemu pokazanego na rys. 2-21.

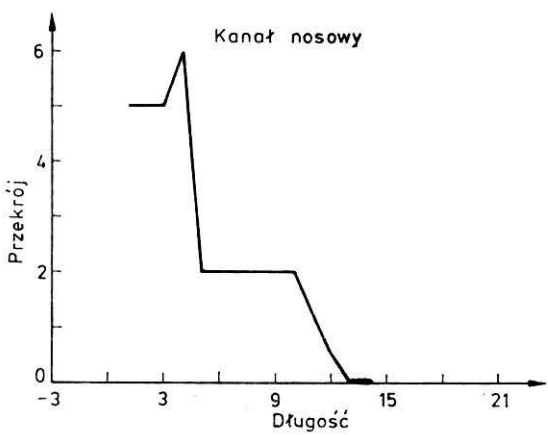
Zestawienie konkretnych danych, pozwalających efektywnie wykorzystać przytoczone wyżej wzory, rozpoczniemy od określenia wartości występujących we wzorach stałych. I tak dla warunków panujących w ustach

$\varrho = 1{,}14\ \text{kg m}^{-3}$
$c = 350\ \text{m s}^{-1}$
$\mu = 1{,}86 \cdot 10^{-5}\ \text{N m}^{-2}\ \text{s}$
$\lambda = 2{,}302 \cdot 10^{-2}\ \text{m}^{-1}\ \text{s}^{-1}\ \text{K}^{-1}$
$c_p = 1{,}005 \cdot 10^{3}\ \text{kg}^{-1}\ \text{K}^{-1}$
$\eta = 1{,}4$
$r_s = 16 \cdot 10^{3}\ \text{kg m}^{-2}\ \text{s}^{-1}$
$m_s = 15\ \text{kg m}^{-2}$

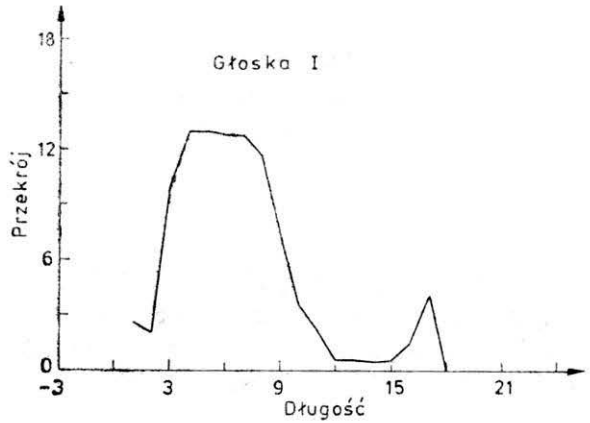
Wybór długości odcinka rury $l$ jest w istocie kompromisem między dokładnością modelu a jego złożonością. Oczywiście korzystne jest wybieranie możliwie najkrótszych odcinków, aby przybliżenie kanału głosowego superpozycją rur było możliwie wierne. Z drugiej jednak strony wzrost złożoności modelu nie pozwala kontynuować tego sposobu przybliżania zbyt długo, gdyż pozostające do dyspozycji środki badania własności modelu — w szczególności dostępne komputery — nie pozwalają na efektywne korzystanie ze zbyt złożonych modeli. Przyjmując jako rozsądne wymaganie, aby model mógł być wykorzystywany do symulacji zjawisk akustycznych o częstotliwościach nie przekraczających $f_{\max} = 5000$ Hz, możemy przyjąć

$$l = \frac{c}{8 f_{\max}} \tag{2.7}$$

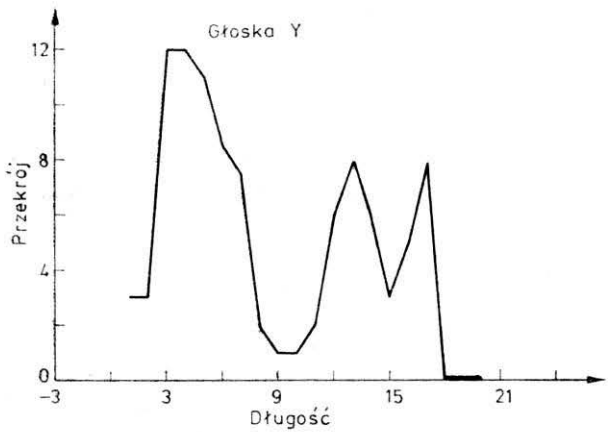
co daje w przybliżeniu wartość $l = 1$ cm, typowo przyjmowaną w tego rodzaju modelach. Oczywiście tak drobna dyskretyzacja przestrzenna potrzebna jest tam, gdzie kształt narządów mowy, a w szczególności ich przekrój zmienia się bardzo szybko; te części, w których przekrój na dużej długości może być uważany za stały, mogą być modelowane jako całość za pomocą pojedynczego segmentu — czyli pojedynczego czwórnika. Taka sytuacja może być brana pod uwagę w przypadku prób — nie uwzględnianych tutaj — modelowania wpływu tchawicy i oskrzeli. Cała tchawica, o długości od rozwidlenia drzewa oskrzelowego do poziomu szpary głośni (średnio około 12 cm), może być traktowana jako jednakowego przekroju



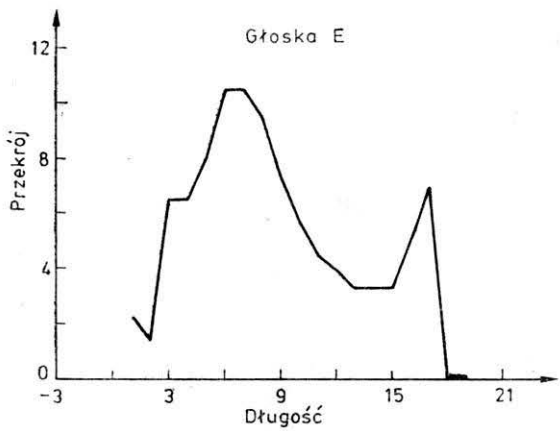

2-22. Zmiana przekroju [cm$^2$] w funkcji długości [cm] dla kanału nosowego, przyjmowana dla modelowania funkcji artykulacji głosek nazalizowanych. Wykres pochodzi z komputerowego systemu modelowania procesu artykulacji mowy i dlatego wartości pośrednie, między sąsiednimi ustalonymi wartościami (por. tekst) są zadane w formie interpolacji liniowej, co nie odpowiada skokowym zmianom przekroju, zakładanym na rys. 2-18



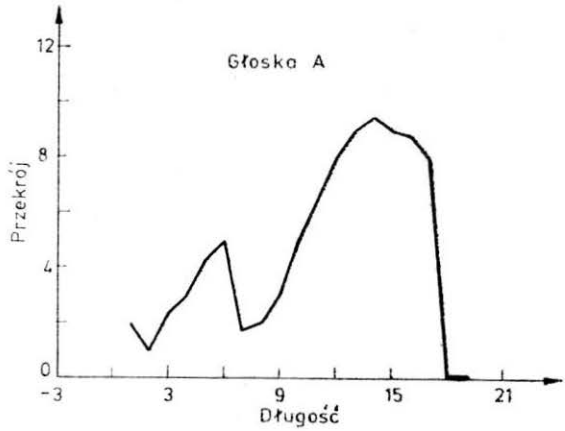

2-23. Zmienność przekroju traktu głosowego (części ustnej i gardłowej) charakterystyczna dla artykulacji głoski *i*



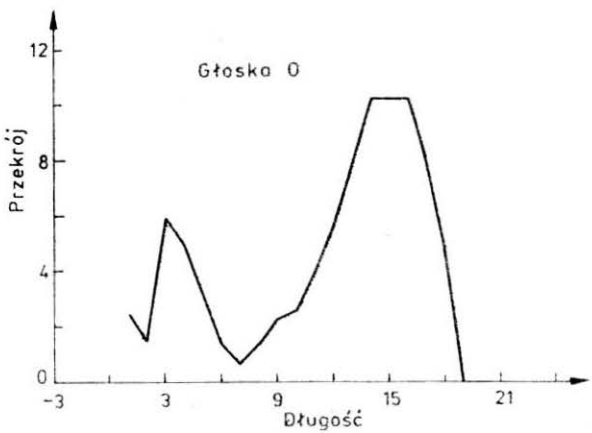

2-24. Przekrój narządów mowy przy artykulacji głoski *y* (w transkrypcji oznaczanej jako *ɨ*). Warto zauważyć różnice między tym rysunkiem a schematem z rys. 2-23 oraz kolejnymi dalszymi rysunkami



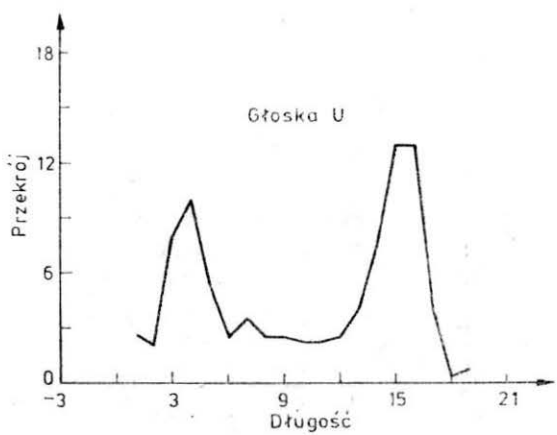

2-25. Zmiana powierzchni przekroju traktu głosowego przy artykulacji głoski *e*




2-26. Zmiana przekroju traktu głosowego przy artykulacji głoski *a*



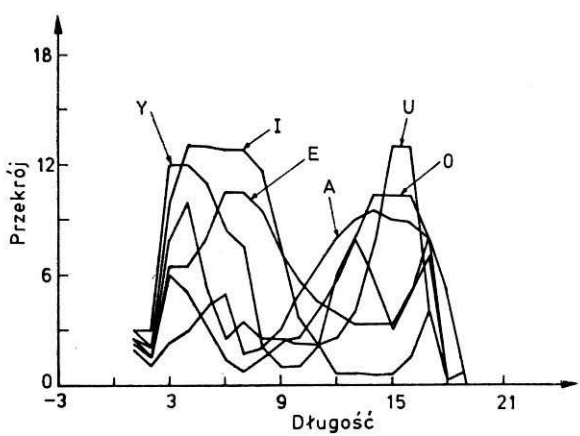

2-27. Zmiana przekroju traktu głosowego w funkcji jego długości dla głoski *o*




2-28. Profil artykulacyjny narządów mowy przy artykulacji głoski *u*

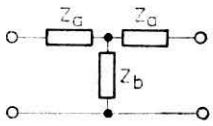
eliptyczna rura o średnicach odpowiednio 20 i 16 mm wzdłuż dłuższej i krótszej osi elipsy. Podobnie oskrzela od rozwidlenia tchawicy do początku rozgałęzienia na elementy drzewa oskrzelowego w płucach mogą być traktowane jako jednorodne rury o długości 3 cm (prawe) i 5 cm (lewe) i przekrojach odpowiednio $18 \times 12$ mm oraz $15 \times 10$ mm. W jamie nosowej również można wyróżnić fragmenty, których przekrój zmienia się na tyle wolno, że przyjmowanie jednocentymetrowego „kwantu" długości jest nieuzasadnione. I tak dla kanału nosowego (12,5 cm długości) można wydzielić aż 6 cm liczącą część centralną, której przypiszemy stałą powierzchnię przekroju, wynoszącą 2 $cm^2$. Pozostały odcinek można podzielić na część tylną, w której początkowy, 3 cm liczący odcinek ma zmienny przekrój z uwagi na ruchomość tylnego języczka podniebienia miękkiego otwierającego i przymykającego ten kanał w trakcie procesu artykulacji. Następny odcinek, o długości 1,5 cm, ma przekrój stały wynoszący 6 $cm^2$. Dalej, idąc ku przodowi napotyka się wymieniony wcześniej, długi na 6 cm, odcinek o stałym przekroju i następnie, przy końcu jamy nosowej, dwa odcinki o przekroju odpowiednio 1,2 i 0,5 $cm^2$, obydwa o długości $l = 1$ cm. Taki opis jamy nosowej sprawdził się w badaniach nad symulowaną komputerowo syntezą mowy i może być przyjęty jako model tego fragmentu traktu głosowego pomimo znacznych uproszczeń w stosunku do rzeczywistej jamy nosowej (rys. 2-22).

Opis pozostałych fragmentów traktu głosowego, a konkretnie kanału gardłowego i kanału ustnego, a także punktu rozwidlenia kanałów: ustnego i nosowego, musi być uzależniony od konkretnej konfiguracji narządów mowy,




2-29. Porównanie (w jednakowej skali) zmienności przekroju narządów artykulacyjnych w funkcji długości traktu głosowego dla wszystkich samogłosek języka polskiego. Widać duże zróżnicowanie przebiegów. Jeszcze większe różnice można zauważyć wprowadzając do rozważań odpowiednie profile także dla spółgłosek, przy czym analiza przebiegów spółgłoskowych jest utrudniona ze względu na to, że dla większości z nich charakterystyczny jest ruch narządów mowy, brak natomiast możliwego do narysowania, ustalonego nieruchomego przekroju — typowego dla samogłosek

związanej z artykulacją danej głoski. Łączna długość wymienionych fragmentów traktu głosowego wynosi od 17 do 19 cm i zależy od stopnia labializacji (wydłużenia warg) przy artykulacji odpowiednich głosek. Długość ta musi być podzielona na odcinki 1 cm długości, gdyż zmienność przekroju narządów mowy przy artykulacji wszystkich głosek jest tu bardzo duża. Na rysunkach 2-23 ÷ 2-28 pokazano obraz zmian przekroju narządów mowy w funkcji długości wzdłuż osi traktu głosowego. Pierwszy odcinek wszystkich wykresów o długości 7 cm odpowiada odcinkowi gardłowemu, następny fragment pośredni długości 1 cm odpowiada rozwidleniu kanałów: nosowego i ustnego, dalszy zaś — odcinkowi ustnemu. Przekroje podane na rys. 2-23 ÷ 2-28 odpowiadają — zgodnie z opisem na rysunkach — artykulacji poszczególnych samogłosek języka polskiego, a na rys. 2-29 pokazano na jednym wykresie, jak bardzo te przekroje się różnią w poszczególnych punktach. Podobne wykresy można sporządzić także dla artykulacji innych głosek, w szczególności szumowych, nosowych i — w mniejszym stopniu z uwagi na istotny udział czynnika ruchu — dla zwartych, drżących itd. Dysponując wymiarami poszczególnych fragmentów traktu głosowego oraz zakładając niezmienność stałych $\varrho$, $c$, $\mu$, $\lambda$ itd. można obliczyć parametry zastępczych czwórników reprezentujących kształt kolejnych odcinków przewodu akustycznego odpowiednio: kanału gardłowego, ustnego i nosowego, a także — gdyby zaszła potrzeba — struktur podgłośniowych: tchawicy i oskrzeli. Zastępcze impedancje podłużne i poprzeczne czwórnika




2-30. Struktura czwórnika zastępczego w konfiguracji T. Czwórnik taki zastępuje w analizie matematycznej odpowiednie fragmenty traktu głosowego, dzięki czemu możliwe staje się opisywanie funkcjonowania narządów artykulacyjnych za pomocą bardzo rozwiniętego i dobrze znanego aparatu matematycznego teorii obwodów

w konfiguracji $T$ (rys. 2-30) można wyliczyć przy założeniu, że w torze głosowym rozchodzi się fala płaska — co istotnie jest spełnione w przypadku wyboru długości odcinków zastępczych zgodnie ze wzorem (2.7). W tym celu wygodnie jest wprowadzić i obliczyć najpierw impedancję charakterystyczną (falową) każdego czwórnika, korzystając ze wzoru

$$Z_0 = \sqrt{\frac{R+j\omega L}{G+G_s+j\omega(C+C_s)}} \tag{2.8}$$

Warto zauważyć, że impedancja $Z_0$ nie zależy od długości aproksymowanego odcinka rury $l$, natomiast od długości tej zależy tamowność falowa $\gamma$ dana wzorem:

$$\gamma = \sqrt{(R+j\omega L)[(G+G_s)+j\omega(C+C_s)}} \tag{2.9}$$

Parametry czwórnika $T$ wyraża się za pomocą impedancji charakterystycznej $Z_0$ i tamowności falowej $\gamma$ w sposób szczególnie prosty, gdy spełniony jest wspomniany warunek małych rozmiarów odcinków zastępczych rur. Wówczas występujące we wzorach na impedancję podłużną $Z_a$ i poprzeczną $Z_b$ funkcje hiperboliczne można zastąpić ich argumentami:

$$Z_a = Z_0 \operatorname{tgh}(\gamma/2) \approx Z_0\gamma/2 = \frac{R+j\omega L}{2} \tag{2.10}$$

$$Z_b = \frac{Z_0}{\sinh\gamma} \approx \frac{Z_0}{\gamma} = \frac{1}{G+G_s+j\omega(C+C_s)} \tag{2.11}$$

Według przytoczonych wzorów można obliczać — wstawiając odpowiednie parametry ze wzorów (2.1)÷(2.6) — parametry wszystkich czwórników łańcucha, zarówno modelujących tor gardłowy, jak i tor nosowy i ustny. Jedyny punkt, w którym wymagana jest pewna uwaga, dotyczy miejsca rozwidlenia kanałów: nosowego i ustnego. Można przyjąć, że w tym punkcie czwórnik modelujący kolejny — ósmy licząc od otworu głośni — odcinek traktu głosowego musi mieć parametry skorygowane w stosunku do wartości wynikających z wymiarów geometrycznych, gdyż jego dane muszą uwzględniać bocznikujący wpływ impedancji wejściowej kanału nosowego $Z_n$. Skorygowane parametry tego czwórnika można wyliczyć ze wzorów:

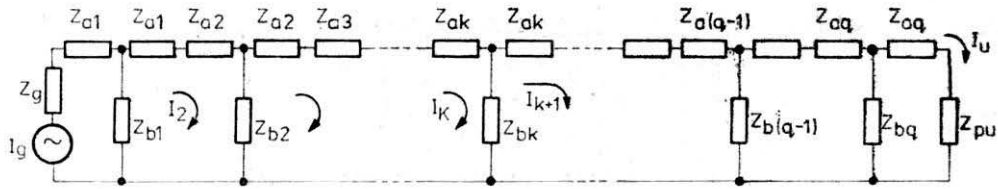
$$Z_a = Z_0 \frac{Z_0 + Z_n\dfrac{\gamma}{2}}{2(Z_0+Z_n\gamma)} \tag{2.12}$$

$$Z_b = \frac{Z_0}{\gamma}\,\frac{Z_n}{Z_n+Z_0/\gamma} \tag{2.13}$$

gdzie wzór na określenie impedancji $Z_n$ będzie podany dalej (2.32).

Dysponując modelem torów: gardłowego, ustnego i nosowego w postaci połączonych łańcuchowo czwórników o omówionej wyżej strukturze i parametrach możemy, korzystając z teorii obwodów elektrycznych, określić transmitancję tych łańcuchów, a tym samym podać opis matematyczny najważniejszego elementu modelu. Transmitancje te wygodnie jest określić osobno dla toru gardłowo-ustnego, a osobno dla toru nosowego.

Zgodnie z przyjmowanym w modelu systemem analogii akustyczno-elektrycznych napięciom w poszczególnych węzłach układu odpowiadają ciśnienia akustyczne, a natężeniom prądu w gałęziach — prędkości objętościowe fali akustycznej. Źródło tonu krtaniowego, które w naszym modelu pełni funkcję elementu wymuszającego, dostarcza odpowiednio zmiennej w czasie fali o dającej się obliczać prędkości objętościowej $V_g(t)$. W modelu odpowiadać temu będzie źródło prądowe o natężeniu $I_g(t)$. Krtań charakteryzuje się także określoną impedancją $Z_g$, której określeniu poświęcone będzie dalej nieco miejsca. Opisane źródło zasila układ, wywołując przepływ prądów oczkowych (rys. 2-31). Ostatni czwórnik jest zwierany przez impedancję




2-31. Wykorzystujący czwórniki zastępcze typu T układ zastępczy, wykorzystywany przy określaniu transmitancji kanału gardłowo-ustnego. Łańcuch czwórników zasilany jest (po lewej stronie) ze źródła prądowego $I_g$ o impedancji szeregowej $Z_g$, zawiera $q$ oczek zbudowanych z czwórników zastępczych i domknięty jest impedancją promieniowania ust $Z_{pu}$

obciążenia $Z_{pu}$ (impedancja promieniowania ust). Stosunek natężenia prądu w tej impedancji $I_u$, odpowiadający w przyjętej skali prędkości objętościowej fali w otworze ust $V_u$, do odpowiednich wartości $I_g$ oraz $V_g$ na wysokości głośni stanowi interesującą nas transmitancję. Oznaczając zatem transmitancję toru gardłowo-ustnego przez $H_{gu}(j\omega)$, możemy więc zapisać:

$$H_{gu}(j\omega) = \frac{V_u(j\omega)}{V_g(j\omega)} = \frac{I_u(j\omega)}{I_g(j\omega)} \tag{2.14}$$

Równania Kirchhoffa dla obwodu o schemacie podanym na rys. 2-31 można zapisać w postaci

$$Z_{11}I_1 + Z_{12}I_2 = Z_gI_g \tag{2.15}$$

$$Z_{21}I_1 + Z_{22}I_2 + Z_{23}I_3 = 0 \tag{2.16}$$

$$Z_{32}I_2 + Z_{33}I_3 + Z_{34}I_4 = 0 \tag{2.17}$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$Z_{(q-1)q}I_{q-1} + Z_{qq}I_q = 0 \tag{2.18}$$

Liczba równań wynika z liczby oczek utworzonych przez $q-1$ czwórników odwzorowujących tor gardłowo-ustny. Oznaczenie $I_k$ użyte we wzorach (2.15)÷(2.18) oznacza prąd oczkowy w $k$-tym oczku ($k = 1, 2, \ldots, q$), impedancje zaś własne oczek $Z_{kk}$ oraz wzajemne $Z_{(k-1)k}$ i $Z_{k(k+1)}$ są sumami odpowiednich impedancji czwórników:

$$Z_{kk} = Z_{a(k-1)} + Z_{b(k-1)} + Z_{ak} + Z_{bk} \tag{2.19}$$

$$Z_{(k-1)k} = Z_{k(k-1)} = -Z_{b(k-1)} \tag{2.20}$$

$$Z_{k(k+1)} = Z_{(k+1)k} = -Z_{bk} \tag{2.21}$$

gdzie indeksy $k-1$, $k$, $k+1$ są numerami odpowiednich czwórników, których impedancje $Z_a$ i $Z_b$ są uwzględniane w sumie. Przyjęto zasadę numeracji, że gałąź poprzeczna $k$-tego czwórnika, mająca impedancję $Z_{bk}$, stanowi granicę pomiędzy oczkami o numerach $k$ oraz $k+1$. Dla skrajnych oczek można zapisać

$$Z_{11} = Z_{a1} + Z_{b1} + Z_g \tag{2.22}$$

oraz

$$Z_{qq} = Z_{a(q-1)} + Z_{b(q-1)} + Z_{pu} \tag{2.23}$$

W dalszych rozważaniach wygodnie będzie rozważać układ równań (2.15)÷(2.18) w postaci macierzowej:

$$\mathbf{ZI} = \mathbf{U} \tag{2.24}$$

gdzie macierz Z ma postać trójdiagonalną:

$$\mathbf{Z} = \begin{bmatrix} Z_{11} & Z_{12} & 0 & 0 & 0 & \dots & 0 & 0 \\ Z_{21} & Z_{22} & Z_{23} & 0 & 0 & \dots & 0 & 0 \\ 0 & Z_{32} & Z_{33} & Z_{34} & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & 0 & \dots & Z_{(q-1)q} & Z_{qq} \end{bmatrix} \tag{2.25}$$

a wektor wymuszeń U ma postać dogodną do obliczeń, gdyż jedynie pierwszy jego element jest niezerowy. Z równań (2.24) można bez trudu wyliczyć prąd w dowolnym oczku wzoru

$$I_k = \frac{\Delta_{1k}}{\Delta} Z_g I_g \tag{2.26}$$

gdzie $\Delta$ jest wyznacznikiem głównym macierzy Z, a $\Delta_{1k}$ jest jej podwyznacznikiem względnym (kofaktorem) dla elementu $Z_{1k}$. Transmitancję (2.14) można teraz wyznaczyć bez trudu opierając się na fakcie, że $I_u = I_q$ a także zakładając dla uproszczenia, że $I_g \approx I_1$. Wówczas

$$H_{gu}(\mathrm{j}\omega) = \frac{\Delta_{1q}}{\Delta_{11}} \tag{2.27}$$

i może być dla każdej konfiguracji przestrzennej obliczona na podstawie przytoczonych wyżej wzorów i wymiarów traktu głosowego.

W identyczny sposób można przeprowadzić obliczenia dla kanału nosowego, którego macierz impedancyjna $\mathbf{Z}_N$ jest prostsza i zawiera elementy odpowiadające jedynie sześciu oczkom:

$$Z_N = \begin{bmatrix} Z_{11} & Z_{12} & 0 & 0 & 0 & 0 \\ Z_{21} & Z_{22} & Z_{23} & 0 & 0 & 0 \\ 0 & Z_{32} & Z_{33} & Z_{34} & 0 & 0 \\ 0 & 0 & Z_{43} & Z_{44} & Z_{45} & 0 \\ 0 & 0 & 0 & Z_{54} & Z_{55} & Z_{56} \\ 0 & 0 & 0 & 0 & Z_{65} & Z_{66} \end{bmatrix} \tag{2.28}$$

Warto zwrócić dodatkowo uwagę, że liczba oczek dla kanału nosowego jest stała, w przeciwieństwie do liczby oczek modelu toru gardłowo-ustnego, która musiała być traktowana jako zmienna i oznaczana przez $q$ z uwagi na zmienną długość toru gardłowo-ustnego przy artykulacji różnych głosek.

Struktura równań opisujących tor nosowy różni się także i tym od struktury dla toru gardłowo-ustnego, że wymuszenie w torze nosowym ma charakter ciśnieniowy (w analogu elektrycznym — napięciowy) i odpowiada różnicy potencjałów na impedancji $Z_b$ czwórnika, który modeluje rozgałęzienie torów ustnego i nosowego. W rozważanym modelu rozgałęzienie przypada w miejscu odpowiadającym położeniu czwórnika nr 8, którego parametry były wcześniej dyskutowane — por. wzory (2.12) i (2.13). Tak więc zapisując dla toru nosowego równania analogiczne do (2.24)

$$\mathbf{Z}_N \mathbf{I}_N = \mathbf{U}_N \tag{2.29}$$

musimy jako wektor wymuszeń $\mathbf{U}_N$ przyjąć wektor o pierwszej składowej wyliczonej ze wzoru

$$U_8 = Z_{b8}\left[\frac{\Delta_{18}}{\Delta} - \frac{\Delta_{19}}{\Delta}\right] Z_g I_g \tag{2.30}$$

i pozostałych składowych wynoszących zero.

Rozwiązując układ równań (2.29) można określić w sposób analogiczny do wyżej opisanego transmitancję kanału nosowego ze wzoru

$$H_n(j\omega) = \frac{\Delta_{16}(N)}{\Delta_{11}(N)} \tag{2.31}$$

gdzie indeks $N$ oznacza, że odpowiednie wyznaczniki są obliczane dla macierzy $\mathbf{Z}_N$, a nie jak uprzednio $\mathbf{Z}$. Rozwiązując równania (2.29) można bez trudu wyznaczyć potrzebną nam wcześniej wartość impedancji wejściowej toru nosowego $Z_n$ (por. wzory (2.12) i (2.13)). Wartość ta może być wyznaczona ze wzoru

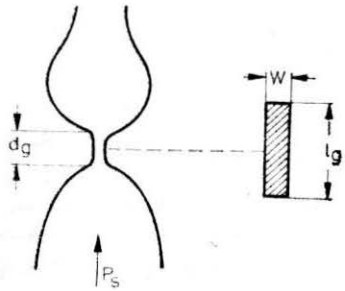
$$Z_n = \frac{U_8}{I_1(N)} = \frac{\Delta(N)}{\Delta_{11}(N)} \tag{2.32}$$

Przy obliczaniu transmitancji kanału nosowego można zwykle uprościć strukturę czwórnika zastępczego pomijając w nim składniki $G_s$ i $C_s$, gdyż ściany jamy nosowej są na ogół znacznie bardziej sztywne niż ściany gardła czy jamy ustnej. Impedancja promieniowania nozdrzy, zwierająca na końcu łańcuch czwórników, może być obliczona podobnie jak impedancja promieniowania ust ze wzorów podanych dalej.

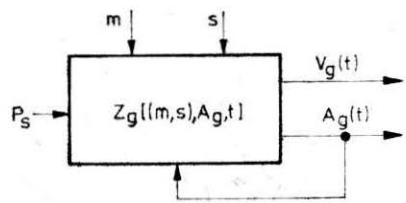
Etap budowy modelu, polegający na określeniu transmitancji układów o zmiennej strukturze, odpowiadających kanałom: gardłowemu, ustnemu i nosowemu, jest najtrudniejszym zadaniem w tworzeniu opisu matematycznego procesu naturalnej artykulacji. W celu opisania całości modelu musimy rozważyć strukturę generatora tonu krtaniowego oraz parametry impedancji promieniowania ust i nozdrzy. Krtań (struny głosowe) jest generatorem aerodynamicznym, którego drgania są warunkowane parametrami mechanicznymi (masa, sztywność, rezystancja strat), geometrycznymi (szerokość i konfiguracja głośni) oraz przepływem powietrza (ciśnieniem podgłośniowym, obciążeniem, impedancją wejściową toru głosowego). Rozmiary geometryczne szpary głośni są na ogół przyjmowane (rys. 2-32) w sposób następujący: długość 18 mm, głębokość (grubość fałdu głosowego) 3 mm, powierzchnia otworu — zmienna od 0 do ok. 20 mm$^2$. Schemat funkcjonowania głośni może być dany jak na rys. 2-33. Czynnikiem bezpośrednio wymuszającym drgania jest podgłośniowe ciśnienie powietrza $p_s$. Powietrze przetłaczane przez szczelinę głośni o powierzchni $A_g$ wprawia w drgania struny głosowe, w wyniku czego powierzchnia szpary głośni zmienia się w czasie w przybliżeniu w ten sposób, że czasowy przebieg $A_g(t)$ tworzy serię quasi-periodycznych impulsów trójkątnych o czasie narastania $\tau_1$, czasie opadania $\tau_2$ i okresie $T_0 > \tau_1 + \tau_2$. Parametr $T_0$, będący odwrotnością częstotliwości podstawowej $F_0$, zmienia się wraz ze zmianami parametrów układu drgającego, przy czym oznaczając masę drgających strun

przez $m$ (rys. 2-34) oraz ich sztywność przez $s$ możemy w przybliżeniu zapisać

$$F_0 = \frac{1}{2\pi}\sqrt{\frac{s}{m}} \tag{2.33}$$

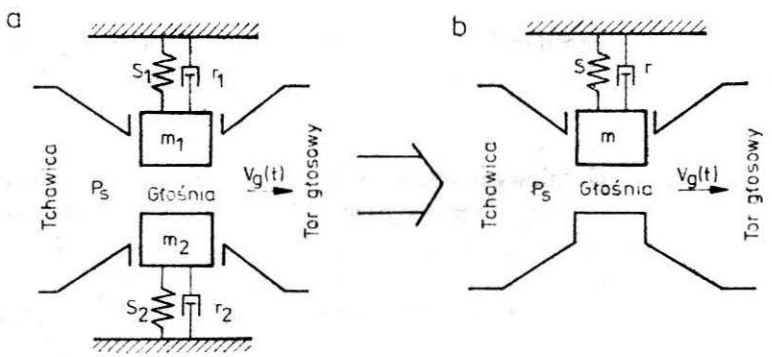



2-32. Uproszczony rysunek przekrojowy krtani. Struny głosowe są traktowane jako zwężenie na drodze przepływu powietrza przepychanego pod ciśnieniem podgłośniowym $p_s$, przy czym ignorując złożone stosunki przestrzenne rzeczywistych strun głosowych opisuje się je jako prostokątną szczelinę o długości $l_g$, szerokości $W$ (zmiennej w czasie!) oraz grubości $d_g$. Podstawowym parametrem uwzględnianym w dalszej analizie jest zakreskowana powierzchnia szpary głośni $A_g$, oczywiście zmieniająca się w czasie




2-33. Schemat blokowy źródła krtaniowego. Ciśnienie podgłośniowe $p_s$ wymusza przepływ powietrza o prędkości objętościowej $V_g$, zależnej od impedancji przepływu $Z_g$. Impedancja ta zależy głównie od powierzchni przekroju szczeliny głośni $A_g$, która jednak zmienia się w czasie na skutek dynamicznego oddziaływania strumienia przepływającego powietrza ze strunami głosowymi. Oddziaływanie to zależy od parametrów mechanicznych strun, głównie od ich masy $m$ i sprężystości $s$

2-34. Szczegółowy (bliższy realnej sytuacji) (a) oraz uproszczony, jednomasowy (b) model strun głosowych. Modele tego typu budowane są dla dokładniejszego przebadania związku między parametrami strun głosowych (masą $m$, sprężystością $s$, tłumieniem wiskotycznym $r$) a parametrami generowanego sygnału dźwiękowego $V_g(t)$ przy różnych wartościach ciśnienia podgłośniowego




Uwzględniając dodatkowo fakt, że działanie mięśni krtani na ogół wpływa równocześnie na parametry $s$ i $m$ możemy przyjąć, że parametr $T_0$ zmienia się w czasie. Są to jednak zmiany na tyle wolne, że rozważając krótkie odcinki czasu możemy przebieg uważać za okresowy.

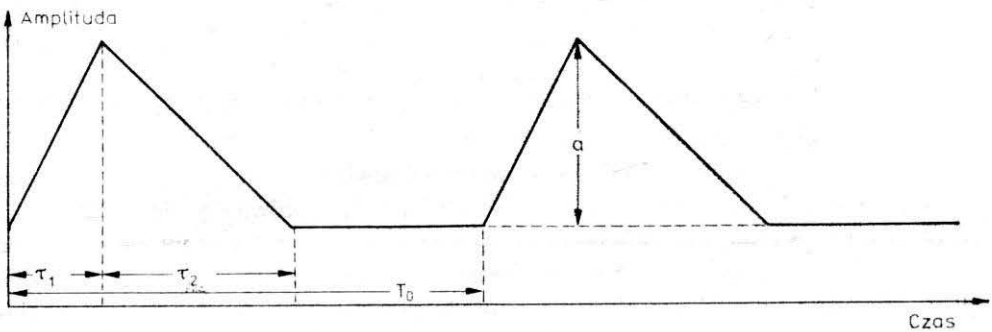
Przebieg czasowy $A_g(t)$ można zapisać w sposób następujący (por. rys. 2-35):

$$A_g(t) = \begin{cases} \dfrac{a}{\tau_1} t & \text{gdy } 0 \leqslant t \leqslant \tau_1 \\ a - \dfrac{a}{\tau_2} t & \text{gdy } \tau_1 \leqslant t \leqslant \tau_1 + \tau_2 \\ 0 & \text{gdy } \tau_1 + \tau_2 \leqslant t \leqslant T_0 \\ A_g(t - nT_0) & \text{gdy } t > T_0;\ n = 1, 2, \ldots \end{cases} \tag{2.34}$$

Transformata Laplace'a takiego przebiegu ma postać

$$A_g(s) = \frac{a}{s^2}\left[\frac{1}{\tau_1} - \left(\frac{1}{\tau_1} + \frac{1}{\tau_2}\right) e^{-s\tau_1} + \frac{1}{\tau_2} e^{-s(\tau_1+\tau_2)}\right] \tag{2.35}$$

Rozważając widmo $A_g(j\omega)$ można zauważyć, że jest ono proporcjonalne do czynnika $f^{-2}$ (gdzie $f$ jest częstotliwością), w wyniku czego widmo sygnału krtaniowego ma obwiednię opadającą (12 dB/oktawę).




2-35. Przebieg czasowy (symulowany przez komputer) powierzchni przekroju szczeliny głośni w czasie procesu artykulacji głoski dźwięcznej. Widoczny jest charakterystyczny, trójkątny kształt impulsów

Zmiany powierzchni $A_g(t)$ (rys. 2-33) wpływają na zmiany impedancji akustycznej $Z_g$ szpary głośni, które modulując przepływ powietrza wywołany ciśnieniem podgłośniowym $p_s$, powodują określony przebieg prędkości objętościowej fali akustycznej $V_g(t)$.

Przebieg omówionych zmian można opisać następująco. Impedancja głośni, wyrażająca się wzorem

$$Z_g = R_k + R_v + j\omega L \tag{2.36}$$

i może być, w zakresie częstotliwości typowych dla tonu krtaniowego ($f < 1000$ Hz) i przy normalnym wysiłku głosowym, wyrażającym się ciśnieniem podgłośniowym $p_s < 1569$ Pa, zadowalająco aproksymowana częścią rzeczywistą*).

$$Z_g \approx R_k + R_v \tag{2.37}$$

gdzie $R_k$ oznacza kinetyczną rezystancję strat, związaną z przemianą ciśnienia na energię kinetyczną przepływu powietrza w głośni, a $R_v$ jest klasyczną rezystancją tarcia powietrza o ściany głośni. Obie składowe, zarówno $R_k$,

*) Wpływ części urojonej impedancji, powodowanej bezwładnością powietrza w szparze głośni, występuje jedynie przy dużych częstotliwościach.

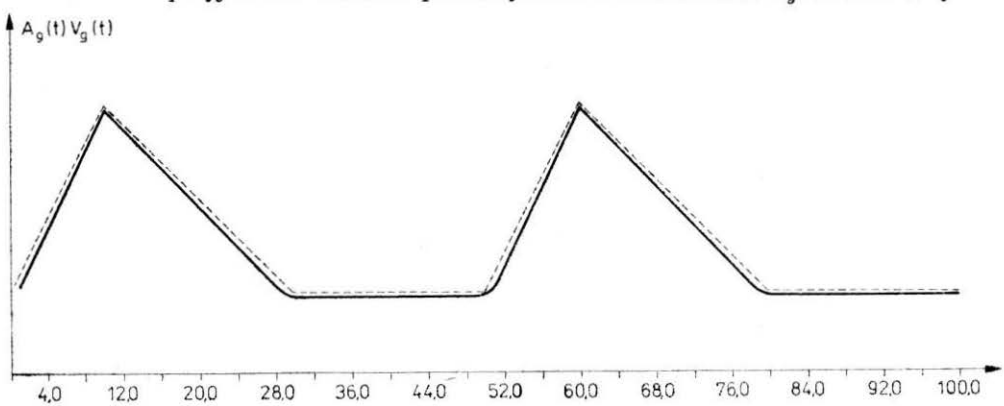
jak i $R_v$ zależą od powierzchni otworu głośni $A_g$, są więc zmienne w czasie. W szczególności dla małych wartości $A_g$ dominujące znaczenie ma rezystancja $R_v$, której wartość może być wyznaczona ze wzoru

$$R_v = \frac{12\mu d_g l_g^2}{A_g^3} \tag{2.38}$$

gdzie $d_g$ jest głębokością szpary głośni (według przytoczonych wyżej danych $d_g = 3$ mm), $\mu$ jest współczynnikiem tarcia powietrza (wprowadzonym już uprzednio), $l_g$ jest długością szpary głośni (zwykle przyjmuje się $l_g = 18$ mm). Dla dużych wartości $A_g$ (konkretnie dla $A_g > 0{,}2\ A_{g\,\max}$) dominujący okazuje się natomiast drugi składnik wzoru (2.37), to znaczy kinetyczna rezystancja strat.

$$R_k = 0{,}875 \frac{\sqrt{2\varrho p_s}}{2A_g} \tag{2.39}$$

Wyliczenia numeryczne oparte na przytoczonych wzorach prowadzą do wniosku, że łączna impedancja $Z_g$ jest rzędu $10^7$ omów akustycznych i może być uznana w prawie całym zakresie częstotliwości za znacznie większą od impedancji wejściowej kanału głosowego. Jest to uzasadnienie dla wcześniej przyjętego założenia, że źródło krtaniowe jest aproksymowane w schemacie zastępczym przez źródło prądowe, co odpowiada wymuszeniu akustycznemu o stałej wartości prędkości objętościowej. Porównując impedancję $Z_g$ z impedancją falową kanału głosowego musimy brać pod uwagę rezonanse w nim występujące. Szczególnie na uwagę zasługują rezonanse przy niskich częstotliwościach, gdyż ze wzrostem częstotliwości impedancja kanału głosowego maleje do wartości około $8{,}5 \cdot 10^5$ omów akustycznych. Natomiast rezonanse niskoczęstotliwościowe, szczególnie odpowiadające tzw. pierwszemu formantowi samogłoskowemu mogą charakteryzować się znacznym zwiększeniem impedancji falowej traktu głosowego, która może wówczas przyjmować wartości porównywalne z wartościami $Z_g$. Zatem w tych za-




2-36. Porównanie przebiegu czasowego powierzchni przekroju szpary głośni $A_g(t)$ (linia przerywana), zadanego zgodnie z rzeczywistym przebiegiem czasowym, z przebiegiem czasowym prędkości objętościowej $V_g(t)$ (linia ciągła). Nawet przy dużej dokładności komputerowych obliczeń różnica pomiędzy przebiegami $A_g(t)$ oraz $V_g(t)$ jest trudno dostrzegalna. Możliwe jest więc traktowanie przebiegu $V_g(t)$ jako fali trójkątnej o widmie opadającym 12 dB/oktawę, gdyż rozbieżność rzeczywistego przebiegu w stosunku do takiego przybliżenia jest pomijalnie mała

kresach niskich częstotliwości charakterystyki modelu wyliczone przy założeniu prądowego charakteru źródła wymuszającego mogą odbiegać od rzeczywistych charakterystyk narządów mowy.

Pomijając te niedokładności możemy obecnie wyznaczyć parametry źródła sygnału, jakim jest krtań. Przebieg czasowy prądu źródła $I_g(t)$ odpowiada (w przyjętym systemie analogii) przebiegowi prędkości objętościowej $V_g(t)$.

$$V_g(t) = \frac{p_s}{Z_g(t)} \cong \frac{p_s}{R_v[A_g(t)] + R_k[A_g(t)]} \tag{2.40}$$

Przebieg czasowy $V_g(t)$ wyliczony ze wzoru (2.40) z uwzględnieniem zależności (2.38) i (2.39) przy trójkątnym przebiegu zależności $A_g(t)$ (por. wzór (2.34)) przedstawiono na rys. 2-36. Jak widać, impulsy $V_g(t)$ nie są — ściśle biorąc — trójkątne, mogą jednak być z zadowalającą dokładnością aproksymowane przebiegiem trójkątnym. Zresztą należy mieć na uwadze także fakt, że przyjęcie trójkątnego kształtu impulsów (rys. 2-35 i wzór (2.34)) miało także charakter pewnego przybliżenia.

Powracając do schematu blokowego z rys. 2-33 należy stwierdzić, że ciśnienie podgłośniowe $p_s$ wraz ze zmienną w czasie impedancją $Z_g$ $(t)$ określają wartość prędkości objętościowej $V_g(t)$, a ta z kolei oddziałuje na zmiany $A_g(t)$ kształtujące wartości $Z_g(t)$. Generator krtaniowy jest więc typowym generatorem pracującym ze sprzężeniem zwrotnym, przy czym parametry generowanego sygnału są określone przez takie własności układu jak parametry mechaniczne drgających strun głosowych. Parametry te są regulowane, jak to wcześniej omówiono, odpowiednimi mięśniami sterowanymi przez wymienione fragmenty systemu nerwowego. Działanie krtani może więc być w pełni opisane przez podane wyżej wzory. Na zakończenie prezentowanych rozważań warto zwrócić uwagę na jeszcze jeden fakt. Impedancja krtani jest zmienna, co bardzo komplikuje model źródła krtaniowego i związane z nim rozważania. Przy obliczeniach przyjmuje się często uproszczony model ze stałą impedancją źródła, wyliczoną dla spoczynkowej wartości stopnia otwarcia fałdów głosowych, co odpowiada wartości impedancji wyliczonej ze wzoru (2.39) przy wstawieniu doń wartości $A_{go} = 5$ mm$^2$. Badania symulacyjne wykazały, że nie ma istotnego znaczenia, czy będzie przyjęta stała wartość impedancji $Z_g$, czy też będzie zastosowany pełny model ze zmienną w czasie impedancją źródła, o ile tylko modelowaniu podlega proces artykulacji przy niezbyt wielkim wysiłku głosowym (wyrażającym się małą wartością ciśnienia podgłośniowego $p_s$).

W celu podsumowania rozważań nad modelem naturalnego procesu artykulacji należy rozważyć jeszcze kilka prostych problemów szczegółowych. Pierwszą sprawą jest wyznaczenie impedancji promieniowania odpowiednio ust $Z_{pu}$ i nosa $Z_{pn}$. Jak pamiętamy impedancje te domykały łańcuchy czwórników modelujących odpowiednio kanał nosowy i kanał ustny. W literaturze są omawiane różne modele akustyczne procesu emisji sygnału mowy z ust. Na podstawie tych modeli dochodzi się do różnych na ogół wzorów do obliczenia wartości $Z_{pu}$ oraz $Z_{pn}$. Nie wnikając tu w zasadność różnych

modeli można przyjąć i zaakceptować ostateczny rezultat rozważań w postaci wzoru

$$Z_{pu} = \frac{\varrho\omega^2}{4\pi c} K(\omega) + j\omega \frac{8\varrho}{3\pi^2 r_u} \tag{2.41}$$

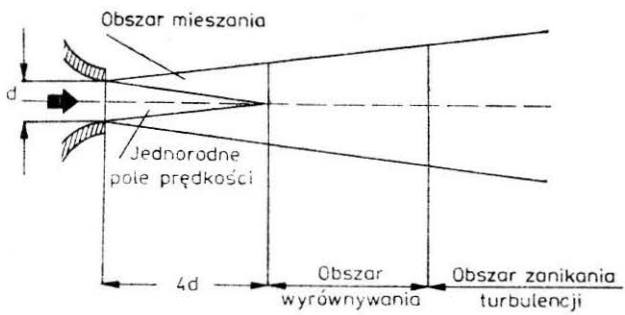
gdzie $r_u$ oznacza promień otworu ust (przy założeniu, że otwór ten można uważać za kołowy, $K(\omega)$ — czynnik korekcyjny mający na celu uwzględnienie, przy niskich częstotliwościach, faktu, że usta znajdują się w kulistej głowie mającej niewielkie rozmiary, a nie są drgającym tłokiem w nieskończonej płaskiej odgrodzie. Czynnik korekcyjny wyliczany jest ze wzoru

$$K(\omega) = \begin{cases} \dfrac{0{,}6\omega}{2\pi 1600} + 1 & \text{dla } \omega \leqslant 2\pi\ 1600 \text{ Hz} \\ 1{,}6 & \text{dla } \omega > 2\pi\ 1600 \end{cases} \tag{2.42}$$

Wzór dla impedancji promieniowania nozdrzy jest prostszy, gdyż z jednej strony małe rozmiary nozdrzy pozwalają na stosowanie przybliżonych wzorów, a ponadto powierzchnia nozdrzy $A_n$ nie ulega zmianom tak jak promień ust $r_u$. Impedancję $Z_{pn}$ można więc obliczyć ze wzoru

$$Z_{pn} = \frac{\varrho\omega^2}{2\pi c} + j\omega \frac{8\varrho}{3\pi\sqrt{A_n}} \tag{2.43}$$

Wszystkie elementy modelu procesu artykulacji głosek dźwięcznych są już określone. Pozostaje jeszcze rozważenie artykulacji głosek szumowych,




2-37. Model źródła szumowego, używany do prezentacji zjawisk zachodzących przy artykulacji głosek szumowych. Model ma postać dyszy o średnicy $d$

przy wytwarzaniu których źródło dźwięku znajduje się w określonym punkcie wzdłuż osi traktu głosowego — powyżej krtani, która ze swojej strony dodaje lub nie składową dźwięczną. Nie można w tym przypadku wyznaczyć przebiegu czasowego sygnału źródła dźwięku, gdyż ma on przypadkowy (stochastyczny) charakter. Rozpatrując wypływ powietrza przez zwężenie w narządach mowy, będące źródłem szumu, możemy posłużyć się modelem dyszy o średnicy $d$, przez którą powietrze o gęstości $\varrho_s$ i temperaturze $T_s$ wypływa z prędkością $v$ do otoczenia, w którym panuje temperatura $T_0$ i gęstość powietrza wynosi $\varrho_0$. Jak widać na rys. 2-37, przy takim modelu procesu generacji szumu można wyróżnić obszar mieszania (zachodzi tu mieszanie powietrza wypływającego z nieruchomym powietrzem otaczającym), w którym generowane jest 49,9% energii akustycznej. Częstotliwość

środkowa emitowanego szumu jest zależna od odległości rozważanego punktu od miejsca przewężenia $x$ i może być wyznaczona ze wzoru

$$f_{śr} = \frac{v}{d}\left(\frac{0{,}2}{x/d}\right)^{0{,}38} \tag{2.44}$$

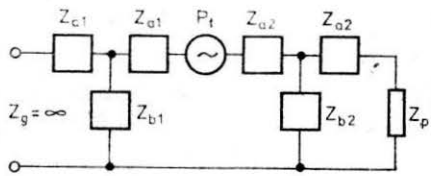
Drugi ważny obszar, zaczynający się w odległości $x > 4d$ od miejsca przewężenia, odpowiada obszarowi wyrównywania. W obszarze tym jest promieniowane 48,4% energii całkowitej szumu, a jej widmo jest określane przez częstotliwość środkową $f_{śr}$

$$f_{śr} = \frac{v}{d}\left(\frac{1{,}8}{x/d}\right)^{1{,}43} \tag{2.45}$$

Pozostałą częścią energii fali, wynoszącą zaledwie 1,7%, nie będziemy się tu zajmowali.

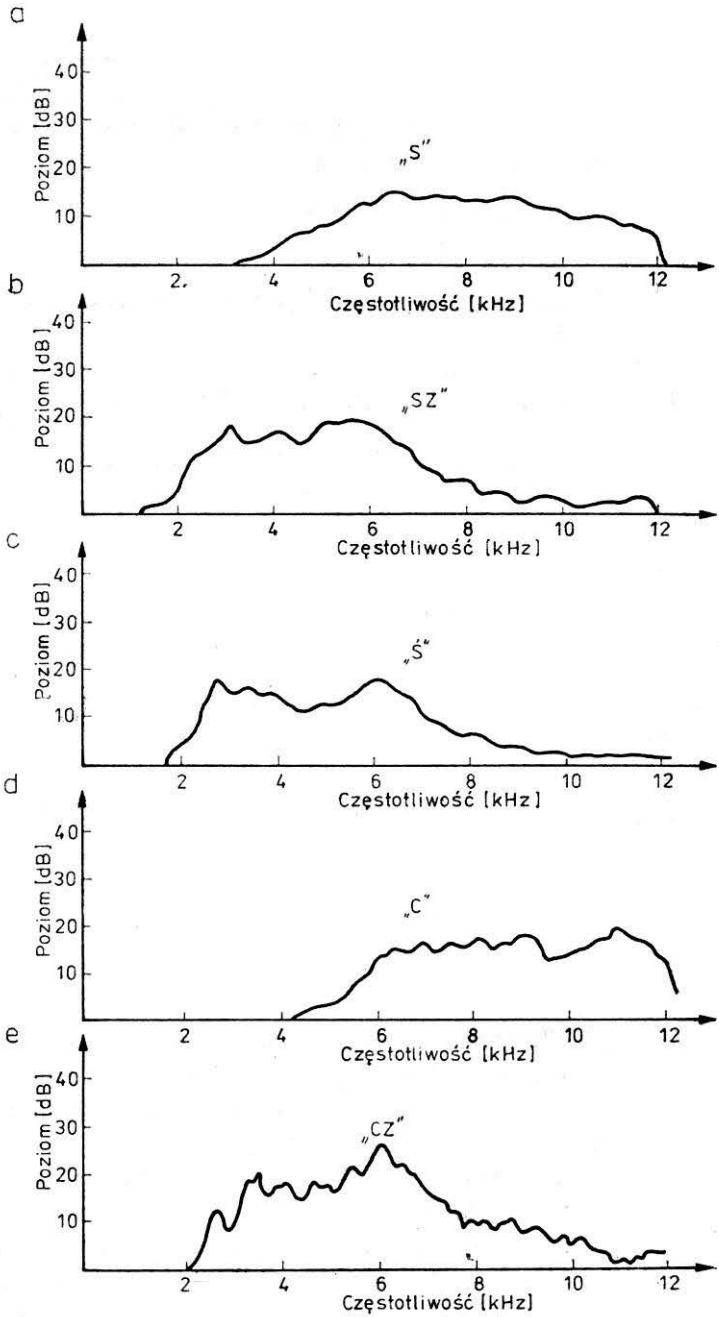
Łączna moc akustyczna szumu generowanego przez strumień wypływającego powietrza wyznaczona może być ze wzoru

$$P = 3 \cdot 10^{-5} \frac{\varrho_s^2 v^8 d^8}{\varrho_0 c^5 \left(\frac{T_0}{T_s} \cdot 0{,}6 + 0{,}4\right)^2} \tag{2.46}$$

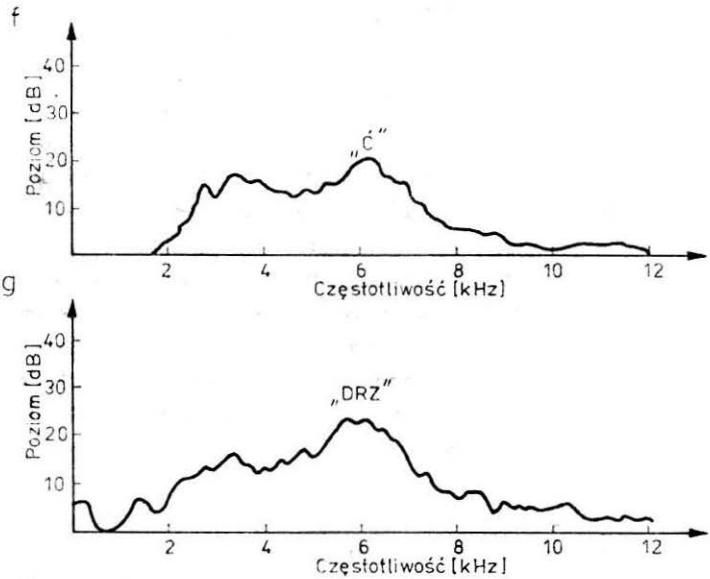



2-38. Maksymalnie uproszczony model procesu artykulacji głosek szumowych. Charakterystyczne jest umieszczenie źródła pobudzenia $p_t$ (w rozważanym przypadku — szumu) wewnątrz łańcucha czwórników modelujących trakt głosowy, a nie na jego początku, jak przy artykulacji głosek z pobudzeniem krtaniowym (dźwięcznych)

Źródło szumu o podanych charakterystykach jest umieszczone ponadkrtaniowo, w określonym punkcie łańcucha czwórników zastępczych modelujących tor gardłowo-ustny (rys. 2-38). Rozwiązując równania tego obwodu możemy w każdym konkretnym przypadku określić transmitancję toru i jego łączną charakterystykę, a tym samym modulujący wpływ, jaki na szum (i ewentualnie towarzyszący mu ton) ma określona konfiguracja narządów mowy. Zasadniczo nowym elementem, jakiego możemy przy tym oczekiwać w stosunku do wcześniej przeprowadzonych rozważań, jest pojawianie się w tym przypadku znaczących zer transmitancji, czyli głębokich minimów charakterystyki amplitudowo-częstotliwościowej, w której uprzednio dominowały bieguny (będące źródłem maksimów rezonansowych). Łatwo to zauważyć na skrajnie uproszczonym modelu traktu głosowego ze




2-39. Widma głosek polskich:
a — szumowej głoski *s*, b — *sz* (zapisywanej w transkrypcji jako *ʃ*), c — *ś* (zapisywanej w transkrypcji jako *ɕ*), *d* — *c* (zapisywanej w transkrypcji jako t͡s), e — *cz* (zapisywanej w transkrypcji jako t͡ʃ), f — *ć* (zapisywanej w transkrypcji jako t͡ɕ), g — *drz* (zapisywanej w transkrypcji jako d͡ʒ) będącej szumową i dźwięczną równocześnie; ma to natychmiastową i oczywistą konsekwencję w postaci pojawienia się w widmie składowych o niskich częstotliwościach, nieobecnych w widmach poprzednio pokazanych głosek




wzbudzeniem ponadkrtaniowym, przedstawionym na rys. 2-38. Transmitancję tego układu można przedstawić w postaci

$$H(j\omega) = \frac{Z_{b2} Z_p}{(Z_{b1} + Z_{a1} + Z_{a2} + Z_{b2})(Z_{b2} + Z_{a2} + Z_p)} \tag{2.47}$$

Transmitancja ta ma minima w punktach będących pierwiastkami licznika. Oczywiście przy bardziej realistycznym odwzorowaniu kształtu traktu głosowego wraz z przewężeniem postać transmitancji będzie bogatsza, a jej bieguny i zera będą determinować obwiednię generowanego szumu, decydując o zróżnicowanym kształcie widma różnych głosek szumowych (por. rys. 2-39).

Podsumowując można stwierdzić, że istnieje możliwość opisania i matematycznego modelowania procesu naturalnej artykulacji sygnału mowy, co dowodzi, że proces ten jest już wystarczająco dobrze poznany i że wiedza na ten temat jest spójna i wewnętrznie niesprzeczna. Rozumiejąc dokładnie proces naturalnej artykulacji mowy możemy też podejmować próby jego naśladowania konstruując urządzenia i algorytmy komputerowe wytwarzające mowę w sposób sztuczny. Teoretycznie najbardziej oczywistą drogą takiego sztucznego generowania sygnału mowy jest skonstruowanie modelu — na przykład symulacyjnego — opisanych wyżej procesów i uzyskiwanie z niego potrzebnych przebiegów czasowych, emitowanych następnie z wykorzystaniem technik przetwarzania cyfrowo-analogowego i typowego wyposażenia elektroakustycznego. Opisana droga jest jednak dla współczesnej techniki zbyt złożona. Nakład obliczeń wymagany przy symulacji rzeczywistego procesu artykulacji nie pozwala na uzyskiwanie rezultatów w czasie rzeczywistym, a ponadto koszt takiej syntezy jest zbyt duży, aby

mógł być akceptowany. Potrzebne są więc specjalistyczne metody sztucznej syntezy sygnału mowy, dostarczające sygnału o zadowalających parametrach — tańszym kosztem.

## 2.5. Wytwarzanie mowy z wykorzystaniem systemów technicznych

Jak pokazano na rysunku 2-1, wytwarzanie mowy w systemach technicznych może polegać bądź to na odtwarzaniu sygnału zapisanego w określonej postaci, bądź na generacji sygnału z wykorzystaniem specjalistycznej aparatury. W pierwszym przypadku mamy do czynienia z mową rekonstruowaną i podstawowy problem polega na tym, jak zmniejszyć objętość informacyjną zarejestrowanej mowy, aby nie zajmować zbyt dużych obszarów pamięci w urządzeniu odtwarzającym. Technika ta jest prymitywna, ale gwarantuje szybkie osiągnięcie potrzebnych efektów. Na tej zasadzie działa większość komercyjnych systemów syntezy mowy (niektóre z nich będą omówione).

Alternatywne podejście polega na tym, by stworzyć syntezator o możliwie prostej strukturze, a równocześnie o parametrach i możliwościach zbliżonych do naturalnego traktu głosowego człowieka. Ta droga jest trudniejsza, ale pozwala na syntezę sygnału odpowiadającego dowolnym — a nie tylko uprzednio zarejestrowanym — wypowiedziom. Problem polega tu głównie na opracowaniu metod sterowania parametrami syntezatora mało obciążających dla systemu sterującego (w sensie nakładu obliczeń). Wydaje się, że ostatnio na tej drodze notuje się sporo interesujących rozwiązań praktycznych i zapewne technika generacyjna zdominuje wkrótce rynek systemów syntezy mowy.

Przechodząc do rozważań bardziej szczegółowych zaczniemy od urządzeń odtwarzających. Sygnał mowy można zapamiętać w formie analogowej (na przykład w postaci nagrania magnetofonowego) i następnie odtworzyć w razie potrzeby w całości lub składając z kilku odpowiednio dobranych fragmentów całą wiadomość. W ten sposób funkcjonuje mnóstwo urządzeń informacyjnych od telefonicznych automatów informacyjnych poczynając (np. „zegarynka"), a na informacji dworcowej i domowych „sekretarzach" kończąc. Wadą takiego sposobu odtwarzania mowy jest mała elastyczność: asortyment możliwych wypowiedzi jest ograniczony i ściśle zdeterminowany zawartością „banku informacji", a zmiana odtwarzanych komunikatów jest kłopotliwa. Szczególnie istotne jest ograniczenie dotyczące utrudnień montażu dłuższych komunikatów z elementów składowych. Możliwości popularnej zegarynki, w której oddzielnie nagrane poszczególne godziny są montowane z cyklicznie odtwarzaną liczbą minut, wyznaczają zakres możliwych operacji. W systemach, w których liczba możliwych komunikatów musi być większa, a ich różnorodność także przekracza ramy najprostszych kilkuwyrazowych anonsów, pojawiają się kłopoty z efektywnym magazynowaniem, wyszukiwaniem i łączeniem ze sobą elementów. Pierwszy problem, który się przy tym wyłania, dotyczy rodzaju użytych do składania

elementów. Niewątpliwie najłatwiej jest zbudować system, w którym elementami podlegającymi „montażowi" są całe wyrazy. Są one w naturalny sposób odizolowane jeden od drugiego, mogą być więc montowane przy minimalnym jedynie uwzględnieniu zjawisk powstających na styku segmentów. Wprawdzie mowa tak rekonstruowana będzie monotonna, pozbawiona wszelkiej intonacji i wysoce nienaturalna w odbiorze, ale w końcu można się zgodzić z pewnymi niedogodnościami, jeśli w ślad za nimi iść będą prostota konstrukcji i niska cena. Praktyka wykazała bowiem, że komunikaty wytwarzane w wyżej omówiony sposób będą zrozumiałe. Jednak liczba wyrazów, którą trzeba przy takim systemie zgromadzić jako „budulec" do syntezy, jest bardzo duża. Nawet przy drastycznych ograniczeniach podstawowy zasób słów niezbędnych do w miarę elastycznej budowy tworzonych wypowiedzi musi zawierać kilka tysięcy wyrazów. Można przyjąć, że słownik zawierający 4000 wyrazów byłby w podstawowych zastosowaniach wystarczający. Jednak język polski odznacza się bardzo niewygodną własnością: jest fleksyjny, co oznacza, że obok podstawowych form poszczególnych wyrazów trzeba dysponować także formami odmiennymi, a to powiększa słownik o dalsze 12 000 wyrazów. Pełny słownik języka polskiego zawiera natomiast ponad 100 000 wyrazów. W sumie jest tego zbyt wiele, aby można było taką „taśmoteką" swobodnie operować.

Może więc rozwiązaniem jest użycie mniejszych fragmentów lingwistycznych — na przykład sylab? Jest ich w języku polskim około 2000, a przez odpowiednie ich zestawianie można wygenerować każdy wyraz. Jednak jest to nadal liczba zbyt duża dla operatywnego działania systemu, a ponadto mowa zestawiona z oddzielnych sylab bez zastosowania „łagodnego" przejścia sygnału od jednej sylaby do drugiej jest bardzo nieprzyjemna w odbiorze i mało zrozumiała. Wydaje się, że zamiast kompromisowo wybierać sylaby lepiej pójść w „rozdrabnianiu" sygnału mowy jeszcze dalej i zdecydować się od razu na to, aby opierać system odtwarzania na głoskach, czyli mówionych odpowiednikach liter. To nieprecyzyjne określenie wymaga oczywiście uściślenia, gdyż często kilka głosek jest kodowanych tą samą literą, a niekiedy kilka liter koduje jedną głoskę. W dalszej części książki pojęcie głoski będzie bardziej istotne dla rozważań — szczególnie w kontekście zadań rozpoznawania mowy — i wówczas będzie dokładniej przedyskutowane*).

*) W literaturze dotyczącej zagadnień analizy, syntezy i rozpoznawania mowy często jest używany termin f o n e m. Pod pewnymi względami pomiędzy głoską a fonemem zachodzą daleko idące analogie i dlatego w literaturze technicznej, w której efektywność praktycznych rozwiązań ceni się zwykle wyżej, niż precyzję wysławiania — traktuje się niekiedy terminy „fonem" i „głoska" wymiennie. Takie postępowanie jest naturalnie nieprawidłowe, gdyż pojęcie fonemu jest bardziej abstrakcyjne od pojęcia głoski — często przyjmuje się na przykład, że desygnatem pojęcia fonem jest klasa głosek, między którymi występują jedynie różnice osobnicze (tj. wynikające z indywidualnych cech głosu lub wymowy) lub kontekstowe (tj. wynikające z wpływu głosek sąsiednich). Dla technika fonem może zatem być idealnym wzorcem głoski, od którego każda konkretna realizacja w pewnym stopniu odbiega. Jednak w systemach syntezy mowy komputer tworząc głoski posługuje się pewnymi wzorcami, które ściśle biorąc fonemami nie są — jednak bywają tak nazywane. Podobnie w systemach rozpoznawania rejestrowane głoski są porównywane ze wzorcami, które także nie spełniają rygorystycznych wymogów definicji fonemu — a jednak z braku innego terminu mówi się także w tym przypadku o rozpoznawaniu fonemów. Te niedoskonałości języka techniki znajdują swoje częściowe usprawiedliwienie w fakcie, że lingwiści, którzy wprowadzili pojęcie fonemu i którzy nim chętnie operują (zarzucając

W chwili obecnej istotne jest, że głosek jest niewiele — około 40 dla języka polskiego. Z głosek można zbudować każdy wyraz lub grupę wyrazów, jest to więc (pozornie) idealne „tworzywo" do reprodukcji mowy. Niestety, głoski w różnych kontekstach miewają różne brzmienie i różnice te są istotne dla zrozumienia treści wypowiedzi. Ponadto głosek nie można już pod żadnym pozorem łączyć mechanicznie ze sobą, gdyż łagodne przejście od głoski do głoski — w warunkach naturalnej artykulacji zabezpieczone przez łagodny ruch narządów mowy od jednej pozycji do drugiej, odpowiadającej artykulacji kolejnej głoski — jest koniecznym warunkiem odbierania subiektywnego całości sygnału jako zrozumiałego sygnału mowy. Zresztą doświadczenia psychologów dowodzą, że stany przejściowe, to znaczy te partie sygnału, które odpowiadają przejściu między jedną głoską a następną lub poprzednią, bywają bardziej istotne dla zrozumienia analizowanej głoski, niż jej część stacjonarna. Dotyczy to głównie niektórych spółgłosek, które stają się niezrozumiałe dla człowieka, jeśli wysłuchuje się ich w izolacji, gdyż zasadnicze informacje potrzebne do ich identyfikacji mieszczą się w charakterystycznych deformacjach sygnału samogłosek poprzedzających je lub następujących po nich. Co więcej, można przeprowadzić doświadczenie polegające na słuchaniu fragmentu sygnału akustycznego, z którego usunięto fragment odpowiadający badanej spółgłosce, pozostawiając stany przejściowe samogłosek poprzedzających i następujących po usuniętej głosce. Efekt jest zadziwiający: słuchacz „słyszy" i rozpoznaje nieistniejącą głoskę prawidłowo! Należy podkreślić, że nie ma możliwości domyślania się — na podstawie kontekstu — o jaką głoskę chodzi, gdyż badania tego typu prowadzi się z wykorzystaniem tzw. logatomów, to znaczy zestawień głosek pozbawionych sensu.

Wniosek z przytoczonych rozważań jest tylko jeden. Analogowe metody wytwarzania sygnału mowy nie mają perspektyw. Będą istniały jeszcze przez jakiś czas w prostych systemach powiadamiających (zegarynka itp.), ale zapewne i tam wyprą je w końcu doskonalsze pod każdym względem systemy cyfrowe. Operując techniką cyfrową można dokonywać takich manipulacji na zapamiętanych fragmentach sygnału mowy, których nigdy i przy użyciu żadnej aparatury analogowej nie uda się nawet w przybliżeniu naśladować. O szczegółach odwzorowania sygnału mowy w systemie cyfrowym, a także o operacjach, które można wykonywać na sygnale mowy dysponując systemem cyfrowym, będzie mowa w rozdz. 4, a także — w kontekście konkretnych zastosowań — w rozdz. 5 i 6. Teraz problematyka cyfrowej reprezentacji sygnału — odtwarzanego lub syntetyzowanego —

przy tym bezustannie technikom niepoprawne jego używanie), sami od około stu lat toczą spory na temat definicji tego pojęcia, nie mogąc zgodzić się na żadną z kilkudziesięciu będących w użyciu, opublikowanych i wielostronnie uzasadnionych propozycji. Pewien przegląd tego zagadnienia i związanej z nim literatury dokonany jest w referacie plenarnym XXII Otwartego Seminarium z Akustyki (W. Jassem: Wstępne założenia akustycznej teorii fonemu. Materiały OSA'85, Kraków 1985, str. 61—64). W książce przyjęto ze względów praktycznych nazwę „głoska", niekiedy jednak będzie także mowa o fonemach traktowanych jako klasy głosek lub ich wzorce. Zwolennicy bardziej precyzyjnych definicji muszą sami wybrać jedno z konkurujących określeń proponowanych przez lingwistów i dzielnie odpierać ataki zwolenników innych wyjaśnień tego terminu.

zostanie tylko zasygnalizowana. Czasowy przebieg sygnału może być zapisany w postaci ciągu wartości liczbowych, odpowiadających amplitudom sygnału mierzonym w ustalonych, zwykle jednakowo odległych od siebie momentach czasu. Mając przebieg czasowy możemy zawsze dokonać zamiany na wspomniany zbiór dyskretnych wartości liczbowych. Co więcej, jeśli tylko odstępy czasu między kolejnymi próbkami są dostatecznie małe, to sygnał cyfrowy mieści w sobie dokładnie tę samą informację, co sygnał oryginalny, gdyż możliwe jest całkowicie dokładne odtwarzanie sygnału analogowego z zarejestrowanego sygnału cyfrowego. Wszystkie wiążące się z tym uwarunkowania i wiadomości teoretyczne podano w p. 4.1.

Cyfrowe metody odtwarzania mowy stawiają przed konstruktorami odpowiedniej aparatury problem sposobu reprezentacji sygnału mowy w systemie cyfrowym. Wspomniana wyżej metoda bezpośredniego zapisu przebiegu czasowego sygnału akustycznego w postaci cyfrowej jest najprostsza, ale niesłychanie pamięciochłonna. Można wykazać, że sygnał mowy odtwarzany z całą dokładnością reprezentuje strumień informacji 240 000 bit/s (bodów). Taki strumień informacji sprawia trudności przy przesyłaniu go w formie cyfrowej na większą odległość, a ponadto w błyskawicznym tempie wypełnia pamięć urządzenia przetwarzającego. Pamięć operacyjna największych dostępnych w Polsce komputerów wystarczy przy tak rozrzutnym kodowaniu na zapamiętanie niecałej minuty transmisji sygnału, a pamięć przeciętnego mini- czy mikrokomputera (na przykład popularnych obecnie w kraju i za granicą komputerów osobistych) może pomieścić zaledwie około 1 sekundę sygnału — i to pod warunkiem, że nie będzie w niej żadnych programów, które też zajmują miejsce. A jak tu operować sygnałem, łączyć go i przekształcać bez odpowiednich programów?

Problem oszczędnego kodowania sygnału mowy jest więc centralnym zagadnieniem warunkującym efektywność cyfrowego odtwarzania mowy. Różne systemy syntezy mowy rozwiązują to zagadnienie rozmaicie; sam problem jest zresztą nie taki nowy, jak się może wydawać, gdyż przed informatykami, chcącymi nauczyć mowy swoje komputery, borykali się z problemem zmniejszenia informacyjnej objętości sygnału mowy specjaliści z zakresu telekomunikacji, ponieważ przy przesyłaniu mowy na odległość także można odnieść niebagatelne korzyści, jeśli się pasmo sygnału odpowiednio ograniczy. Do zagadnienia tego powrócimy w rozdz. 6.

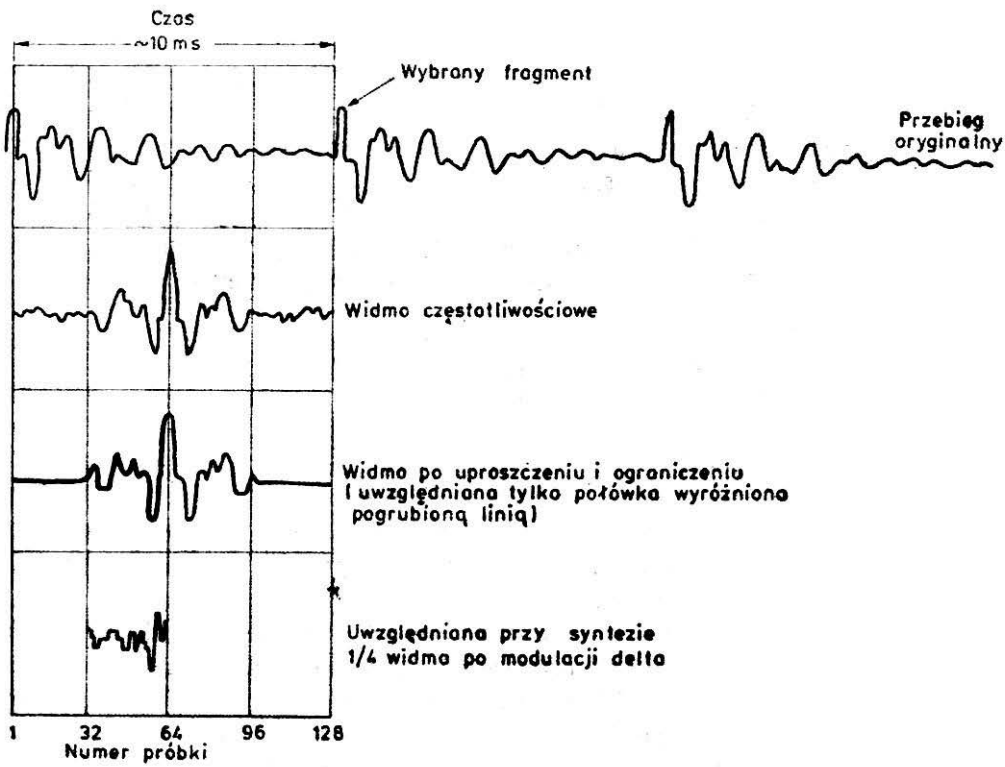
Obecnie omówimy pewien konkretny system oszczędnego kodowania sygnału mowy w systemie cyfrowym, stosowany w komercyjnym systemie odtwarzania mowy firmy National Semiconductor. System ten, nazywany DIGITALKER, operuje całymi wyrazami zapisanymi w pamięci cyfrowej w tak zagęszczonej postaci, że pojedynczy układ scalony z pamięcią o pojemności 128 kbit wystarcza do zapamiętania około 120 słów odpowiednio dobranych do zastosowania „mówiącej końcówki" w systemie komputerowym, w kasie sklepowej, w samochodzie lub w domu. Oczywiście odtwarzanie sygnału mowy wymaga odpowiedniej interpretacji zgromadzonych w pamięci zapisów, służy do tego specjalny scalony mikroprocesor nazywany SPC (ang. *Speech Processor Chip*), odczytujący zapisy w pamięci i na ich

podstawie generujący sygnał mowy. Jakość uzyskanej mowy jest zadowalająca, gdyż odpowiednie oprogramowanie procesora pozwala także na pewną modulację sygnału zarówno w zakresie tonacji, jak i natężenia. Pracę procesora synchronizuje zegar kwarcowy.

Problemem jest oczywiście sposób oszczędnego, upakowanego zapisu sygnału w pamięci. W syntezatorze DIGITALKER redukcja informacyjnej objętości zapamiętanego i przystosowanego do odtworzenia sygnału mowy przebiega czteroetapowo. W pierwszej kolejności z zarejestrowanego i przeznaczonego do odtworzenia sygnału wybiera się kilka najbardziej charakterystycznych fragmentów. Wykorzystuje się przy tym znaną właściwość sygnału mowy, że prawie wszystkie głoski mają charakterystyczne przebiegi czasowe, których cykliczne powtarzanie może być subiektywnie ocenione jako ciągła artykulacja rozważanej głoski. Szczególnie łatwe jest wykrycie i wskazanie takich fragmentów w samogłoskach i w spółgłoskach szumowych. W innych głoskach zachodzi niekiedy konieczność wybrania więcej niż jednego fragmentu do zapamiętania, przy czym najtrudniejsze do odtwarzania spółgłoski plozyjne mają na tyle krótki czas trwania, że cały ich przebieg czasowy łącznie z fazą zwarcia musi być użyty jako wzorzec, bez konieczności powtarzania. Dzięki wybraniu wspomnianych charakterystycznych fragmentów sygnału i stosowaniu ich powtarzania (typowo około 5 do 15 powtórzeń zapamiętanego fragmentu imituje pojedynczą głoskę) można oszczędzić — jak się oszacowuje — ponad 75% pamięci, która byłaby potrzebna przy przechowywaniu nie przetworzonego sygnału. W niektórych systemach to wystarcza. Na przykład, jeden ze znanych systemów syntezy mowy polskiej zakładał jedynie zapamiętywanie owych charakterystycznych fragmentów i ich cykliczne odtwarzanie zgodnie z założonym programem, dostarczając dobrej jakości mowy przy stosunkowo niewielkim zajęciu pamięci komputera. W systemie DIGITALKER zastosowano kolejne przekształcenie, powodujące dalszą, wydatną redukcję informacyjnej objętości sygnału. Wykorzystano mianowicie wymieniony przy omawianiu naturalnego procesu artykulacji fakt, że położenie i ruchy narządów artykulacyjnych kształtują głównie widmo sygnału mowy. Poza tym ucho ludzkie jest mało wrażliwe na wartości przesunięcia fazowego, zatem większość interesujących informacji jest zawartych w jego charakterystyce amplitudowo-częstotliwościowej, łatwej do uzyskania z zarejestrowanego przebiegu czasowego — na przykład na drodze obliczeniowo przeprowadzonej transformacji Fouriera. Wybierając zatem określony fragment sygnału mowy, będący „reprezentantem" pewnej głoski (przy czym jego typowy czas trwania odpowiada około 10 ms naturalnego trwania sygnału mowy i jest reprezentowany w systemie przed dokonaniem dalszej redukcji przez 128 próbek wartości chwilowych w jednakowych odległych momentach czasu) możemy dokonać jego transformacji, otrzymując w wyniku 128 wartości amplitud sygnału dla wybranych 128 pasm częstotliwości. Na razie oszczędności nie widać: było 128 liczb i jest nadal 128 — tyle że w pierwszym podejściu są to liczby zespolone, reprezentujące amplitudowe i fazowe składowe widma. Możliwość redukcji informacji wynika dopiero z przeanalizo-

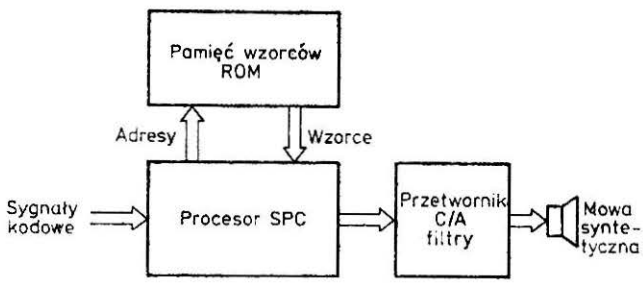
wania struktury tego widma. Po pierwsze jest ono symetryczne, wystarczy więc zapamiętanie jedynie połówki widma, bo druga jest możliwa do odtworzenia na podstawie symetrii. Po drugie łatwo zauważyć, że składniki o dużej amplitudzie grupują się (na ogół!) przy niskich częstotliwościach, wystarczy więc dla odtworzenia sygnału brać pod uwagę jedynie centralną część widma. W praktyce zatem brana jest pod uwagę 1/4 widma, czyli 32 próbki zamiast 128. W ten sposób uzyskuje się kolejne 75% oszczędności pamięci.

W celu uzyskania dalszego ograniczenia informacyjnej objętości zapamiętywanych danych wykorzystuje się modulację typu delta. Polega ona na tym, że zapamiętywane są w kolejnych próbkach przyrosty wartości sygnału (w rozważanym przypadku — widma), a nie same wartości. Ponieważ widmo nie zmienia się zbyt szybko, więc przyrosty wyrażają się mniejszymi wartościami niż same próbki. W ten sposób nie zyskuje się wprawdzie na liczbie próbek, która pozostaje taka sama, ale wartości do zapamiętania mieszczą się w mniejszym przedziale i mogą być reprezentowane mniejszą liczbą bitów. Cały ten złożony proces przedstawiono na rys. 2-40. Łatwo zauważyć, że zapamiętana forma sygnału silnie odbiega od jego rzeczywistego przebiegu i dlatego w procesie odtwarzania musi uczestniczyć specjalny procesor SPC, odtwarzający sygnał na podstawie szczególnego zapisu w pamięci.




2-40. Zasada kompresji sygnału mowy, stosowana przy syntezie mowy w systemie DIGITALKER

Schemat systemu odtwarzania przedstawiono w uproszczonej postaci na rys. 2-41. Centralną rolę odgrywa procesor SPC, do którego jest podawany początkowy adres obszaru w pamięci ROM, w którym jest umieszczony zapis potrzebnego komunikatu. Synteza może przebiegać na podstawie wzorców całych wypowiedzi zapisanych w wyżej omówiony oszczędny sposób w pamięci ROM lub może polegać na montażu wyrazów z zapisanych




2-41. Uproszczona struktura syntezatora mowy działającego według schematu systemu DIGITALKER. Upakowane wzorce w pamięci ROM, których objętość informacyjną zminimalizowano techniką przedstawioną na rysunku 2-40 wymagają dla odtworzenia czasowego przebiegu sygnału mowy użycia specjalnego procesora SPC

elementów (głosek) z wykorzystaniem specjalnego programu generacji i dodatkowych informacji na temat czasu trwania poszczególnych elementów wypowiedzi i ich modulacji amplitudowych (akcent) i częstotliwościowych (intonacja). Problem, jaki przy tym powstaje, polega na zapewnieniu łagodnego przejścia od generacji jednej głoski do generacji następnej. Stosunkowo prosty koncepcyjnie sposób polega na tym, aby w pewnym odcinku czasu od chwili $T_1$ do chwili $T_2$ określać wypadkowy sygnał $U(t)$ jako sumę ważoną przebiegu czasowego głoski kończącej swoje brzmienie $U_1(t)$ i głoski pojawiającej się jako następna $U_2(t)$. W zapisie matematycznym operacja ta jest prosta:

$$U(t) = U_1(t)\left(1 - \frac{t-T_1}{T_2-T_1}\right) + U_2\frac{t-T_1}{T_2-T_1} \tag{2.48}$$

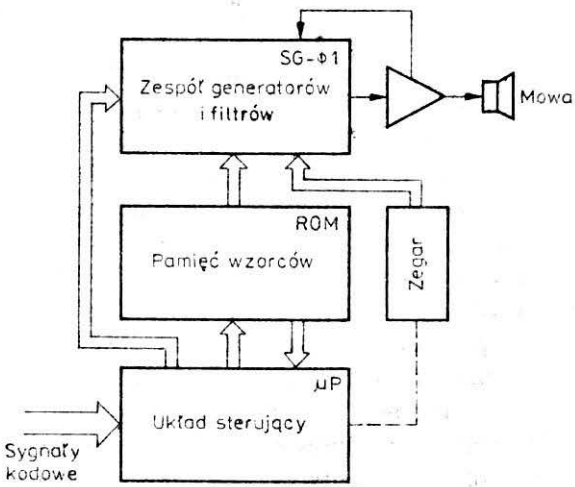
dla $T_1 \leqslant t \leqslant T_2$.

Rzeczywista realizacja tej koncepcji jest uciążliwa ze względu na konieczność wykonywania mnożeń (dwu w każdym kroku czasowym), co w warunkach stosowania mikroprocesorów może prowadzić do trudności z pracą w czasie rzeczywistym.

Podsumowując można więc stwierdzić, że przytoczona metoda dobrze nadaje się do stosowania przy ograniczonym i z góry ustalonym słowniku, który wówczas musi być w całości umieszczony w pamięci wzorców (naturalnie w oszczędnej postaci). Przy próbach zastosowania omówionej metody odtwarzania do generacji dowolnych wypowiedzi albo trzeba godzić się na bardzo niską jakość odtwarzanej mowy, spowodowaną niedoskonałoś-

ciami procesu łączenia*) albo należy liczyć się z potrzebą zastosowania komputera o dużej mocy obliczeniowej.

Znacznie korzystniejsze własności mają syntezatory parametryczne. Proces syntezy polega w nich na programowanym wybieraniu parametrów generatora sygnału, którego budowa w większym lub mniejszym stopniu jest wzorowana na schemacie i zasadach funkcjonowania omówionego wyżej traktu głosowego człowieka. Przykładem systemu tego rodzaju jest syntezator Votrax SC-01. Parametrami sygnału mowy, sterującymi generator, są tzw. formanty, to znaczy rezonanse powstające w narządach mowy i charak-




2-42. Schemat syntezy mowy używany w systemie VOTRAX. Wzorce są w tym przypadku opisem reguł generacji dźwięku, a nie upakowanym zapisem czasowego przebiegu sygnału. Sercem układu jest sterowany cyfrowo system generujący, zawierający programowalne generatory, filtry i elementy formujące obwiednię dźwięku

teryzujące się maksimami obwiedni widma sygnału emitowanego podczas naturalnej artykulacji. Zamiast więc generować czasowy przebieg sygnału mowy — co jak wskazano wyżej wymaga dużych ilości informacji — można generować sygnał o uproszczonym widmie, kształtowanym przez zespół przestrajanych generatorów i filtrów regulowanych pod względem częstotliwości środkowej i szerokości pasma. Korzysta się tu z faktu, że zmiany widma, wywołane ruchem narządów mowy i wynikającymi z tego zmianami geometrii traktu głosowego, przebiegają stosunkowo wolno i do ich śledzenia wystarcza strumień informacji rzędu tysiąca bitów na sekundę — a więc wielokrotnie mniej niż w najbardziej nawet „upakowanych" rozwiązaniach, w których sygnał jest odtwarzany z wzorca zapamiętanego w formie przebiegu czasowego.

Kontrolowanie procesu generacji sygnału w przypadku odrębnego sterowania każdego generatora i każdego filtru może być dość złożone i czasochłonne. Na szczęście nie wszystkie kombinacje parametrów układu syntetyzującego są jednakowo prawdopodobne, przeciwnie — interesują nas wyłącznie te, które odpowiadają konkretnym głoskom rozważanego języka**).

*) Najczęściej w celu uniknięcia kłopotów na stykach głosek separuje się je sztucznie wstawianymi krótkimi pauzami (okresami ciszy), co pogarsza naturalność sygnału, ale ułatwia jego zrozumienie.

**) W przypadku systemu Votrax SC-01 chodzi oczywiście o język angielski, którego głoski różnią się znacznie od fonemów języka polskiego!

Przyjmując (z pewnym nadmiarem wynikającym z konieczności uwzględnienia różnych wariantów tej samej głoski w różnych kontekstach), że będzie rozważany zestaw 64 dźwięków (w systemie Votrax 45 podstawowych form głosek, 16 dodatkowych brzmień specjalnych oraz 3 odcinki ciszy o różnej długości) możemy każdy z tych dźwięków zakodować za pomocą 6-bitowego kodu. Stosując mikroprocesor 8-bitowy pozostają jeszcze 2 bity, które w systemie Votrax służą do kodowania częstotliwości podstawowej każdego z dźwięków i dają możliwość swobodnego kształtowania intonacji wypowiedzi.

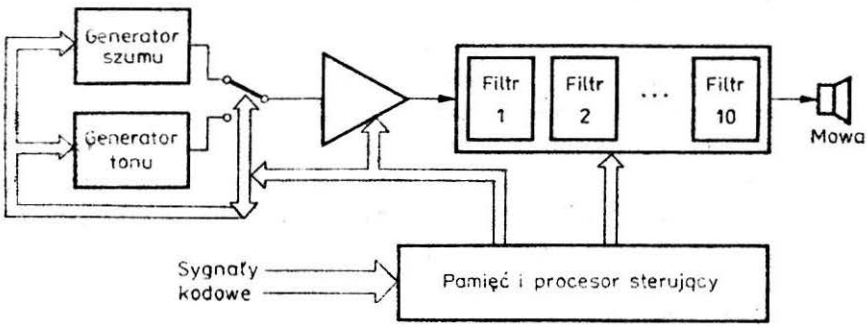
Naturalnie w omawianym systemie występuje również problem zapewnienia łagodnego, płynnego przejścia od jednej głoski do drugiej. Układ syntezy musi zapewnić „miękkie" przejście wartości parametrów, na podstawie których jest dokonywana synteza, od jednej wartości ustalonej do następnej. Układ musi ponadto uwzględniać fakt, że pewne przejścia — na przykład początek artykulacji głosek plozyjnych (takich jak *p* lub *c*) musi charakteryzować się gwałtowną zmianą parametrów. Nie można więc ograniczyć się do „wolno przestrajanych" generatorów i filtrów — układy formujące widmo muszą mieć możliwość szybkiej zmiany parametrów i zapewniać odpowiednią kontrolę programową procesu przejściowego.

Struktura blokowa syntezatora jest dość prosta. W pamięci ROM są zgromadzone wzorce głosek, zawierające parametry powodujące przestrojenie generatorów i filtrów w celu wygenerowania potrzebnego sygnału. Układ sterujący pamięcią wybiera z niej i przesyła do systemu generującego SG-01 kolejne elementy według reguł wynikających z syntetyzowanej wypowiedzi. System generujący wytwarza sygnał będący złożeniem odpowiednich częstotliwości, imitujący naturalną mowę. Sygnał musi być poddany procesowi wzmocnienia we wzmacniaczu akustycznym i moduł SC-01 zapewnia wzmacniaczowi dodatkowy sygnał akustyczny pełniący rolę sprzężenia zwrotnego do większej stabilności generowanego sygnału. Cały układ jest synchronizowany zewnętrznym zegarem o częstotliwości 720 kHz, przy czym chcąc uzyskać efekt modulacji mowy (dla bardziej naturalnej intonacji) trzeba ten zegar programowo przestrajać.

Omówiony system Votrax leży w istocie na pograniczu między systemami odtwarzania mowy a systemami jej syntezy. Z punktu widzenia sposobu generacji sygnału można tu mówić o syntezie, gdy sterowane w sposób parametryczny generatory i filtry tworzą sygnał, a nie tylko go odtwarzają. Z punktu widzenia sterowania tym procesem mamy jednak do czynienia z procesem odtwarzania: pamięć ROM zawiera dla każdej wypowiedzi dokładny schemat sterowania, będący również (podobnie jak w systemie DIGITALKER i pokrewnych) — pewną formą zapisu oryginalnego sygnału, który należy odtworzyć. Naturalnie elastyczność systemu parametrycznego i jego możliwości są znacznie większe niż systemu odtwarzania przebiegów czasowych, niemniej o prawdziwej syntezie trudno tutaj mówić. Syntezatorem z prawdziwego zdarzenia jest natomiast przyrząd firmy Texas Instruments nazywany Voice Synthesis Procesor (w skrócie VSP), opracowany w postaci układu scalonego dużej skali integracji. Na rynku (amery-

kańskim) są dostępne zarówno proste wersje syntezatora TMS 5100, stosowane głównie w zabawkach, jak i złożone kosztowne systemy TMS 5200 wykorzystywane w systemach komputerowych.

Struktura syntezatora TMS 5200 jest właściwie modelem traktu głosowego (rys. 2-43) realizowanym cyfrowo i sterowanym w sposób całkowicie para-




**2-43. Schemat parametrycznej syntezy mowy z wykorzystaniem procesora VSP, charakterystycznej między innymi dla generatorów firmy Texas Instruments. Taki system syntezy stawia najwyższe wymagania sprzętowi użytemu do generacji sygnału mowy, jest jednak najbardziej oszczędny, jeśli idzie o pojemność pamięci wymaganą do zapamiętania określonego odcinka czasowego sygnału mowy, a także dostarcza sygnału mowy o bardzo dobrej jakości**

metryczny. Sterowanie tym systemem odbywa się za pomocą 50-bitowych rozkazów podawanych z częstotliwością 40 Hz, a więc bardzo wolno. Do sterowania syntezatora zaprojektowano specjalną pamięć o dużej pojemności (zestaw 16 układów TMS 6100 może pomieścić do 30 minut nieprzerwanej rozmowy) i małej szybkości działania — układ scalony ROM TMS 6100. Układ syntezatora obejmuje: generatory tonu krtaniowego i szumu zrealizowane cyfrowo i sterowane (przestrajane) za pomocą pierwszych 6 bitów słowa rozkazowego, przełącznik „ton/szum" ustawiający jeden z dwu dostępnych generatorów (do jego kontroli służy kolejny, siódmy bit słowa rozkazowego), regulowany wzmacniacz określający jeden z 15 możliwych poziomów głośności dźwięku (kolejne 4 bity słowa rozkazowego), wreszcie 10 filtrów o regulowanych charakterystykach, modelujących transmitancję toru głosowego, do których sterowania wykorzystuje się pozostałe bity słowa rozkazowego (zależnie od wpływu na brzmienie sygnału poszczególne filtry są sterowane przy użyciu od 3 do 5 bitów). W celu ograniczenia wpływu małej częstotliwości aktualizacji danych (40 Hz) na jakość sygnału mowy zastosowano technikę interpolacyjną do zapewnienia płynnego procesu przechodzenia od jednych wartości danych do kolejno napływających. Interpolacja ta opiera się na technice predykcji liniowej używanej w analizie mowy do jej opisu i rozpoznawania. Jest to technika (patrz p. 4.5) wymagająca dużej mocy obliczeniowej (wystarczy powiedzieć, że realizacja 10 filtrów syntezatora TMS 5200 wymaga wykonania 200 000 dodawań i tyluż operacji mnożenia w ciągu jednej sekundy), ale dająca najlepsze rezultaty, jeśli idzie o płynność i naturalność sygnału mowy. Sygnał wyjściowy z syntezatora ma postać cyfrową i jest zbiorem słów 8-bitowych podawanych z częstotliwością 8 kHz. Pozwala to po zastosowaniu przetwor-

nika cyfrowo-analogowego na uzyskanie dobrej jakości sygnału mowy o parametrach lepszych niż w telefonii.

Syntezator TMS 5200 ma wiele dalszych udoskonaleń, pozwalających na jego wygodniejsze i bardziej oszczędne — z punktu widzenia systemu sterującego — wykorzystanie. W szczególności programowanie powtarzania pewnych sekwencji dźwiękowych, a także oszczędne kodowanie szeregu dźwięków (na przykład głosek szumowych, w których nie trzeba przestrajać generatora tonów ani tak dokładnie kształtować charakterystyki traktu głosowego) pozwala zaoszczędzić pamięć i upakować znacznie dłuższe fragmenty sygnału, niżby wynikało z przemnożenia częstotliwości 40 Hz przez długość słowa (50 bitów). Ponadto dla wygody sterowania pracą syntezatora dostępna jest „biblioteka" programów sterujących artykulacją 128 typowych odmian głosek, a także gotowe programy tworzenia typowych komunikatów.

Przytoczone rozwiązania systemów syntezy mowy należy traktować jako przykładowe. Syntezatory o różnych parametrach budują firmy: AMI, General Instruments, Hitachi, Intel, ITT, Matsushita, Philips, TSI i inne. Znane są też liczne i udane próby konstruowania systemów syntezy mowy polskiej, poczynając od syntezatorów parametrycznych, jak np. Synfor profesora Kacprowskiego, a na badaniach Politechniki Wrocławskiej kończąc. Nie ma możliwości ani celu omawiania wszystkich istniejących rozwiązań, szczególnie że podany wyżej przegląd podawał charakterystyczne cechy, wspólne dla większości konstrukcji. Ważny jest właściwie jeden wniosek. Proces generacji mowy jest na tyle dobrze znany, że można do jego modelowania użyć urządzeń technicznych zapewniających przy rozsądnych kosztach dobrą jakość syntetyzowanego dźwięku. Innymi słowy droga porozumiewania się maszyny z człowiekiem za pomocą głosu jest otwarta.

# 3

# Percepcja mowy

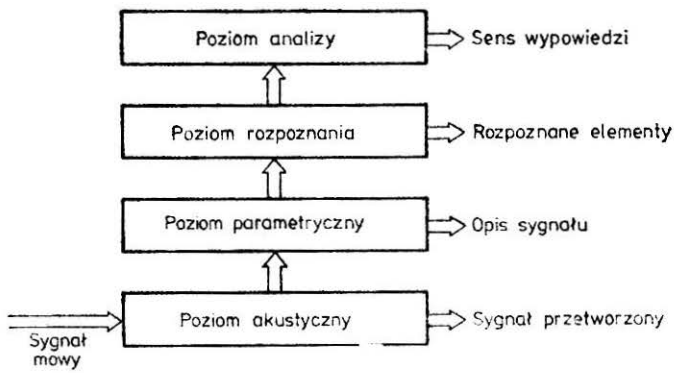
## 3.1. Wprowadzenie

Zagadnienie percepcji mowy jest znacznie bardziej złożonym problemem niż jej artykulacja i to zarówno z punktu widzenia opisu naturalnego analizatora słuchowego człowieka, jak i w zakresie technicznych systemów rozpoznających mowę. W rozdziale będą przedstawione wybrane elementy opisu struktury i funkcji ucha i systemu nerwowego analizującego dźwięki mowy, przy czym podobnie jak dla systemu artykulacji podstawą rozważań będzie model matematyczny. Wzmiankowane będą także systemy techniczne służące do automatycznego rozpoznawania mowy, z tym że ten ostatni problem znacznie szerzej będzie przedstawiony w rozdz. 6.

Złożoność zadania rozpoznawania mowy wynika — niezależnie od tego, czy rozpoznającym obiektem jest mózg człowieka, czy automat — z kilku podstawowych własności sygnału mowy jako nośnika informacji. Przedstawiona dalej lista problemów, z którymi trzeba się uporać budując system rozpoznawania mowy, pozwoli własności te przeanalizować w sposób systematyczny i uporządkowany, ułatwi rozumienie metod rozpoznawania mowy i wyjaśni, dlaczego tak trudno zbudować naprawdę efektywne metody kompresji informacyjnego nadmiaru sygnału mowy w telekomunikacji. Lista ta jest równocześnie wykazem problemów badawczych zarówno dla

biologów studiujących funkcjonowanie analizatora słuchowego człowieka, jak i dla inżynierów dążących do skonstruowania systemu technicznego receptora mowy.

Przystępując do budowy wspomnianej listy warto spojrzeć na proces rozpoznawania mowy z punktu widzenia teorii systemów i wyróżnić w nim kilka hierarchicznie powiązanych poziomów (rys. 3-1). Na podstawowym,




3-1. Uproszczona struktura systemu percepcji mowy. Takiego hierarchicznego schematu można się doszukać zarówno w urządzeniach technicznych służących do rozpoznawania mowy, jak i w naturalnym systemie biologicznym, służącym do percepcji mowy: uchu człowieka i współpracujących z nim strukturach nerwowych

akustycznym poziomie pozyskiwana jest informacja o rozpoznawanym sygnale dźwiękowym. Powstają przy tym między innymi następujące problemy:

1. W jakiej postaci należy sygnał wprowadzać do systemu?
2. Jeśli sygnał ma być wprowadzony bezpośrednio w formie przebiegu czasowego, to jak szerokie powinno być rozważane pasmo częstotliwości i wynikająca z niego częstotliwość próbkowania sygnału? Z jaką dokładnością odwzorowywać amplitudę sygnału, czy stosować równomierny, czy poddany kompresji rozkład poziomów dyskretyzacji amplitud? Jaką zastosować technikę kodowania? Czy i w jaki sposób dokonywać preemfazy sygnału? itp.
3. Jeśli sygnał jest wprowadzany w postaci przetworzonej, to jaka ma być reguła tego przetwarzania, aby nie tracić istotnej informacji, a równocześnie ograniczyć informacyjną pojemność sygnału, utrudniającą jego zmieszczenie w pamięci systemu rozpoznającego lub/i transmisję przez kanały telekomunikacyjne?
4. Jeśli przetwarzaniem, o którym mowa w punkcie 3, jest transformacja widmowa, to jak jej dokonywać (analogowo, z wykorzystaniem filtrów czy cyfrowo za pomocą algorytmu FFT)?
5. Ile powinno być i jak winny być rozmieszczone wyróżnione pasma częstotliwości?
6. Jak dobrać czas całkowania sygnału w poszczególnych pasmach (stałe czasowe demodulatorów za odpowiednimi filtrami)?
7. Z jaką częstotliwością próbkować sygnały na wyjściach poszczególnych filtrów?

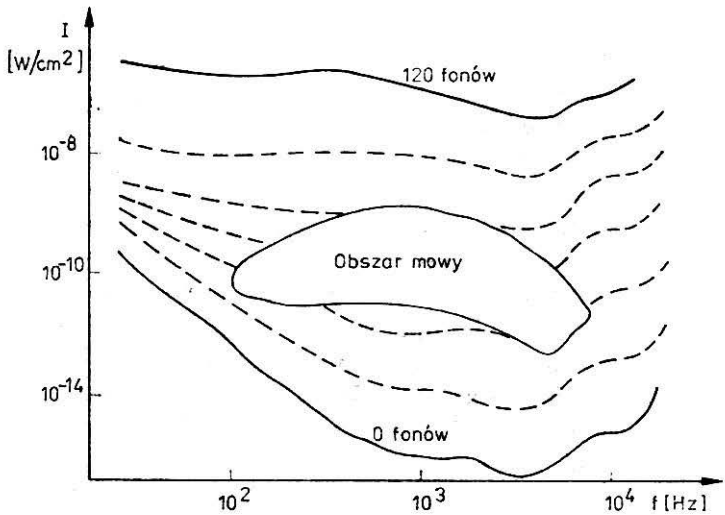
Dyskusja niektórych spośród wymienionych wcześniej problemów przeprowadzona będzie w rozdz. 4, w chwili obecnej należało je tylko wymienić, aby mieć świadomość, na jakie aspekty zwracać uwagę przy studiowaniu omawianych w tym rozdziale biologicznych i technicznych systemów analizy i percepcji mowy.

Oczywiście, zgodnie ze schematem podanym na rysunku 3-1, analiza akustyczna jest zaledwie wstępnym etapem w hierarchicznej strukturze przetwarzania sygnału mowy, która towarzyszy każdej próbie jej rozpoznawania. Następny, parametryczny poziom jest źródłem kolejnych problemów. Jego zadaniem jest opisanie sygnału mowy przez określenie jego parametrów, które pozwolą na jednoznaczną i pewną jego identyfikację, a równocześnie będą zawierać możliwie mało zbędnej (to znaczy nieprzydatnej przy rozpoznawaniu) informacji. Warto podkreślić, że omawiany poziom ma kluczowe znaczenie dla efektywności procesu rozpoznawania, gdyż wybór niewłaściwych cech spowoduje nieuchronnie bądź nieodwracalną stratę niezbędnych do rozpoznawania informacji, bądź w zbyt małym stopniu ochroni nas przed „zalewem" informacji zbytecznych. Niestety, ani teoria rozpoznawania, ani akustyka mowy nie dostarczają wystarczających przesłanek do wyboru najwłaściwszego zestawu cech. Z tego między innymi powodu tak wiele zainteresowania (i miejsca w książce) zajmuje problematyka modelowania procesu artykulacji mowy i badania nad naturalną percepcją mowy. Śledząc sterowanie procesu artykulacji oczekujemy bowiem odpowiedzi na pytanie, które własności sygnału są świadomie kształtowane, a które są wynikiem zbiegu okoliczności. Podobnie analizując proces rozpoznawania mowy przez ucho i mózg człowieka możemy odnotować własności sygnału, które w procesie tym odgrywają pierwszoplanową rolę, a następnie możemy oczekiwać, że oparcie technicznego systemu rozpoznawania na podobnych cechach jest racjonalnie uzasadnione. Istnieją bowiem przesłanki świadczące, że sygnał mowy w procesie swego formowania został tak ukształtowany, by optymalnie odpowiadał możliwościom naszego systemu percepcyjnego. Wszelkie dźwięki, jakie może wydawać narząd mowy, które jednak nie prowadzą do poprawnej percepcji przekazywanych treści, są z mowy eliminowane. Warto zwrócić uwagę, że przedstawiona tu teza jest odmienna od rozpowszechnionego i łatwego do zakwestionowania poglądu, że to słuch człowieka jest optymalnie dostosowany do odbioru mowy. Argumentacja na rzecz tezy o dopasowaniu słuchu do mowy, a nie odwrotnie, jest problematyczna. Zagadnienie to staje się zupełnie jasne, jeśli odpowie się na podstawowe pytanie, co było wcześniej: słuch z jego własnościami, czy mowa z jej parametrami potrzebnymi przy rozpoznawaniu. Na rysunku 3-2 pokazano położenie amplitudowo-częstotliwościowych charakterystyk sygnału mowy na tle obszaru najlepszego słyszenia człowieka.

Jak wynika z przedstawionych uwag, na parametrycznym poziomie procesu rozpoznawania mowy rodzą się kolejne problemy i kolejne trudności. Oto niektóre z nich*):

*) Zachowano ciągłą numerację rozważanych zagadnień badawczych dla zaznaczenia faktu, że stanowią one w istocie jedną całość a podział na poziomy ma charakter umowny.

8. Jakie parametry wybrać, aby odpowiadały wymaganiom minimalnej reprezentacji sygnału i jego pewnej identyfikacji?
9. Jak wyznaczać wybrane parametry opierając się na posiadanych środkach technicznych przyjętej na poprzednim poziomie reprezentacji sygnału i w warunkach obecności zakłóceń zniekształcających obraz sygnału?




3-2. Charakterystyki określające czułość systemu słuchowego człowieka z zaznaczonym obszarem, jaki na płaszczyźnie natężenia dźwięku $I$ oraz jego częstotliwości $f$ zajmuje naturalny sygnał mowy. Łatwo zauważyć, że charakterystyki sygnału mowy są takie, aby jego percepcja odbywała się maksymalnie łatwo. Zarówno wymiar częstotliwościowy, jak i amplitudowy sygnału mieści się dokładnie w rejonie najlepszego słyszenia

10. Czy parametry wyznaczone w procesie opisu mowy mają własności wymagane przez procedury identyfikacji elementów wypowiedzi, czy też należy je dodatkowo poddać transformacji?
11. Jakiego rodzaju transformacja parametrów (jeśli uznano celowość jej stosowania) może zapewnić optymalną geometrię przestrzeni obiektów z punktu widzenia metod rozpoznawania?
12. W jaki sposób i z użyciem jakich środków technicznych dokonywać transformacji parametrów? W szczególności, czy wykorzystywać obliczenia realizowane techniką cyfrową w głównym komputerze, czy też raczej stosować układy przekształcające, realizowane analogowo lub z użyciem specjalizowanych procesorów?
13. Czy wyniki procesów wydobywania parametrów (i ewentualnego ich transformowania) zapisywać w pamięci urządzenia rozpoznającego w formie bezpośredniej (łatwiejszej do dalszych obliczeń), czy w formie zakodowanej, wykorzystując metody przystosowane do maksymalnej oszczędności miejsca w pamięci operacyjnej komputera?
Rozwiązanie przytoczonych problemów (lub — co się niestety częściej praktykuje — arbitralne podjęcie potrzebnych decyzji), nie kończy prezentowanej listy trudności, które trzeba pokonać, lecz prowadzi do kolejnych problemów, związanych tym razem z trzecim poziomem systemu rozpoznawania mowy — z systemem rozpoznawania elementów sygnału.
Sytuacja, która stanowi punkt wyjścia do rozważań na tym poziomie, może być scharakteryzowana w sposób następujący. Sygnał mowy został już zarejestrowany i przetworzony do postaci zbioru odpowiednich parametrów

(pierwotnych lub przetransformowanych). Ponieważ jednak sygnał zmienia się, więc wyliczone cechy nie pozostają stałe, lecz zmieniają się w czasie, tworząc w przestrzeni parametrów złożone trajektorie. Trajektorie takie mogą być rozpoznawane w całości, tworząc system rozpoznawania kompletnych wypowiedzi, jednak znacznie bardziej celowe jest rozpoznawanie elementów wypowiedzi i składanie całości z poszczególnych rozpoznanych segmentów.

Systemy „całościowe" obecnie są stosowane do rozpoznawania ograniczonego słownika, jedynie w tych systemach komercyjnych, w których główną rolę odgrywa szybkość działania, a nie wysoka jakość procesu rozpoznawania i uniwersalność zastosowań. Z tego względu w dalszych rozważaniach skupimy się na omawianiu systemów działających na zasadzie etapowej: najpierw rozpoznawane są oddzielne segmenty, a dopiero później ich ciągi są identyfikowane z określonymi wypowiedziami.

Przy takim postawieniu zadania możliwe jest jednak wyróżnienie dalszych problemów, z których niektóre (numerując je kolejno dalej) wyszczególniono niżej:

14. Jakie segmenty mają być podstawą rozpoznawania (wyrazy, sylaby, głoski, tzw. mikrofonemy — patrz rozdz. 5 — pojedyncze próbki czasowe sygnału)?
15. Jak dokonać podziału ciągłego sygnału mowy na wskazane segmenty?
16. Jakimi metodami rozpoznawać wydzielone segmenty? (W teorii rozpoznawania obrazów istnieje kilkadziesiąt różnych możliwych algorytmów, a wiele spośród tych algorytmów ma swoje odmiany — por. rozdz. 5).
17. Jakimi metodami scalać segmenty w całe wypowiedzi?
18. Jak korygować błędy rozpoznawania?

Odpowiedź na przytoczone pytania daje w efekcie konkretną realizację systemu rozpoznawania wypowiedzi na płaszczyźnie leksykalnej (identyfikacja elementów słownika). Problem rozpoznawania mowy na tym jednak się nie kończy. Pozostaje analiza syntaktyczna wypowiedzi i jej semantyczna identyfikacja, aby ustalić sens wypowiedzianego polecenia i adekwatnie do niego działać (por. rozdz. 5).

## 3.2. Zbiorczy model niższych pięter systemu słuchowego człowieka

### 3.2.1. Wstęp

W stosunkowo licznej literaturze dotyczącej prób modelowania systemu słuchowego człowieka przeważają publikacje dotyczące analizy funkcjonowania tego systemu, głównie ucha wewnętrznego i narządu Cortiego, a także wybranych fragmentów niższych pięter części nerwowej tego systemu. W dalszym rozdziale opisano strukturę modelu zbiorczego, obejmującego całość mechanicznej składowej systemu oraz niższe piętra (do *nucleus cochlearis* włącznie) części nerwowej. Model ten jest oparty na wynikach

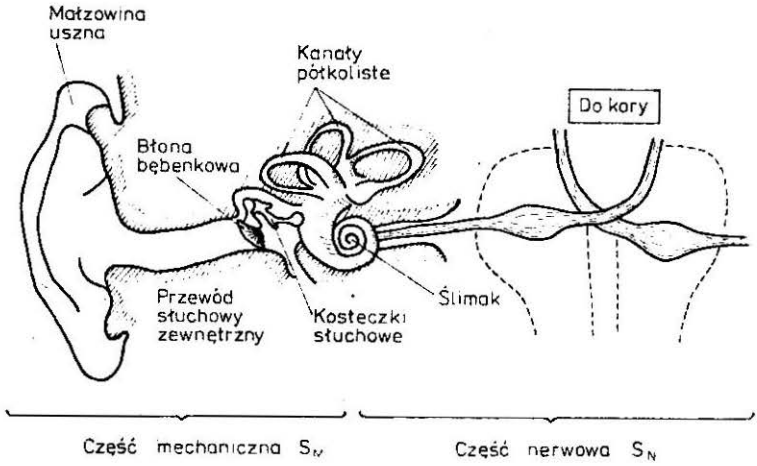
wcześniejszych opracowań dotyczących prób modelowania fragmentów systemu i poszukiwań optymalnego (z punktu widzenia symulacji komputerowej) modelu poszczególnych elementów systemu: ucha zewnętrznego i środkowego, błony podstawowej ucha wewnętrznego, komórek rzęskowych narządu Cortiego, spiro- i ortoneuronów zwoju spiralnego, komórek nerwowych tworzących jądra ślimakowe oraz hipotetycznej struktury sieci neuronowej jąder ślimakowych, spełniającej założoną funkcję „wyostrzania" (polepszania selektywności). Ma on na celu sprawdzenie współdziałania badanych uprzednio oddzielnie fragmentów systemu. Model oprogramowano w postaci pakietu programów symulacyjnych i pomocniczych (przeznaczonych głównie do graficznej prezentacji wyników za pomocą plottera Calcomp) w językach MIMIC i FORTRAN EXTENDED i badano z wykorzystaniem komputera Cyber 72. Wszystkie prezentowane w tym rozdziale wykresy pochodzą z wyników symulacji i zostały wykonane za pomocą wspomnianego plottera.

## 3.2.2. Założenia i ograniczenia przyjęte przy budowie modelu

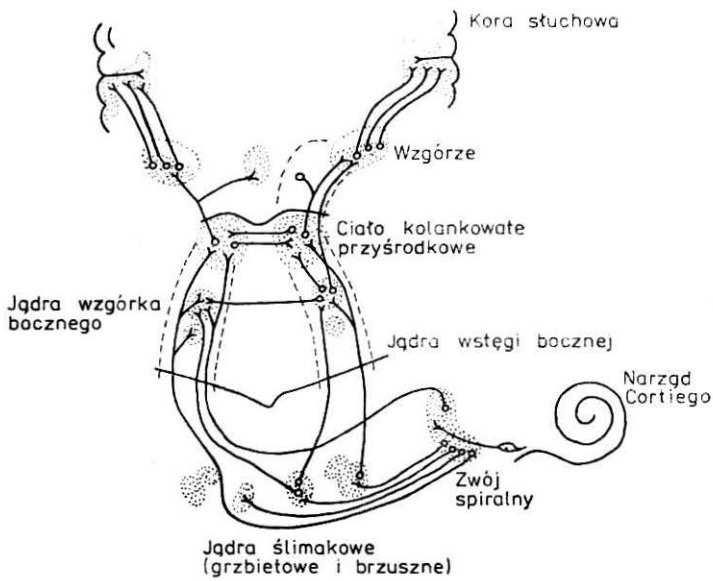
System słuchowy człowieka, drugi pod względem złożoności po systemie analizatora wzrokowego, jest zbyt skomplikowany i zbyt mało poznany, aby mógł być przedmiotem modelowania uwzględniającego wszystkie aspekty jego działania. Przystępując do budowy modelu trzeba więc ograniczyć zakres rozważanych zjawisk, świadomie rezygnując z części znanych faktów na rzecz dostosowania go do założonego celu. W przypadku omawianego systemu słuchowego celem modelowania jest poznanie procesów, przetwarzania i redukcji ilości informacji, zachodzących w systemie słuchowym w celu wykorzystania ich przy budowie automatycznych urządzeń rozpoznających mowę. Z tego powodu skupiono uwagę wyłącznie na przekazywaniu informacji dźwiękowej pomijając modelowanie wszystkich innych zjawisk związanych z funkcjonowaniem systemu słuchowego, a także brano pod uwagę wyłącznie drogi aferentne, prowadzące od ucha do mózgu, pomijając symulowanie funkcji dróg aferentnych, głównie regulacyjnych (m.in. pominięto w modelu funkcje mięśni napinacza błony bębenkowej oraz strzemiączkowego) oraz zrezygnowano z modelowania funkcji (słabo zresztą znanych) autonomicznego systemu nerwowego, którego zakończenia synaptyczne w komórkach rzęskowych ucha wewnętrznego wydają się odgrywać istotną rolę adaptacyjną.

## 3.2.3. Struktura modelu

Na rysunku 3-3 przedstawiono budowę części mechanicznej systemu słuchowego, na rys. 3-4 — schemat powiązań w obrębie części nerwowej tego systemu, a na rys. 3-5 — powiązania między strukturą omawianego modelu a przedstawioną w uproszczeniu budową systemu słuchowego.




3-3. Schematyczny obraz części mechanicznej systemu słuchowego. Punktem styku z częścią nerwową jest ślimak, a dokładniej — znajdujący się w nim narząd Cortiego, przekazujący informacje o dźwięku do neuronów zwoju spiralnego




3-4. Schematyczny, bardzo uproszczony obraz powiązań w zakresie nerwowej części analizatora słuchowego. Informacja dźwiękowa, zarejestrowana przez receptory zlokalizowane w ślimaku, jest przekazywana do neuronów zwoju spiralnego, następnie do jąder ślimakowych (grzbietowych i brzusznych), wstęgi bocznej i wzgórka bocznego, ciała kolankowatego przyśrodkowego wzgórza i do korowych ośrodków mowy w płacie skroniowym

Poddawaną modelowaniu część systemu słuchowego można rozpatrywać jako parę

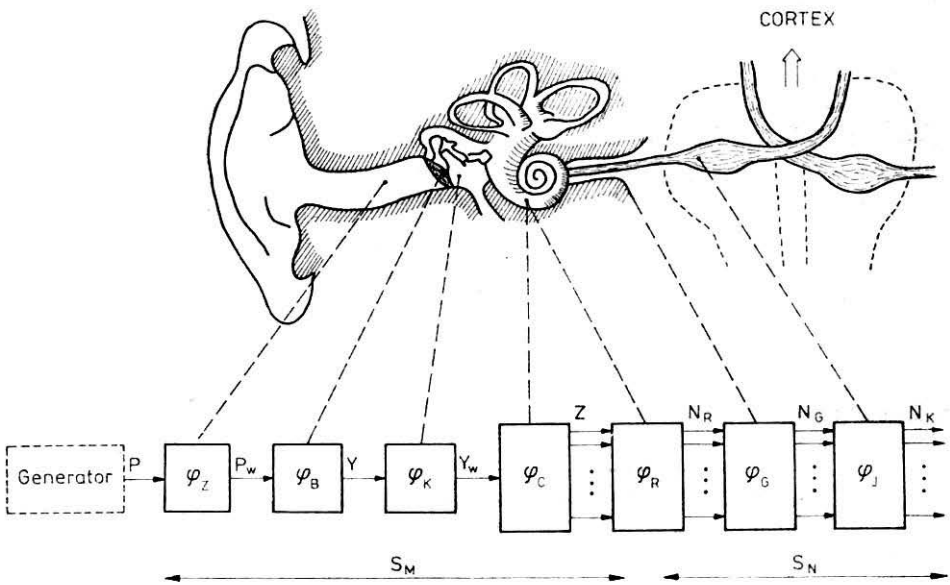
$$S_s = \langle S_m, S_n \rangle \tag{3.1}$$

gdzie $S_m$ jest częścią mechaniczną, a $S_n$ częścią nerwową systemu. System realizuje odwzorowanie

$$\varphi_s : p \to N_k \tag{3.2}$$

gdzie $p \subset E^T$ jest zbiorem funkcji czasu, reprezentujących przebiegi ciśnienia akustycznego na zewnątrz przewodu słuchowego zewnętrznego, a $N_k \subset X^{N \times T}$ jest dynamicznym rozkładem pobudzeń neuronów ostatniej warstwy jąder ślimakowych (ostatniego uwzględnionego w modelu elementu drogi nerwowej).

Użyte w przytoczonych określeniach zbiory scharakteryzować można w następujący sposób. Zbiór chwil czasu $T = \{t: t \in \mathcal{R} \wedge t \geqslant 0\}$, zbiór chwilowych wartości ciśnienia akustycznego $E = \{e: e \in \mathcal{R} \wedge 0 \leqslant e_d \leqslant e \leqslant e_g\}$, zbiór chwilowych wartości pobudzeń neuronów (które można utożsamiać z chwilową częstotliwością impulsów w ustalonym punkcie aksonu)




3-5. Powiązanie elementów rzeczywistej struktury analizatora słuchowego i omawianych w książce elementów jego modelu. Model nie obejmuje wszystkich elementów rzeczywistego systemu. Uzupełnieniem modelu jest generator, wykorzystywany do badania jego własności; generator podobnie jak pozostałe elementy modelu jest realizowany w postaci odpowiedniego modułu programu symulacyjnego dla komputera. Opis pokazanych na rysunku bloków będzie sukcesywnie wprowadzany w kolejnych punktach

$X = \{x;\ x \in \mathcal{R} \wedge 0 \leqslant x \leqslant x_g\}$ oraz zbiór numerów komórek nerwowych (w pewnej arbitralnie przyjętej, ale ustalonej numeracji) $N \in \mathcal{N}$.
Rozważany model systemu realizuje odwzorowanie

$$\varphi_s^*: p^* \to N_k^* \tag{3.3}$$

przy czym (pomijając efekt kwantowania zbioru liczb rzeczywistych wynikający ze skończonej długości komórki pamięci używanego komputera) możemy zapisać: $p^* \subset E^{T^*}$ oraz $N_k^* = x^{N^* \times T^*}$, gdzie $T^* = \{\eta: \eta \in \in \mathcal{N} \wedge \eta \Delta \in T \wedge \Delta \in T\}$ jest zbiorem numerów równoodległych chwil czasowych (model jest typu synchronicznego), zaś $N^* \subset N$ jest zbiorem numerów nielicznych wybranych komórek nerwowych, których działanie jest przedmiotem symulacji. Różnica między odwzorowaniami $\varphi_s$ i $\varphi_s^*$ nie ogranicza się wyłącznie do odmiennego charakteru zbiorów $p$ i $p^*$ oraz $N_k$ i $N_k^*$, gdyż model odwzorowuje jedynie wybrane funkcje oryginału.
Przyjmując określony funkcjonał $Q: \mathcal{R}^T \times \mathcal{R}^T \to \mathcal{R}$ można dla każdego $e(t) \in p$ zapisać implikację:

$$Q[e(t), e^*(\eta)] < \varepsilon_1 \Rightarrow \bigwedge_{\xi \in N^*} Q[x(t, \xi), x^*(\eta, \xi)] < \varepsilon_2 \tag{3.4}$$

gdzie $\varepsilon_1$ i $\varepsilon_2$ są ustalonymi wartościami. Zgodnie z podziałem anatomicznym systemu funkcję $\varphi_s$ można przedstawić jako złożenie szeregu odwzorowań. Z zależności (3.1) wynika, że

$$\varphi_s = \varphi_n \cdot \varphi_m \tag{3.5}$$

gdzie $\varphi_m$: $p \to z$, zaś $\varphi_n$: $z \to N_k$.

Zbiór $z \subset Y_w^{L \times T}$ reprezentuje drgania błony podstawnej ucha wewnętrznego, będące funkcją czasu $T$ i odległości rozważanego punktu błony od helikotremy $L = \{l\colon 0 \leqslant l \leqslant l_{\max} \wedge l \in \mathscr{R}\}$. Zbiór wielkości wychyleń rozważanych punktów błony $Y_s = \{y_s\colon |y_s| \leqslant y_{sm} \wedge y_s \in \mathscr{R}\}$ ma podobną charakterystykę jak zbiory $Y$ (chwilowych wartości wychyleń centralnego punktu błony bębenkowej) oraz $Y_w$ (chwilowych wartości wychyleń podstawy strzemiączka w okienku owalnym ślimaka), przy czym różnią się jedynie ograniczenia amplitud: $y_m > y_{wm} > y_{sm}$. W modelu zbiór $z$ jest zastąpiony zbiorem $z^* \subset \subset Y_w^{N^* \times T^*}$. Należy zwrócić uwagę, że zbiór $N^*$ stanowiący dyskretną wersję zbioru ciągłego $L$ jest identyczny ze zbiorem $N^*$ występującym w definicji $N_k$. Jest to bardzo istotne ograniczenie modelu, w ogromnym stopniu upraszczające konstrukcję odwzorowania $\varphi_n$, będącego modelowym odwzorowaniem funkcji nerwowej części systemu $\varphi_n$.

Odwzorowania $\varphi_n$ oraz $\varphi_m$ mają złożony charakter; dla ich uproszczenia dokonano ich dalszej dekompozycji opierając się na kryteriach anatomicznych

$$\varphi_m = \varphi_c \cdot \varphi_k \cdot \varphi_b \cdot \varphi_z \tag{3.6}$$

gdzie:

$\varphi_z\colon p \to p_w$

$p_w \subset E^T$ jest zbiorem czasowych przebiegów ciśnienia akustycznego na wysokości błony bębenkowej w głębi przewodu słuchowego wewnętrznego,

$\varphi_b\colon p_w \to w$

$w \subset Y^T$ jest zbiorem przebiegów czasowych drgań błony bębenkowej,

$\varphi_k\colon w \to w_w$

$w_w \subset Y_w^T$ jest zbiorem chwilowych przebiegów drgań podstawy strzemiączka ucha środkowego w okienku owalnym ślimaka,

$\varphi_c\colon w_w \to z$

Identyfikując poszczególne odwzorowania z odpowiednimi elementami anatomicznymi systemu słuchowego możemy stwierdzić, że $\varphi_z$ odpowiada funkcjom ucha zewnętrznego, $\varphi_b$ — błony bębenkowej, $\varphi_k$ — systemu kosteczek słuchowych ucha środkowego zaś $\varphi_c$ — ucha wewnętrznego, a głównie błony podstawowej ślimaka. Realizacja poszczególnych odwzorowań w modelu wprowadza dyskretyzację zbioru $T$ do postaci $T^*$ oraz wymaga uproszczenia odpowiednich zależności do postaci nadającej się do modelowania. Wydaje się, że najbardziej istotne uproszczenia są przy tym wprowadzane do odwzorowania $\varphi_k$.

Analogicznie do podziału odwzorowania $\varphi_m$ jest prowadzona dekompozycja odwzorowania $\varphi_n$ na elementy odpowiadające kolejno warstwie receptorów

(komórek rzęskowych) $\varphi_r$, spiro- i ortoneuronów zwoju spiralnego $\varphi_g$ oraz neuronów jąder ślimakowych $\varphi_j$:

$$\varphi_n = \varphi_j \cdot \varphi_g \cdot \varphi_r \tag{3.7}$$

gdzie:

$\varphi_r: z \to N_r$

$N_r \subset X^{N \times T}$ jest zbiorem dynamicznych rozkładów pobudzeń na poszczególnych receptorach,

$\varphi_g: N_r \to N_g$

$N_g \subset X^{N \times T}$ jest zbiorem dynamicznych rozkładów pobudzeń neuronów zwoju spiralnego.

$\varphi_j: N_g \to N_k$

Realizacja tych odwzorowań w modelu sprowadza się, obok dyskretyzacji czasu, do wprowadzenia we wszystkich rozważanych elementach drogi słuchowej tej samej liczby tak samo ponumerowanych elementów symulujących funkcjonowanie komórek nerwowych i receptorowych. Liczba ta, wynosząca w konkretnym modelu 30, jest znacznie mniejsza od liczebności odpowiednich zbiorów w rzeczywistym obiekcie. Jest to kolejne, bardzo istotne ograniczenie modelu.

## 3.2.4. Model części mechanicznej systemu słuchowego

Omawiając w niniejszym punkcie poszczególne wymienione wyżej odwzorowania będziemy stosować dla uproszczenia notacji oraz zwiększenia czytelności zapis operatorowy Laplace'a (dotyczy to odwzorowań składających się na $\varphi_m$) i będziemy przytaczać od razu transmitancje odpowiednich członów (przy założeniu ich liniowości). Innymi słowy opisując dowolne odwzorowanie $\varphi: a \to b$, gdzie $a \subset u^T$, zaś $b \subset v^T$, zapisywać je będziemy jako $\hat{\varphi} = G(s) \stackrel{\text{def}}{=} \frac{A}{B}$, gdzie: $A \subset U^S$, $B \subset V^S$, $S = \{s \in \mathscr{C}\}$. Zależności między $a$ i $A$ oraz $b$ i $B$ określa znany wzór całkowy Laplace'a:

$$A(s) = \int_0^\infty a(t) \mathrm{e}^{-st} \mathrm{d}t$$

Omawiając rolę i działanie poszczególnych fragmentów części mechanicznej systemu słuchowego należy pamiętać, że jego rola w całości analizatora jest pomocnicza. Wspomniane elementy są konieczne ze względu na funkcjonowanie całego systemu, jednak z punktu widzenia procesów przetwarzania sygnałów wprowadzają zniekształcenia, deformując między innymi widmo podlegającego analizie dźwięku. Ucho zewnętrzne, reprezentowane w modelu przez odwzorowanie $\varphi_z$, to małżowina uszna wraz z przewodem słuchowym zewnętrznym, tworzące razem rodzaj elastycznej tuby, wprowadzającej dźwięk z otoczenia do błony bębenkowej. U wielu zwierząt jedną z czynności wskazanego układu jest polepszenie kierunkowych charakterystyk słuchu; dotyczy to zwłaszcza tych ssaków, u których małżowina uszna ma duże rozmiary i jest ruchoma. U człowieka funkcja ta ma znaczenie szczątkowe.

Zmiany poziomu sygnału, spowodowane kierunkiem jego docierania w stosunku do małżowiny usznej nie przekraczają kilku decybeli, są więc mało przydatne przy lokalizacji źródła dźwięku*). Ważniejsza jest rola przewodu słuchowego zewnętrznego polegająca na ochronie delikatnej i łatwej do uszkodzenia błony bębenkowej w głębi wąskiego kanału, który dodatkowo zapewnia niezbędny dla pracy błony „mikroklimat" (stabilna temperatura i wilgotność). Niestety, ceną za ten komfort jest deformacja struktury częstotliwościowej sygnału, gdyż kanał słuchowy jest rezonatorem.

Jeszcze bardziej skomplikowane jest uwzględnianie własności ucha środkowego. Zespół kosteczek słuchowych: młoteczek, kowadełko i strzemiączko przekazuje drgania błony bębenkowej do okienka owalnego ślimaka, tworząc złożony układ kinematyczny o wielu stopniach swobody i skomplikowanych własnościach dynamicznych. Rola tego układu sprowadza się do dopasowania impedancji środowiska, z którego fala dźwiękowa nadchodzi, do impedancji środowiska, w którym fala dźwiękowa będzie się dalej rozprzestrzeniać. U zwierząt żyjących w wodzie problem ten nie występuje, gdyż impedancja akustyczna płynów wypełniających ich ucho wewnętrzne jest praktycznie taka sama, jak impedancja środowiska. Z tego powodu narządy słuchu tych zwierząt są bardzo uproszczone, a nawet bywają zredukowane do postaci receptorów skórnych. Natomiast u człowieka, podobnie jak u większości**) zwierząt lądowych, konieczne jest dopasowanie warunków propagacji fali dźwiękowej w płynie wypełniającym ślimak ucha wewnętrznego do warunków rozchodzenia się dźwięków w powietrzu. Energia fali powinna być przekazana z jednego ośrodka do drugiego pomimo drastycznych różnic gęstości, sprężystości, tłumienności i bezwładności obydwu wymienionych środowisk. Brak takiego dopasowania powoduje, że fala dźwiękowa rozchodząca się w jednym ze wskazanych środowisk nie może przedostać się do drugiego, gdyż ulega na ich granicy niemal 100% odbiciu. Jest to między innymi powód rozpowszechnionego błędnego mniemania o braku możliwości wydawania dźwięków przez ryby i o ciszy głębi oceanicznych.

Upośledzenie mechanizmów dopasowania struktury ucha środkowego powoduje przeciętne podwyższenie progu słyszalności o ponad 40 dB, co jest równoważne niemal całkowitej głuchocie. Dopasowanie impedancji zachodzi z jednej strony w przekładni mechanicznej, gdyż praca kosteczek słuchowych jako systemu dźwigni powoduje ok. 8-krotne zwiększenie ciśnienia akustycznego, z drugiej strony ze względu na stosunek powierzchni błony bębenkowej i błonki zamykającej okienko owalne ślimaka ciśnienie to zwiększa się dodatkowo w stosunku ok. 1:15. Łącznie oba mechanizmy zwiększają ciśnienie akustyczne w perylimfie ślimaka w stosunku do ciśnienia w przewodzie słuchowym zewnętrznym w stosunku 1:100, zmniejszając oczywiście w identycznej proporcji prędkość objętościową. Układ, o którym

*) Lokalizacja ta przebiega u człowieka opierając się na słyszeniu dwuusznym, na podstawie różnic fazowych sygnałów.

**) Wyjątek stanowią owady, których narządy słuchu (rozmieszczone u różnych gatunków w najbardziej nieoczekiwanych miejscach na głowie, korpusie i odnóżach) są zazwyczaj przetwornikami akcelerometrycznymi pracującymi bez przetwarzania impedancji.

mowa, jest więc istotnie biernym systemem dopasowania impedancji, a nie aktywnym systemem zwiększającym moc sygnału.

Układ ucha środkowego ma jeszcze wiele własności, o których można tu jedynie skrótowo wspomnieć. I tak za pomocą mięśni: napinacza błony bębenkowej i strzemięczkowego możliwe jest takie wpływanie na pracę systemu, by nadchodząca fala dźwiękowa była tłumiona — w stopniu potrzebnym do adaptacji słuchu do dźwięków o dużym poziomie natężenia. Tylko dzięki temu mechanizmowi słuch może pokrywać swoim zakresem czułości (ogromny!) obszar natężenia dźwięku — ponad 120 dB. Dalej, jama bębenkowa mająca kontakt z powietrzem atmosferycznym poprzez trąbkę słuchową, otwierającą się w części nosowej gardła, pełni funkcję układu wyrównującego spoczynkowe ciśnienie po obydwu stronach błony bębenkowej, która w przeciwnym przypadku może ulegać naprężeniom*) utrudniającym percepcję słuchową. Na koniec układ ucha środkowego (a przynajmniej ostatnia z kosteczek słuchowych — strzemiączko) uczestniczy w słyszeniu na drodze tzw. przewodnictwa kostnego.

Wszystkie wymienione układy, a zwłaszcza elementy ucha środkowego, w tym łańcuch kosteczek słuchowych, wnoszą do percepowanego sygnału dźwiękowego dynamiczne zniekształcenia i zakłócenia, których charakter ujęty będzie w przytoczonym dalej modelu.

W wielu próbach opisu i modelowania systemu słuchowego przyjmuje się, że $\hat{\varphi}_z = 1$. Jednak analiza budowy ucha zewnętrznego pozwala upewnić się, że podejście takie jest błędne. Przewód słuchowy zewnętrzny (*meatus acousticus ext*) ma wprawdzie dość złożony kształt, można go jednak utożsamiać (z wystarczającą do celów modelowania dokładnością) ze sztywną rurką o przekroju kołowym i długości $D = 27$ mm, zamkniętą na końcu sztywną przegrodą. W tym przypadku

$$\hat{\varphi}_z = \frac{\omega_0^2}{s^2 + 2\xi s\omega_0 + \omega_0^2} \tag{3.8}$$

gdzie: $\omega_0 = \dfrac{\pi c}{2D}$

$c$ — prędkość fali dźwiękowej w powietrzu.

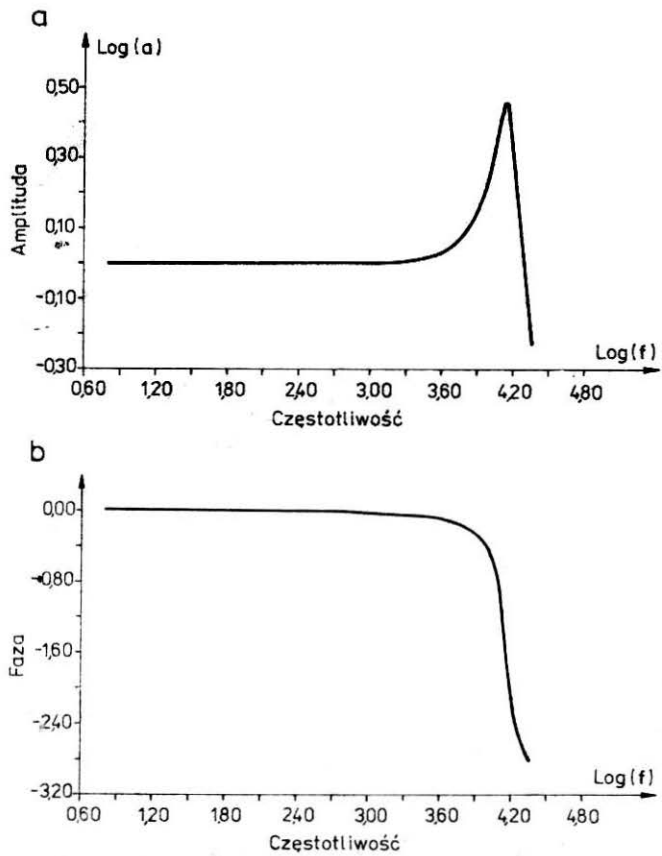
W modelu przyjęto $\omega_0 = 2\pi$ 2300 Hz oraz $\xi = 0{,}2$ uzyskując charakterystykę częstotliwościową przedstawioną na rys. 3-6. Charakterystyka ta jest zgodna z podawanymi przez wielu badaczy charakterystykami empirycznymi.

Odnośnie odwzorowania $\hat{\varphi}_b$ istnieje stosunkowo niewiele danych empirycznych i równie mało prób modelowania. W omawianym modelu przyjęto arbitralnie $\hat{\varphi}_b = K = \text{const}$, jakkolwiek jest to niewątpliwie znaczne uprosz-

*) Naprężenia te powstają przy zmiennym ciśnieniu zewnętrznym, na przykład podczas lotu samolotem lub przy nurkowaniu. Rola trąbki słuchowej nie ogranicza się jednak wyłącznie do tych przypadków, gdyż powietrze zawarte w zamkniętych jamach ciała ulega tzw. wsysaniu, rozpuszczając się w krążącej krwi i limfie. W tej sytuacji niedrożność trąbki słuchowej (wywołana np. obrzmieniem błony śluzowej ujścia gardłowego) prowadzi w krótkim czasie do upośledzenia słuchu na skutek różnicy ciśnień po obu stronach błony bębenkowej.

czenie. Odwzorowanie $\hat{\varphi}_k$ bywało przedmiotem badań symulacyjnych, dostępnych jest także stosunkowo wiele danych na temat charakterystyk częstotliwościowych ucha środkowego, wobec tego wybór transmitancji $\hat{\varphi}_k$ był silnie zdeterminowany wynikami wcześniejszych prac. W modelu stosowano postać transmitancji

$$\hat{\varphi}_k = \frac{C_0}{(s+a)[(s+a)^2+b^2]} \tag{3.9}$$



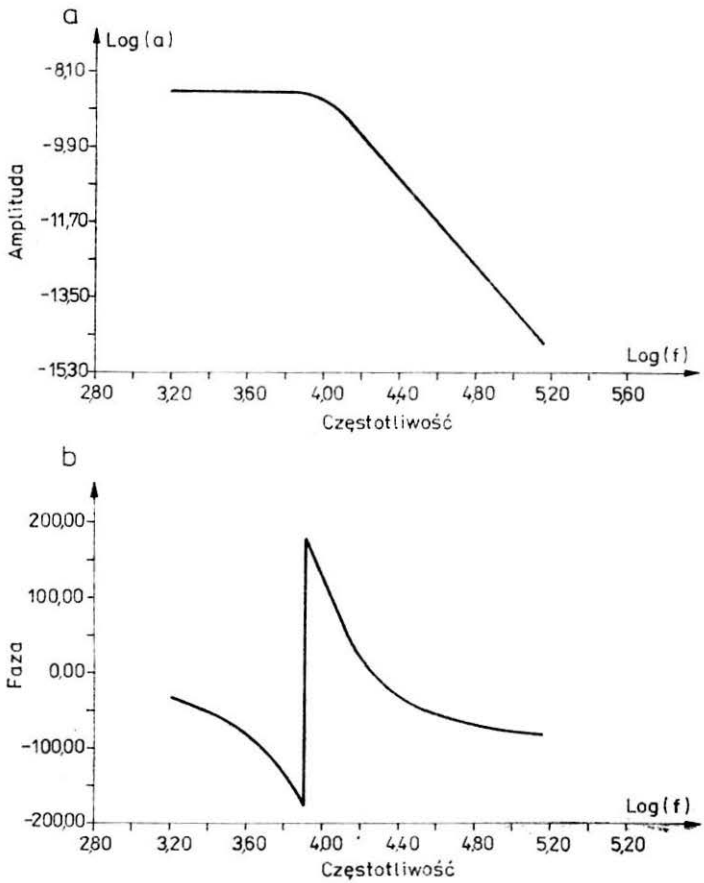

3-6. Uzyskane z symulacji komputerowej charakterystyki częstotliwościowe (amplitudowa (a) i fazowa (b)) ucha zewnętrznego

z parametrami $b = 2a = 2\pi\ 1500$ oraz $C_0 = 0{,}3a(a^2+b^2)$. Charakterystykę częstotliwościową modelu przedstawiono na rys. 3-7.

Odnośnie odwzorowania $\hat{\varphi}_c$ istnieje najwięcej wątpliwości, a jego realizacja w postaci modelu nastręcza największych trudności.

Przewód ślimakowy będący jednym z kilku kanałów ucha wewnętrznego spełnia rolę analizatora dźwięku. Światło kanału ślimaka o średnicy około 3 mm jest przedzielone poprzecznie na całej długości blaszką spiralną, tworząc odcinek górny i dolny. Elastyczna część blaszki jest nazywana błoną podstawną. Od blaszki spiralnej biegnie ukośnie do zewnętrznej ściany ślimaka błona Reisnera, tworząc tym samym trzy kanały: schody przedsionka, schody bębenka oraz przewód ślimakowy. Kanały te łączą się ze sobą na

szczycie ślimaka małym otworkiem zwanym helikotremą. Błona podstawna spełnia rolę analizatora częstotliwościowego drgań akustycznych.
Na błonie znajduje się skupisko komórek zwane narządem Cortiego. Wyróżniamy w nim dwa typy komórek słuchowych (rzęsatych): wewnętrzne i zewnętrzne. Komórki rzęsate wewnętrzne ułożone są w jednym rzędzie



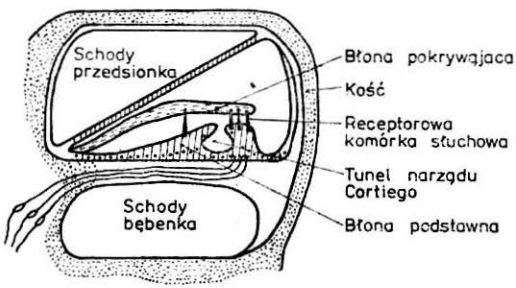

3-7. Charakterystyki amplitudowo-częstotliwościowe (a) i fazowo-częstotliwościowe (b) struktur ucha środkowego (uproszczone w stosunku do rzeczywistości)

i w pewnej odległości od komórek rzęsatych zewnętrznych tworzących trzy, cztery lub pięć równoległych rzędów. Każda komórka rzęsata jest zaopatrzona w szczecinowate rzęski o różnej długości. W stanie spoczynku dłuższe rzęski dotykają swymi końcami żelowatej struktury zwanej błoną pokrywającą. Błona ta wraz z zespołem komórek tworzy funkcjonalnie zamknięty system przetwarzający sygnały mechaniczne na sygnały nerwowe. Istnienie w analizatorze słuchowym dużej różnicy potencjałów między wnętrzem komórki rzęsatej a kanałem ślimakowym (około 160 mV) stwarza możliwość sterowania systemu na drodze elektrycznej. Struktura ucha wewnętrznego, przedstawiona na rys. 3-8 i 3-9, wskazuje na ogromną rolę, jaką w tym fragmencie systemu słuchowego pełni błona podstawna. Jej funkcja polega na zamianie zmian ciśnienia akustycznego w perylimfie wypełniającej kanały ślimaka (patrz dalej) wywołanych drganiami podstawy strzemiączka w okien-

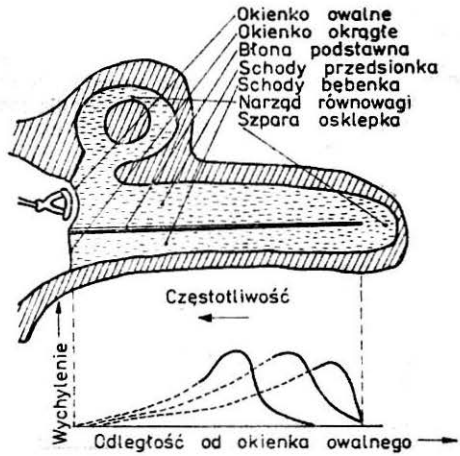
ku owalnym ślimaka $W_w$ na poprzeczne oscylacje poszczególnych punktów błony. Można to opisać następująco:

$$\varphi_c : W_w \to z \tag{3.10}$$

Błona ta wzdłuż kanału ślimaka zmienia swą szerokość (w stosunku 1:12,5), masę (1:50) i sztywność ($1:10^5$), w wyniku czego drgania mechaniczne strzemiączka w okienku owalnym ślimaka, przenosząc się poprzez płyn wypełniający ślimak (perylimfę), wywołują nierównomierne oscylacje poszczególnych punktów błony (część dolna rys. 3-9).




3-8. Ucho wewnętrzne — przekrój poprzeczny jednego z kanałów ślimaka z zaznaczeniem najważniejszych struktur. Dla lepszej rozróżnialności szczegółów powiększono znacznie w stosunku do rzeczywistości narząd Cortiego. Podstawowym elementem jest tu błona podstawna, dzieląca schody przedsionka i przewód ślimakowy od schodów bębenka. Jej drgania, wymuszone rejestrowaną falą dźwiękową, przekazywane są przez receptorowe komórki słuchowe do części nerwowej systemu



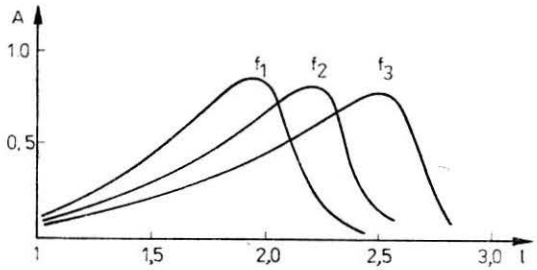

3-9. Uproszczony schemat ucha wewnętrznego — ślimak narysowano w formie rozwiniętej, to znaczy schody przedsionka i schody bębenka (w rzeczywistości ponad 2,5 zwoju wokół osi ślimaka) narysowano jako proste. Nie zachowano również proporcji wymiarów. Wskazano natomiast, że drgania błony podstawnej zależne są od częstotliwości; obwiednia drgań błony, naszkicowana w uproszczeniu pod rysunkiem, charakteryzuje się występowaniem (przy odbieraniu czystego, pojedynczego tonu) maksimum, którego położenie zależne jest od częstotliwości tonu: im niższa częstotliwość, tym dalej (od podstawy ślimaka i okienka owalnego) występuje maksimum

Zjawisko to można ująć ilościowo. Fakt, że maksimum obwiedni drgań przypada przy różnych częstotliwościach w różnych punktach błony (rys. 3-10), pozwala „wyskalować" błonę w jednostkach częstotliwości. Pomiędzy współrzędną przestrzenną punktu na błonie $x$ a częstotliwością fali dźwiękowej $f(x)$, wprawiającej ten punkt w maksymalne drgania, występuje zależność wyrażająca się wzorem Greenwooda:

$$f(x) = b[10^{a(L-x)} - 1] \tag{3.11}$$

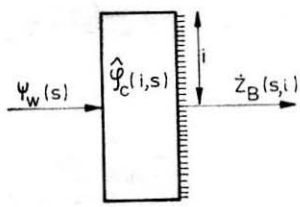
którego parametry, dla ucha człowieka, zwykło się przyjmować:

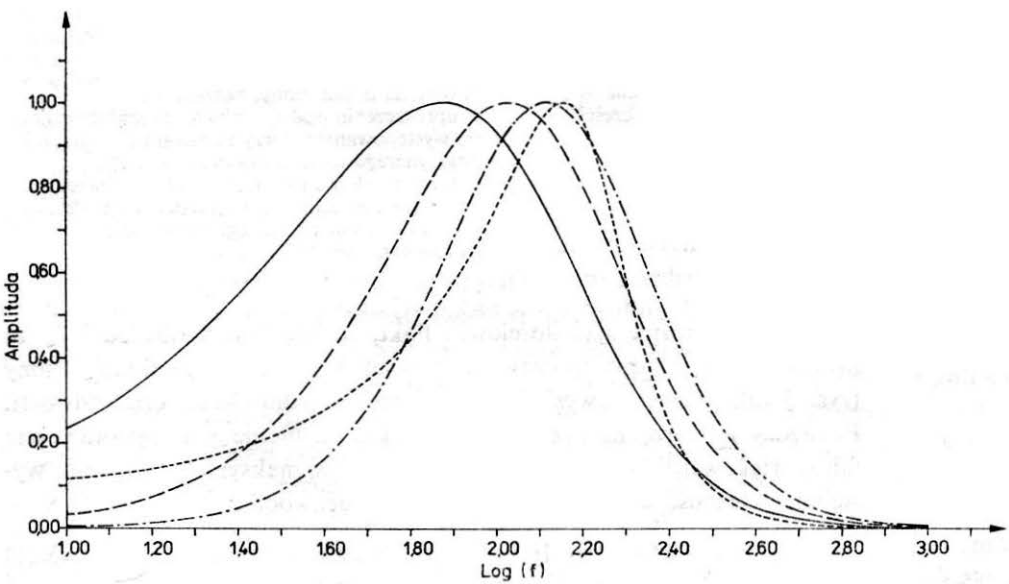
$$a = 0{,}06, \quad b = 165{,}4, \quad L = 35 \tag{3.12}$$




3-10. Obraz obwiedni drgań błony podstawnej przy trzech różnych częstotliwościach $f_1 > f_2 > f_3$. Oś amplitud $A$ wyskalowano w jednostkach względnych, rzeczywiste amplitudy drgań zależą bowiem od intensywności sygnału dźwiękowego i są na ogół bardzo małe. Przykładowo dla intensywności dźwięku odpowiadającego progowi słyszalności amplituda rejestrowanych przez system nerwowy drgań błony podstawnej jest rzędu rozmiarów atomów: $10^{-10}$ m. Oś długości liczonej wzdłuż kanału ślimaka $l$ jest natomiast wyskalowana w centymetrach; dla zachowania właściwej proporcji należało narysować obraz drgań tak, by amplituda drgań w punkcie maksimum była mniejsza od grubości kreski, którą zaznaczono oś $l$

3-11. Schemat modelu błony podstawnej; transmitancja jest w tym przypadku zależna nie tylko od operatora zespolonego $s$, ale również od numeru rozpatrywanego punktu na błonie $i$



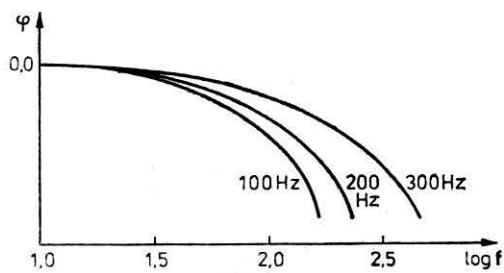

3-12. Wynik komputerowej symulacji błony podstawnej ucha wewnętrznego. Przebiegi charakterystyki amplitudowo-częstotliwościowej sygnału dla trzech wybranych częstotliwości sygnału sinusoidalnego (linia ciągła) oraz dla pojedynczej częstości fali trójkątnej (linia przerywana)

W odróżnieniu od poprzednio omówionych transmitancji, $\hat{\varphi}_c$ reprezentowane jest nie przez jedną, lecz przez zbiór transmitancji, odpowiadających wyróżnionym, kolejno ponumerowanym punktom błony (rys. 3-11). Wprowadzając zbiór $N^* = \{1, ..., 30\}$ i decydując, że kolejnym numerom $i = 1, ..., 30$ będą odpowiadać punkty odległe o $l_i = 35 - 16{,}6 \log(0{,}604i + 1)$ od helikotremy*), otrzymujemy zależność będącą podstawą symulacji

$$\hat{\varphi}_c(i, s) = \left[\frac{c_1 i}{c_2 + i}\right]^{0,8} \frac{(s + c_3 i)\exp\left(-\frac{c_4 s}{i}\right)}{(s + c_5 i)\left(\frac{c_6}{i^2}s^2 + \frac{c_7}{i}s + 1\right)^2} \tag{3.13}$$

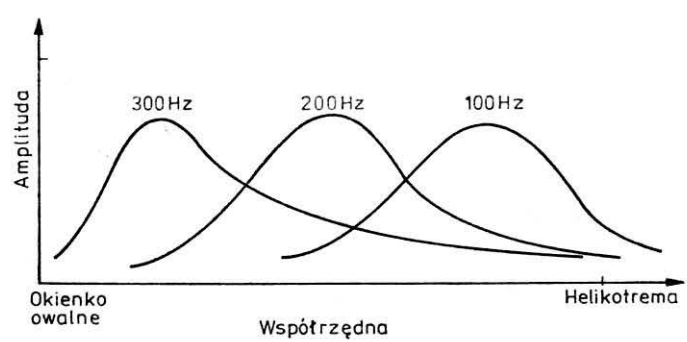
gdzie stałe wynoszą odpowiednio: $c_1 = 402$, $c_2 = 10$, $c_3 = 628$, $c_4 = 0{,}0038$, $c_5 = 6283$, $c_6 = 2 \cdot 10^{-6}$, $c_7 = 0{,}0013$.

Na rysunku 3-12 przedstawiono charakterystyki amplitudowo-częstotliwościowe, a na rys. 3-13 charakterystyki fazowo-częstotliwościowe dyskutowanego modelu błony dla punktów odpowiadających $i = 1$, 2 i 3. Ponieważ dla funkcjonowania modelu błony zasadnicze znaczenie ma lokalizacja punktu o maksymalnych drganiach, podano również charakterystyki amplituda — współrzędna (rys. 3-14) i faza — współrzędna (rys. 3-15) przy wymuszeniu tonem sinusoidalnym o częstotliwości odpowiednio równej 100, 200 i 300 Hz.




3-13. Charakterystyki fazowo-częstotliwościowe błony podstawnej dla trzech wybranych częstotliwości

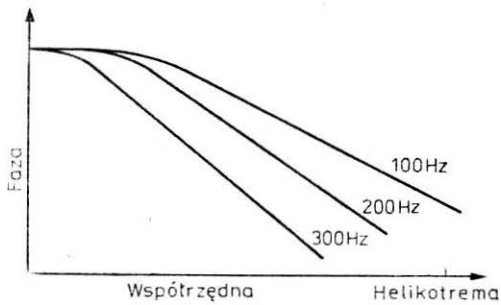
Odwzorowanie $\hat{\varphi}_r$ odpowiadające modelowaniu receptorów (komórek




3-14. Charakterystyki sygnału wyjściowego z błony podstawnej w układzie amplituda—współrzędna na błonie

*) Taki rozkład położeń branych pod uwagę punktów odpowiada równomiernemu rozkładowi częstotliwości charakterystycznych kolejnych punktów w przedziale 100÷300 Hz z krokiem 100 Hz.

rzęskowych narządu Cortiego) będzie opisane bez korzystania z przekształcenia Laplace'a, ponieważ w ich działaniu istotną rolę odgrywa czynnik nieliniowy (typu detekcji impulsowej), spowodowany faktem, że depolaryzacja błony komórki wywoływana jest wyłącznie przez jednokierunkowe uginanie rzęsek. Zagadnienie to zasługuje na obszerniejsze omówienie, gdyż zajmuje stosunkowo mało miejsca w łatwo dostępnej literaturze.

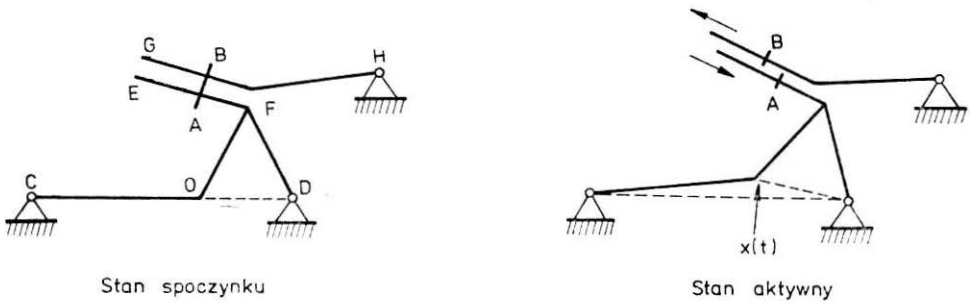



3-15. Charakterystyki sygnału wyjściowego z błony podstawnej w układzie faza — współrzędna na błonie dla różnych częstotliwości sygnału

## 3.2.5. Model receptora słuchowego

Bekesy wraz z grupą współpracowników dokonał w roku 1962 fundamentalnych badań częściowo wyjaśniających mechanizmy przetwarzania informacji w systemie słuchowym. Od tego też czasu wszystkie prace opierają się na podanych przez niego założeniach i tezach. W pracach tych często powraca się do problemu wyjaśnienia procesu przetwarzania sygnałów mechanicznych na sygnały nerwowe w narządzie Cortiego. Dotychczasowe badania opierały się na koncepcji powiązania potencjałów mikrofonowych z czynnościami elektrycznymi narządu Cortiego. Wydaje się, że takie podejście do rozwiązania problemu jest zbyt ogólne i nie w pełni obrazuje rzeczywisty proces przetwarzania informacji przez komórki rzęsate.

Modelowany biologiczny przetwornik sygnałów mechanicznych składa się z dwóch komórek: komórki rzęsatej oraz dwubiegunowej komórki nerwowej rozpatrywanych łącznie. W procesie modelowania zwrócono uwagę na środowisko, w jakim znajdują się te komórki, i odwzorowano wpływ układu




3-16. Geometryczna interpretacja pobudzenia komórek rzęsatych w narządzie Cortiego. Przemieszczenie błony podstawnej wywołane sygnałem dźwiękowym $X(t)$ powoduje w stereociliach komórek rzęsatych naprężenia zginające, wywołane przemieszczeniem punktów $A$ i $B$, między którymi utwierdzone są rzęski

mechanicznego błony podstawowej na przetwarzanie sygnałów i pobudzanie części receptorowych komórki rzęsatej.

Na podstawie rozważań nad budową układu: błona podstawna — narząd Cortiego — błona pokrywająca zaproponowano schemat układu mechanicznego (rys. 3-16). Działanie bodźca w formie wychylenia błony podstawnej CD w punkcie 0 powoduje przemieszczenie części sztywnych układu. Zmiana położenia narządu Cortiego EF względem błony pokrywającej GH stanowi główne źródło rozważanych dalej zjawisk. Należy bowiem zauważyć, że odpowiadające sobie punkty AB w wyniku przemieszczenia $x(t)$ w punkcie 0 ulegają przesunięciu względem siebie, a szerokość szczeliny pomiędzy błoną pokrywającą a kanałem Cortiego maleje. W rezultacie działanie układu mechanicznego można opisać równaniem:

$$P(t) = \begin{cases} (K - \Delta K) \cdot x(t) & \text{dla } x(t) \geqslant 0 \\ 0 & \text{dla } x(t) < 0 \end{cases} \tag{3.14}$$

gdzie:

$P(t)$ — ugięcie rzęsek komórek narządu Cortiego (odcinek AB na rys. 3-16),

$x(t)$ — funkcja drgań błony podstawnej,

$K$ — współczynnik wzmocnienia układu,

$\Delta K$ — adaptacyjna zmiana współczynnika wzmocnienia.

Uwzględnianie w równaniu (3.14) takich czynników jak: wzajemne i przeciwnie skierowane przemieszczanie się narządu Cortiego i błony pokrywającej względem siebie oraz nieliniowość przekształcenia funkcji drgań błony podstawnej $x(t)$ w funkcji zmiany przekroju szczeliny $P(t)$, przy pominięciu innych czynników upraszcza model, nie zubażając go nadmiernie.

Elementem bezpośrednio odbierającym sygnały $P(t)$ jest układ cienkich rzęsek (tzw. stereocilia), zakotwiczonych w płytce kutikularnej, oraz ciałko podstawowe (Hensena), występujące w formie zagęszczenia tworów cytoplazmatycznych bezpośrednio pod powierzchnią płytki kutikularnej (rys. 3-17). W zależności od typu komórki rzęsatej oraz jej lokalizacji rozróżniamy różne formy rozmieszczenia rzęsek ($I$, $V$, $U$, $W$), długości rzęsek zaś zmieniają się równomiernie malejąc w kierunku osi ślimaka (rys. 3-18).

Analizując układ rzęsek oraz ich właściwości receptorowe można zaproponować uproszczony wzór opisujący sumaryczną liczbę pobudzonych rzęsek:

$$n(x) = \int_{Y_0}^{Y} \varrho[y, x(t)]\,dy \tag{3.15}$$

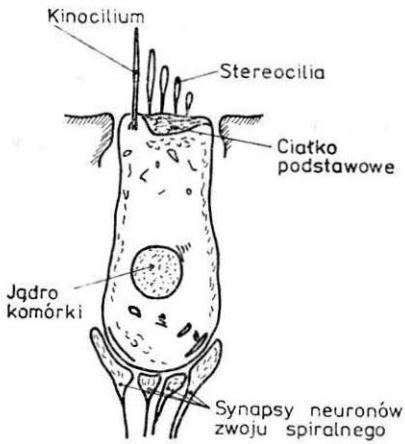
gdzie:

$y$ — odległość od osi ślimaka,

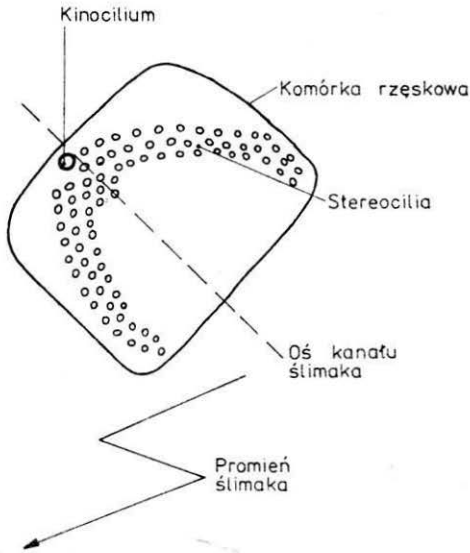
$x(t)$ — funkcja drgań błony podstawnej,

$\varrho$ — funkcja gęstości rozmieszczenia rzęsek.

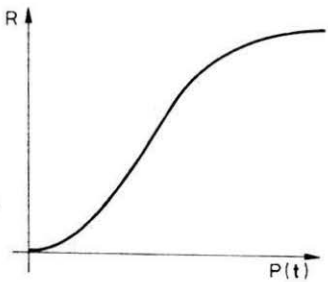
Dotychczasowe rozważania oparte były na ilościowym podejściu do problemu pobudzenia układu rzęsek. Należy wziąć też pod uwagę nieliniowość funkcji naprężenia rzęski w zależności od wielkości pobudzenia oraz nieliniowość występującą w transmisji sumarycznego naprężenia rzęsek do




3-17. Uproszczony schemat przekroju komórki słuchowej (rzęsatej) mieszczącej się w narządzie Cortiego ucha wewnętrznego człowieka. Widoczne u góry komórki rzęski to stereocilia odbierające wrażenia zmysłowe związane z drganiami akustycznymi. Informacje o tych drganiach przekazywane są do synaps komórek nerwowych zwoju spiralnego, widocznych u dołu komórki




3-18. Obraz rozkładu rzęsek na górnej powierzchni komórki rzęsatej, opracowany na podstawie fotografii z mikroskopu elektronowego. Dzięki regularnemu ułożeniu rzęsek liczba zadrażnionych (ugiętych) stereociliów jest nieliniową funkcją amplitudy drgań




3-19. Zmiana oporności elektrycznej $R$ ciałka Hensena w zależności od zmieniającej się w czasie szerokości przekroju szczeliny pomiędzy wierzchołkami komórek rzęsatych w narządzie Cortiego a błoną nakrywkową

ciałka Hensena. W wyniku tego ciałko Hensena na sygnał sumarycznego naprężenia reaguje zmianą swej odporności $\Delta R$, związaną z wartością $P(t)$ nieliniową zależnością

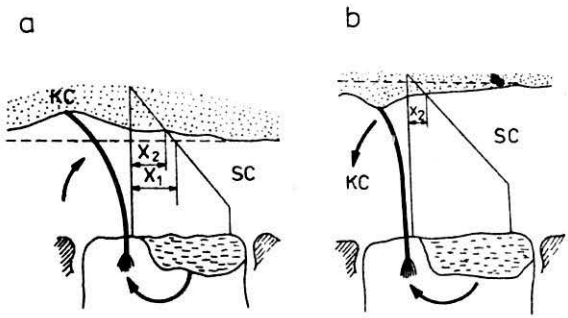
$$\Delta R = F[P(t)] \tag{3.16}$$

Orientacyjny przebieg funkcji $F[P(t)]$ przedstawiono na rys. 3-19. W trakcie badań symulacyjnych przyjmuje się różne parametry rozważanej funkcji i bada się ich wpływ na funkcjonowanie modelu.

Wśród rzęsek komórki rzęsatej wyróżniamy jedną grubszą witkę zwaną kinocilium o odrębnej strukturze i funkcji. Budowa jej jest podobna do budowy rzęsek tzw. brzeżka prążkowanego w przełyku i drogach oddechowych, które to rzęski mają zdolność ruchu. Kinocilium dla zewnętrznych

komórek rzęsatych jednym końcem przytwierdzone jest do błony pokrywającej, natomiast drugi koniec znajduje się wewnątrz komórki rzęsatej, stanowiąc zgrubienie zawierające błoniaste twory cytoplazmatyczne. Budowa ta sugeruje istnienie zamkniętego funkcjonalnie i metabolicznie systemu. Na podstawie podobieństwa struktury kinocilium oraz jego ciałka do innych podobnych rzęsek uznano ją za ośrodek ruchowy w narządzie Cortiego.

Bliskość położenia ciałka podstawowego kinocilium w stosunku do ciałka Hensena oraz specyficzne rozgraniczenie obu obszarów błonkami i tworami cytoplazmatycznymi pozwala sądzić, że istnieje tu sprzężenie zwrotne zawierające następujące układy: błona pokrywająca — stereocilia — ciałko Hensena — ciałko podstawowe — kinocilium — błona pokrywająca (rys.




3-20. Hipotetyczne funkcjonowanie wewnątrzkomórkowego sprzężenia zwrotnego, którego efektem jest regulacja czułości komórki rzęskowej w zależności od amplitudy rejestrowanego dźwięku. W przypadku dźwięku o dużej mocy (a) kinocilium KC odpycha błonę nakrywkową, w wyniku czego obszar ugięcia stereociliów SC zmniejsza się z wielkości $x_1$ do $x_2$. Natomiast w przypadku dźwięku o małej mocy (b) kinocilium przyciąga błonę nakrywkową, powiększając obszar ugięcia stereociliów do wartości $x_2$

3-20). Jego hipotetyczne działanie jest następujące. W wyniku pobudzenia następuje ruch błony pokrywającej oraz powierzchni narządu Cortiego względem siebie. Ruch ten powoduje odchylenie się pobudzonych stereociliów oraz kinocilium, umożliwiając wygięcie się tych rzęsek bez zmiany długości i innych deformacji. Gdy amplituda sygnału wejściowego jest duża, następuje pobudzenie określonej liczby stereociliów SC na długości $x_1$. Kinocilium KC, reagując na bodziec usztywnieniem, powoduje uwypuklenie błony pokrywającej, a tym samym pobudzenie stereociliów do odcinka długości $x_2$ i osłabienie sygnału (rys. 3-20a).

Natomiast dla zbyt małej amplitudy oddziaływanie kinocilium powoduje ruch w stronę przeciwną, uwypuklając błonę pokrywającą oraz zmieniając zakres pobudzenia stereociliów z $x_1$ na $x_2$ (większy). Istnienie obu rodzajów oddziaływań: sprzężenia zwrotnego ujemnego oraz dodatniego, ma wpływ na układ mechaniczny, a ściślej na zmianę współczynnika wzmocnienia $\Delta K(t)$ we wzorze (3.14) oraz jakość detekcji sygnałów.

W modelu dokładnym omówione pętle sprzężenia zwrotnego można opisać transmitancjami $G_1(s)$, $G_2(s)$:

$$z_1(s) = G_1(s) \cdot R(s) \tag{3.17}$$

$$z_2(s) = G_2(s) \cdot R(s) \tag{3.18}$$

gdzie: $R(s) = R_0 + \Delta R(s)$ (3.19)

$R_0$ — oporność ciałka Hensena dla $P(t) = 0$

$$G_1(s) = \frac{1 + T_3 \cdot s}{(1 + T_1 s) \cdot (1 + T_2 s)}$$

$$G_2(s) = \frac{1 + T_6 \cdot s}{(1 + T_4 s) \cdot (1 + T_5 s)}$$

$T_1 \div T_6$ — dobrane doświadczalnie stałe czasowe.

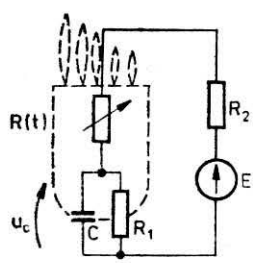
Natomiast adaptacyjną zmianę wzmocnienia układu mechanicznego $\Delta K(t)$ można wyrazić równaniem:

$$z_3(t) = e^{-h \cdot z_2(t)} \tag{3.20}$$

$$\Delta K(t) = c \cdot z_1(t) - d \cdot z_3(t) \tag{3.21}$$

gdzie: $c$, $d$, $h$ — stałe współczynniki.

Na obecnym etapie modelowania istotnym problemem jest transmisja sygnału, odwzorowanego przez zmienną rezystancję ciałka Hensena, a drugi koniec komórki rzęsatej, w rejon dochodzących zakończeń nerwowych. W tym celu rozważono prosty zastępczy obwód elektryczny (rys. 3-21), mo-




3-21. Schemat elektryczny układu, w którym zmiana parametru (rezystancji) ciałka Hensena $R$ jest zamieniana na zmianę napięcia drażniącego synapsę komórki nerwowej $u_c$. Źródło siły elektromotorycznej $E$ zasilane jest różnicą koncentracji jonów elektrolitów w poszczególnych częściach narządu Cortiego, linią przerywaną zaznaczono zarys komórki

gący odwzorowywać pracę komórki rzęsatej w jej środowisku. Zmienna rezystancja $R(t)$ ciałka Hensena powoduje zachwianie równowagi elektrycznej w narządzie Cortiego objawiające się zmianą napięcia na błonie komórkowej $u_c(t)$. W modelu uwzględniono takie wielkości jak: biologiczne źródło zasilania $E$, stałe rezystancje $R_1$, $R_2$ reprezentujące oporności środowiska i błony komórkowej, zmienną rezystancję $R(t)$ ciałka Hensena oraz pojemność $C$ błony komórkowej. Wielkość $u_c(t)$, występującą na końcu komórki rzęsatej, uzyskuje się rozwiązując równanie różniczkowe:

$$\frac{du_c(t)}{dt} + u_0(t) \cdot C \left[ \frac{1}{R_1} + \frac{1}{R_2 + R(t)} \right] - E \frac{C}{R_2 + R(t)} = 0 \tag{3.22}$$

gdzie:

$C$ — pojemność błony komórkowej,

$R = R_0 + \Delta R$ — oporność ciałka Hensena ($R_0$ — oporność spoczynkowa),

$\Delta R$ — zmiana oporności pod wpływem sygnału pobudzającego,

$R_1$ — oporność błony komórkowej,

$R_2$ — oporność środowiskowa,

$E$ — różnica potencjałów między wnętrzem komórki rzęsatej a otoczeniem ($E = 160$ mV = const).

W dolnej części komórki rzęsatej występuje duże zagęszczenie tworów cytoplazmatycznych, zwane ciałkiem Retziusa, wśród których główne znaczenie mają pęcherzyki wypełnione neuromediatorem. W tej części komórki do błony komórkowej przylegają zakończenia nerwowe układu aferentnego i eferentnego. Istnieją trzy rodzaje zakończeń nerwowych różniące się między sobą subtelnymi, lecz istotnymi szczegółami, takimi jak: rodzaj styku z błoną komórkową oraz struktury cytoplazmatyczne, będące w pobliżu styku. W układzie tym każdy nerw aferentny odbiera sygnały od grupy komórek sensorowych i kieruje je w stronę wyższych pięter systemu nerwowego. Natomiast struktura oraz znaczenie równie gęstej sieci włókien eferentnych nie jest obecnie dokładnie znane. Przypuszcza się, że sieć ta stanowi istotny czynnik w procesie kodowania oraz wstępnego przetwarzania (selekcji) informacji w różnych częściach ślimaka.

W modelu omawianego systemu uwzględniono styk komórki rzęsatej z aferentnym zakończeniem nerwowym. Założono przy tym, że ciałko Retziusa w komórce spełnia rolę kolbki presynaptycznej synapsy pobudzającej. W przypadku wystąpienia sygnału pobudzającego następuje wydzielenie neuromediatora z pęcherzyków, a tym samym zadrażnienie błony postsynaptycznej. Dynamikę takiego styku zamodelowano za pomocą transmitancji $G_3(s)$

$$G_3(s) = \frac{w(1+sT_9)e^{-s\tau}}{(1+sT_7)(1+sT_8)} \tag{3.23}$$

gdzie:

$w$ — waga danego styku,

$T_7, T_8, T_9$ — stałe czasowe (dobierane eksperymentalnie),

$\tau$ — opóźnienie.

Pełny model komórki rzęskowej, obejmujący zjawiska opisane równaniami (3.14)÷(3.23), jest jednak zbyt skomplikowany do symulowania go wraz z innymi elementami systemu słuchowego i dlatego zostanie znacznie uproszczony. Można więc zapisać dla $\hat{\varphi}_r$:

$$X_r(n, t) = \frac{1}{2T} \int_{t-T}^{t} y_w(l, t)[1 + \operatorname{sign} y_w(l, t)]\,dt + \mathrm{RND} \tag{3.24}$$

gdzie RND jest wartością przypadkową symbolizującą czynnik losowości, występujący w funkcjonowaniu receptorów. Wyraża się on impulsacją spontaniczną w warunkach braku sygnału dźwiękowego i indeterministyczną

relacją pomiędzy wielkością odpowiedzi receptora a wielkością bodźca. W dalszej dyskusji czynnik losowy pominięto (RND = 0), gdyż wyniki wstępnych badań wykazały, że jego uwzględnianie prowadzi do trudności z interpretacją uzyskiwanych wyników. Ponadto uwzględniono odwzorowanie wyłącznie dla wybranych wartości parametru $l$, odpowiadających wspomnianym wyżej dyskretnym punktom, w których modelowano funkcjonowanie błony. Zatem ostatecznie odwzorowanie $\hat{\varphi}_r$ odpowiada przekształceniu:

$$X_r(i, \eta) = \frac{1}{k} \sum_{\nu=\eta-k}^{\nu} y_w(i, \nu)[1 + \operatorname{sign}(y_w(i, \nu))] \tag{3.25}$$

Odwzorowanie $\hat{\varphi}_g$ ma złożony charakter, gdyż w zależności od typu połączeń między dendrytem neuronu zwoju spiralnego a komórkami receptorowymi pobudzenie uzależnione jest od stanu błony w ustalonym punkcie lub na pewnym, niekiedy dość rozległym obszarze. Zagadnienie to zasługuje na obszerniejsze przedyskutowanie, czemu poświęcony będzie kolejny podrozdział. Zanim jednak przejdziemy do dyskusji szczegółów, warto wskazać, dlaczego poświęcamy tej sprawie tak wiele uwagi.

Ucho wewnętrzne pełni w systemie słuchowym człowieka dwojaką funkcję. Z jednej strony jest ono analizatorem widma (składu harmonicznego) odbieranych sygnałów dźwiękowych, z drugiej stanowi przetwornik, w którym parametry odbieranego sygnału dźwiękowego są przekodowane i przekształcone na impulsy nerwowe, przekazywane do nerwowej części systemu słuchowego. Klasycznie przyjmowano, że obie wymienione funkcje ucha wewnętrznego są od siebie niezależne, gdyż pierwszą z nich wiązano z mechanicznymi własnościami błony podstawowej ślimaka, drugą natomiast z funkcjonowaniem komórek rzęskowych narządu Cortiego i pracą dwubiegunowych neuronów zwoju spiralnego. Wiele faktów zmusza jednak do zrewidowania tego poglądu. Z dokładnych obserwacji fizjologicznych, a także z obliczeń i prób modelowania wynika, że błona podstawowa ucha wewnętrznego jest analizatorem dźwiękowym o bardzo małej dobroci*). W zakresie częstotliwości $1 \div 3$ kHz ocenia się, że $Q_{błony} \approx 1$.

Tymczasem badania mikroelektrodowe prowadzone w kanale słuchowym wewnętrznym dowodzą, że rozkład pobudzeń komórek zwoju spiralnego charakteryzuje się wyraźnie większą dobrocią, rzędu $Q_{zwoju} \approx 20$. Jest to w dalszym ciągu znacznie mniej, niż wynosi dobroć całego systemu słuchowego, dla którego przyjmuje się $Q_{systemu} \approx 200$, jednakże dalszy (w stosunku do $Q_{zwoju}$) wzrost dobroci częstotliwościowej systemu słuchowego można z powodzeniem wytłumaczyć pracą wielowarstwowych asymetrycznych sieci z hamowaniami obocznymi, a nawet udało się określić strukturę takich sieci. Znacznie trudniejsze jest wyjaśnianie efektu wzrostu selektyw-

*) Dobroć analizatora akustycznego rozumiana jest tutaj jako stosunek częstotliwości sygnału wymuszającego $f_0$, będącego czystym tonem sinusoidalnym, do szerokości pasma $\Delta f$, w którym sygnał wyjściowy analizatora jest mniejszy od sygnału dla częstotliwości $f_0$ co najwyżej o 3 dB: $Q = \frac{f_0}{\Delta f}$.

ności na styku między błoną podstawową a zwojem spiralnym (przejście od $Q_{błony}$ do $Q_{zwoju}$). Fenomenu tego nie można wyjaśniać oddziaływaniami międzyneuronowymi, gdyż neurony zwoju spiralnego mają formę dwubiegunową i nie kontaktują się między sobą.

Wyjaśnienie przyczyn wzrostu dobroci analizatora słuchowego ma znaczenie nie tylko teoretyczne. Wszystkie budowane do chwili obecnej analizatory widma dowolnych sygnałów, niezależnie od techniki, w jakiej je wykonano, wykazywały charakterystyczne powiązanie dobroci $Q$ z minimalnym niezbędnym czasem analizy $\Delta t$. Związek ten można wyrazić wzorem

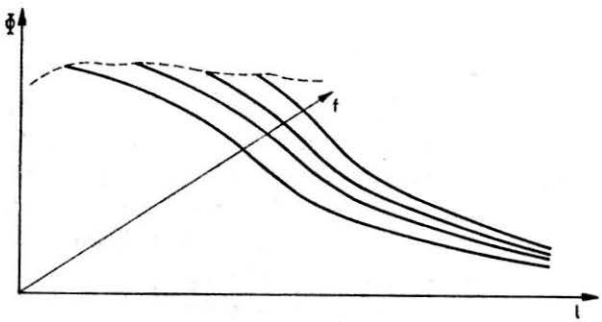
$$\Delta f \cdot \Delta t = k = \text{const} \tag{3.26}$$

gdzie $k$ jest zależne od zastosowanej techniki (zwykle $k \approx 1$). Badania ucha sugerują, że w systemie słuchowym ograniczenie (3.26) wydaje się nie obowiązywać. System ten, jak wspomniano wyżej, cechuje się dużą dobrocią ($Q_{systemu} \approx 200$) przy równoczesnym bardzo małym czasie analizy ($\Delta t \approx$ $\approx$ 10 ms). Żaden ze znanych systemów analizy, włączając w to zastosowanie algorytmu szybkiej transformaty Fouriera, nie zapewnia takiego tempa analizy przy wskazanej selektywności.

## 3.2.6. Model przekazywania informacji do części nerwowej systemu słuchowego

Jak wynika z przytoczonej wyżej dyskusji, na błonie podstawnej tworzona jest „mapa" pobudzeń i fale o różnych długościach wprawiają w maksymalne drgania różne jej rejony. Rozróżnienie dźwięków o różnych częstotliwościach możliwe jest także i na innej zasadzie. Obok charakterystyk amplitudowych, przytoczonych przykładowo na rys. 3-12, rozważać można charakterystyki fazowe błony podstawnej, przedstawione na rys. 3-22.

Z analizy charakterystyk fazowych $\Phi(l, f)$ (por. także rys. 3-13) można wywnioskować, że proces dynamiczny, zachodzący na błonie podstawnej pod

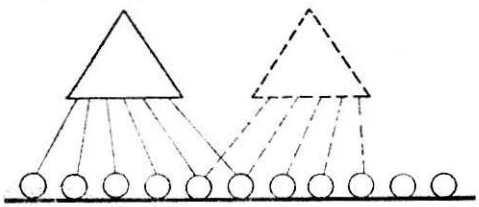



3-22. Charakterystyki błony podstawnej podane w trójwymiarowym układzie współrzędnych: faza, częstotliwość, odległość. Z charakterystyk podanych tu i prezentowanych w innych pracach można wyciągnąć wniosek, że szybkość rozchodzenia się fali pobudzenia akustycznego w błonie podstawnej zależy zarówno od odległości rozważanego punktu od helikotremy, jak i od częstotliwości. Innymi słowy w ustalonym punkcie błony, odpowiadającym położeniu rozważanej komórki rzęsatej, szybkość propagacji fali będzie się zmieniała wraz ze zmianami częstotliwości. Fakt ten można wykorzystać przy próbie wyjaśnienia fonomenu selektywności ucha

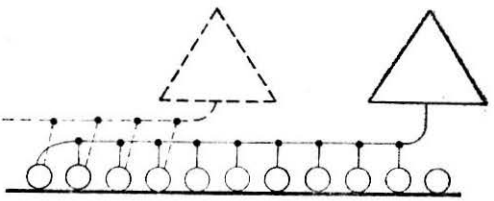
wpływem sygnału dźwiękowego, ma charakter fali mechanicznej, biegnącej wzdłuż ślimaka i gasnącej u jego szczytu. Fala ta rozprzestrzenia się z prędkością, którą można wyznaczyć z wzoru

$$v(l,f) = \frac{2\pi f}{\frac{\partial \Phi(l,f)}{\partial x}} \tag{3.27}$$

W przekazywaniu informacji akustycznej do systemu nerwowego uczestniczą komórki rzęskowe podlegające naprężeniom mechanicznym podczas drgań błony podstawowej. Naprężenia te zamieniane są na impulsy nerwowe, przekazywane do neuronów zwoju spiralnego, a ze zwoju spiralnego nerwem słuchowym — do pnia mózgu. Proces ten zachodzi w sposób niejednorodny. Można wyróżnić dwa typy neuronów zwoju spiralnego. Pierwsze, nazywane ortoneuronami, zbierają pobudzenia z kilku zaledwie blisko sie-



3-23. Połączenie pomiędzy ortoneuronami a pobudzającymi je komórkami rzęsatymi. Widać, że komórka nerwowa (trójkątna) zbiera pobudzenia z małego obszaru błony podstawnej, obsadzonego przez ograniczoną grupę komórek rzęsatych (kółka)



3-24. Schemat połączeń komórek rzęsatych (kółka) ze spironeuronami zwoju spiralnego (trójkąty). Widoczny długi dendryt spironeuronu rozciąga się wzdłuż błony podstawnej i rejestruje pobudzenie łączne, pochodzące od bardzo wielu (typowo — kilkuset) receptorów. Obszar innego dendrytu sąsiedniego spironeuronu (linia przerywana) częściowo zachodzi na obszar pobudzenia prezentowanego spironeuronu, w wyniku czego do każdej komórki rzęsatej dociera kilka dendrytów. Jest ona wobec tego składnikiem wielu podobnych zespołów

bie zlokalizowanych komórek rzęskowych (rys. 3-23), drugie natomiast, zwane spironeuronami, mają długi dendryt, przebiegający wzdłuż błony podstawowej i kontaktujący się z wieloma komórkami rzęskowymi (patrz rys. 3-24). Dendryt spironeuronu na swojej długości kontaktuje się z komórkami rzęskowymi, do których fala mechaniczna na błonie podstawnej dociera w kolejnych chwilach czasu. Odstępy czasowe między kolejnymi pobudzeniami są zależne od momentów, w których wierzchołek fali dociera do odpowiednich punktów, a te zależą od prędkości biegnącej fali. Ponieważ prędkość fali jest uzależniona od częstotliwości, wobec tego rytm pobudzeń zbieranych przez dendryt spironeuronu zależy od częstotliwości analizowanego dźwięku. Równocześnie jednak każdy punkt dendrytu spironeuronu, do którego dotarło pobudzenie od komórki rzęskowej, staje się źródłem sygnału (fali depolaryzacji, rozchodzącej się elektrotonicznie, lub impulsu czynnościowego, rozchodzącego się na drodze aktywnych procesów w błonie komórkowej), którego propagacja w kierunku ciała (perikarionu) komórki

zwoju spiralnego odbywa się z określoną prędkością (zależną od średnicy dendrytu). Zachodzi więc równoległy bieg dwu fal: mechanicznej w błonie podstawnej i elektrycznej (nerwowej) w dendrycie spironeuronu. Istnieje przypuszczenie, że w przypadku synchronizacji tych fal pobudzenie spironeuronu powinno być maksymalne. Jeśli założyć, że wskazana synchronizacja zachodzi w punkcie błony o maksymalnej amplitudzie drgań, to wówczas efekty synfazowego sumowania bodźców od kolejnych komórek rzęskowych przez dendryt spironeuronu mogą dawać zwiększenie dobroci systemu słuchowego, w stosunku do wartości wynikających z rezonansowych charakterystyk błony podstawnej.

Aby dokładniej przeanalizować zjawisko, można odwołać się do modelu symulacyjnego błony podstawnej (odwzorowanie $\hat{\varphi}_c$) oraz zamodelować prawdopodobny przebieg procesów nerwowych. Można także wyprowadzić funkcję $v(x, f)$, której uproszczona postać jest następująca:

$$v(l,f) = a_0 \cdot \frac{b_0 f^{11} + b_1 f^9 10^{-2l} + b_2 f^7 10^{-4l} + b_3 f^5 10^{-6l} + b_4 f^3 10^{-8l} + b_5 f 10^{-10l}}{c b_0 f^{11} 10^{l} + g_1 f^9 10^{-l} + g_2 f^7 10^{-3l} + g_3 f^5 10^{-5l} + g_4 f^3 10^{-7l} + g_5 f 10^{-9l}} \tag{3.28}$$

gdzie:

$$\begin{aligned} &a_0 = \exp(10),\ a_1 = 2\pi 10^5,\ a_2 = -2{,}5(2\pi)^7 a_1^3, \\ &a_3 = 9{,}875 \cdot (2\pi)^5 a_1^5,\ a_4 = -12{,}5(2\pi)^3 a_1^7,\ a_5 = -1{,}125\pi a_1^9, \\ &b_0 = (2\pi)^{11},\ b_1 = -2{,}5(2\pi)^9 \cdot a_1^2,\ b_2 = 10{,}125(2\pi)^7 a_1^4,\ b_3 = \\ &= 2{,}9375(2\pi)^5 a_1^6,\ b_4 = 0{,}984(2\pi)^3 a_1^8,\ b_5 = 0{,}5\pi a_1^{10}, \\ &c = 0{,}375 \cdot 10^{-5},\ g_1 = a_1 + b_1 c,\ g_2 = a_2 + b_2 c, \\ &g_3 = a_3 + b_3 c,\ g_4 = a_4 + b_4 c,\ g_5 = a_5 + b_5 c \end{aligned} \tag{3.29}$$

Trudniej natomiast opisać zjawiska zachodzące w systemie nerwowym.

W przypadku przyjęcia hipotezy elektrotonicznego*) rozchodzenia się pobudzenia w dendrycie spironeuronu, sygnały od poszczególnych synaps, przez które spironeuron styka się z kolejnymi komórkami rzęskowymi, ulegają przed zsumowaniem w perikarionie komórki opóźnieniom proporcjonalnym do odległości synapsy oraz tłumieniu, także zależnemu do odległości. Oznaczając przez $\tau(l - l_0)$ opóźnienie wprowadzane przez odcinek dendrytu między synapsą w punkcie $l$ a perikarionem spironeuronu odpowiadającym punktowi $l_0$, a przez $G_d\ (s, l - l_0)$ transmitancję wyrażającą między innymi tłumienie sygnału przy jego przesyłaniu przez dendryt, możemy zapisać globalne pobudzenie spironeuronu $e(l_0, s)$ jako:

$$e(l_0, s) = \sum_{l=1}^{n} G_d(s, l_i - l_0) e^{-\tau(l_i - l_0)s} x_r(l_i, s) \tag{3.30}$$

gdzie przez $l_i$, $i = 1, 2, \ldots, n$ oznaczono położenie kolejnych komórek rzęs-

*) Alternatywą tej hipotezy jest założenie, że w dendrycie dochodzi do generacji impulsu czynnościowego, który jest przesyłany na drodze aktywnych procesów w błonie komórkowej.

kowych na błonie podstawnej. Sygnał wyjściowy spironeuronu $x_g(i,\eta)$ wyraża się nieliniową funkcją pobudzenia

$$e(l_0,\eta) = \mathscr{L}^{-1}[e(l_0,s_0)]$$

$$x_g(i,\eta) = \begin{cases} 0 & \text{jeśli } e(l_0,\eta) < \Theta(\eta) \\ w\cdot[e(l_0,\eta)-\Theta(\eta)] & \text{jeśli } \Theta(\eta) \leqslant e(l_0,\eta) \leqslant x_{max} \\ w\cdot[x_{max}-\Theta(\eta)] & \text{jeśli } e(l_0,\eta) > x_{max} \end{cases} \tag{3.31}$$

gdzie $w$ i $x_{max}$ są stałymi, zaś $\Theta(\eta)$ oznacza zmienny przebieg progu, uwzględniający zjawiska refrakcji bezwzględnej i względnej

$$\Theta(\eta) = \begin{cases} +\infty & \text{dla } \eta_0 \leqslant \eta \leqslant \eta_0+\varepsilon \\ \Theta_0 + \dfrac{\mathscr{L}}{\eta-(\eta_0+\varepsilon)} & \text{dla } \eta \geqslant \eta_0+\varepsilon \end{cases} \tag{3.32}$$

gdzie $\eta_0$ oznacza moment generacji impulsu czynnościowego, a pozostałe parametry są stałymi, których wartości można obliczyć na podstawie znanych rezultatów eksperymentów neurofizjologicznych. Wartość $x_g(i,\eta)$ nazywana dalej sygnałem wyjściowym spironeuronu, interpretowana jest jako chwilowa częstotliwość impulsów czynnościowych o standardowej postaci

$$h(\eta) = ke^{-a\eta} - ke^{-(a+b)\eta} - ne^{c\eta} + ne^{-d\eta} \tag{3.33}$$

Funkcje $G_d$ oraz $\tau$ występujące we wzorze (3.30) można wyznaczyć w postaci

$$G_d(s,\varrho_i) = \frac{m(\varrho_i)(1+T_3 s)}{s(T_2 s+1)(T_1 s+1)} \tag{3.34}$$

$$\tau(\varrho_i) = \frac{\varrho i}{v_1} \tag{3.35}$$

gdzie:

$$m(\varrho_i) = m_0 \frac{1}{1+\varrho_i} \tag{3.36}$$

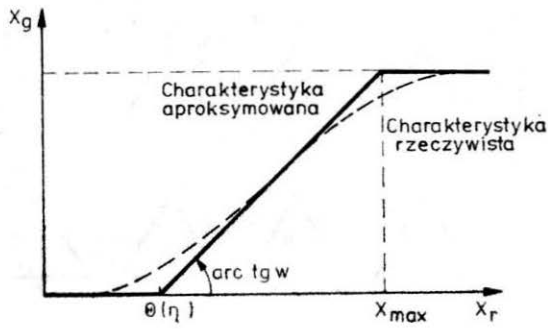
$$\varrho_i = |l_i - l_0| \tag{3.37}$$

Nie jest jeszcze wiadome, czy stałe czasowe $T_1$, $T_2$, $T_3$ są zależne od odległości $\varrho_i$. Ponieważ brak jest przesłanek do określenia postaci zależności $T_j(\varrho_i)$, to do modelowania przyjęto $T_j = \text{const}$. Nie jest również stwierdzone, czy powinno się przyjmować $v_1 = v_1(l)$, wobec czego założono $v_1 = \text{const}$. Przyjęty model umożliwia występowanie wskazanej wyżej interferencji fali pobudzenia i fali mechanicznej w błonie dla spironeuronów. Dla ortoneuronów natomiast przyjęto

$$x_g(i,\eta) = \begin{cases} 0 & \text{gdy } x_r(i,\eta) < \Theta(\eta) \\ w[x_r(i,\eta)-\Theta(\eta)] & \text{gdy } X_{max} > x_r(i,\eta) \geqslant \Theta(\eta) \\ w[X_{max}-\Theta(\eta)] & \text{gdy } x_r(i,\eta) > X_{max} \end{cases} \tag{3.38}$$

gdzie $w$, $X_{max}$ oraz $\Theta(\eta)$ są parametrami nieliniowej charakterystyki modelu neuronu (rys. 3-25). Pominięto przy tym, jak widać ze wzoru (3.38), wszystkie elementy dynamiczne, które mogłyby wystąpić w modelu komórki nerwo-

wej. Założono, że w porównaniu z bezwładnością mechanicznych elementów systemu słuchowego opóźnienia występujące przy propagacji sygnału w dendrytach ortoneuronów czy procesy dynamiczne w ich synapsach nie mają istotnego wpływu na funkcjonowanie modelu.




3-25. Charakterystyka statyczna najprostszego modelu neuronu. Sygnał wyjściowy $X_g$ tylko w pewnym przedziale jest monotonicznie zależny od sygnału wejściowego $X_r$, gdyż poniżej progu zadziałania $\Theta$ oraz powyżej pobudzenia maksymalnego $X_{max}$ sygnał wejściowy praktycznie nie wpływa na sygnał wyjściowy

Ostatni element modelu, sieć warstwowa reprezentująca funkcjonowanie jąder ślimakowych, można opisać funkcją:

$$x_k(i,\eta) = f_2\{by(i-1,\eta) - y(i,\eta)\} =$$
$$= f_2\{bf_1[ax_g(i-1,\eta) - x_g(i,\eta)] - f_1[ax_g(i,\eta) - x_g(i+1,\eta)]\} \quad (3.39)$$

gdzie:

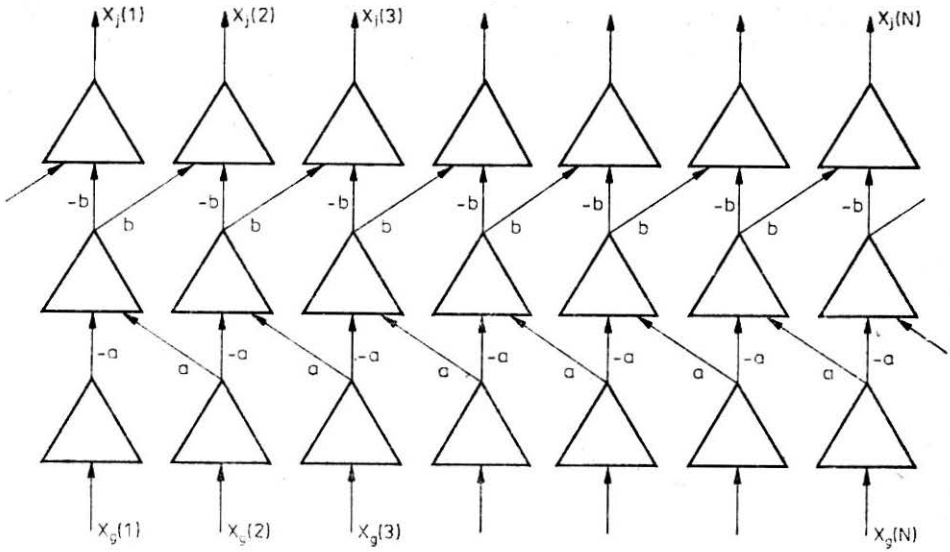
$$f_1(e) = \begin{cases} 0 & e < -p \\ \dfrac{e+p}{2p}k & -p \leqslant e \leqslant p \\ k & e > p \end{cases}$$

$$f_2(e) = \begin{cases} 0 & e < \dfrac{bk}{3} \\ \dfrac{3e-bk}{2bk} & \dfrac{bk}{3} \leqslant e \leqslant bk \\ 1 & e > bk \end{cases}$$

$p$, $b$ oraz $k$ — parametry przyjętych modeli komórki nerwowej odpowiednio realizujących funkcje $f_1$ i $f_2$.

Wartości parametrów $p$, $b$, $k$ dobiera się opierając się na dodatkowych kryteriach (m.in. zdolność eliminacji zakłóceń) na drodze obliczeniowej lub (częściej) empirycznie. Funkcja ta wynika ze specjalnie dobranej struktury sieci neuronowej modelującej funkcje jąder ślimakowych (rys. 3-26). Sieć ta została dobrana w wyniku prac zmierzających do ustalenia takiej jej struktury, aby dobroć na wyjściu była większa niż na wejściu. Sygnał wyjściowy z ostatniej warstwy sieci modelującej jądra ślimakowe jest sygnałem wyjściowym modelu.

Poszczególne elementy modelu, a także całą strukturę, badano przy użyciu generatora sygnałów testowych (por. rys. 3-5 — linia przerywana), generującego przebiegi czasowe




3-26. Najprostsza struktura sieci neuropodobnej polepszającej selektywność rozdziału sygnałów dźwiękowych o różnych częstotliwościach w systemie nerwowym obsługującym funkcjonowanie analizatora słuchowego. Sieci podobnego typu i realizujących zbliżone funkcje można opracować i zaproponować bardzo wiele, jednak przedstawiony model odznacza się prostą i regularną, warstwową budową, co może ułatwiać jego praktyczne wykorzystanie. Nie wiadomo, czy i w jakim stopniu model ten odpowiada rzeczywistej strukturze jąder ślimakowych człowieka

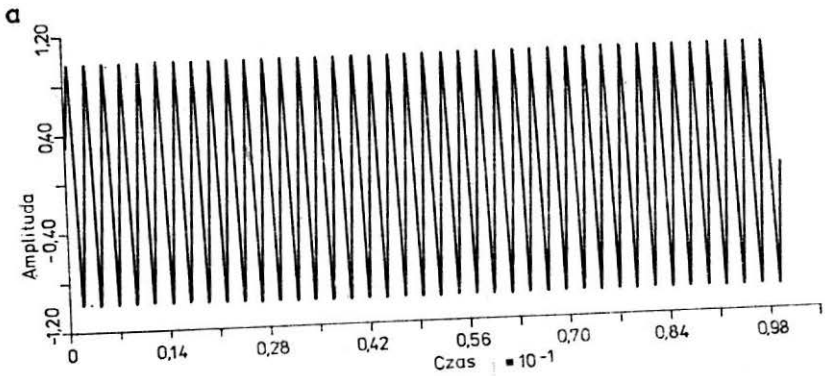
$$p_1(t) = A\sin(2\pi ft+\varphi) \tag{3.40}$$

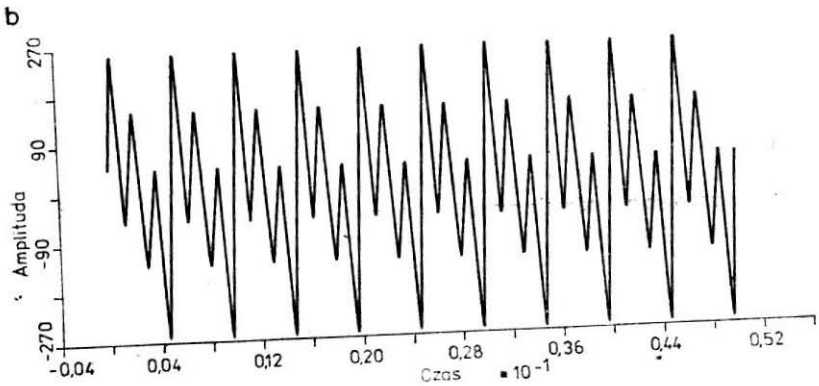
$$p_2(t) = \sum_{i=1}^{n} A_i \sin(2\pi i f_0 t+\psi_i) \tag{3.41}$$

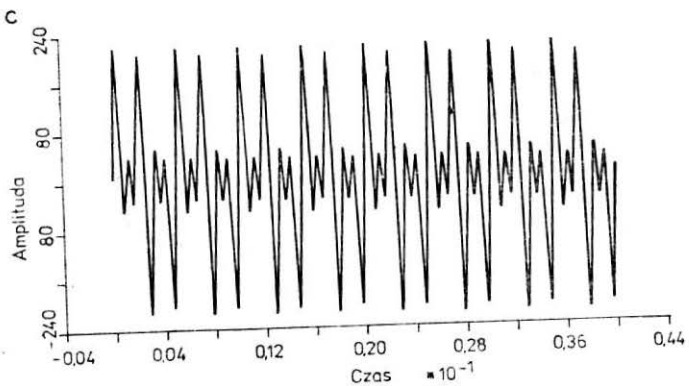
oraz

$$p_3(t) = \sum_{i=1}^{n} A_i \sin(2\pi f_i t+\psi_i) \tag{3.42}$$

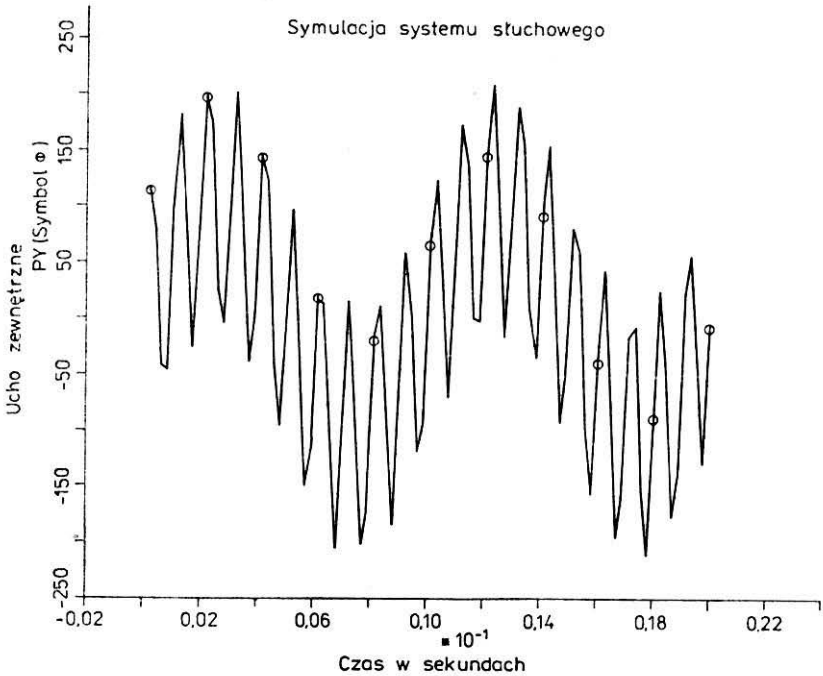
o dowolnie ustalanych wszystkich parametrach. Przebiegi sygnałów testowych przedstawiono na rys. 3-27. Odpowiedź układu modelującego ucho zewnętrzne na wymuszenie postaci jak na rys. 3-27a przedstawiono na rys. 3-28, a odpowiedź modelu ucha środkowego na ten sam sygnał przedstawiono na rys. 3-29. Na rys. 3-30 przedstawiono obraz czasowo-przestrzennych zjawisk modelowanych dla ucha wewnętrznego. Sygnałem wejściowym był sygnał z rys. 3-27c o częstotliwości $f = 1$ kHz. Na rysunku, którego osie odpowiadają czasowi oraz współrzędnej przestrzennej (numerowi rozpatrywanego punktu na błonie), widać powstawanie i propagację fal mechanicznych wzdłuż błony podstawnej, a także modulację amplitudy fali w zależności od położenia rozważanego punktu (tzw. zasada miejsca). To ostatnie zjawisko dokładniej można prześledzić na rys. 3-31, przedstawiającym obwiednię drgań błony oraz obraz biegnącej fali w trzech wybranych momentach czasu. Na rysunku 3-32 przedstawiono w analogiczny sposób jak na rys. 3-30 rozkład pobudzeń na wyjściu jąder ślimakowych.



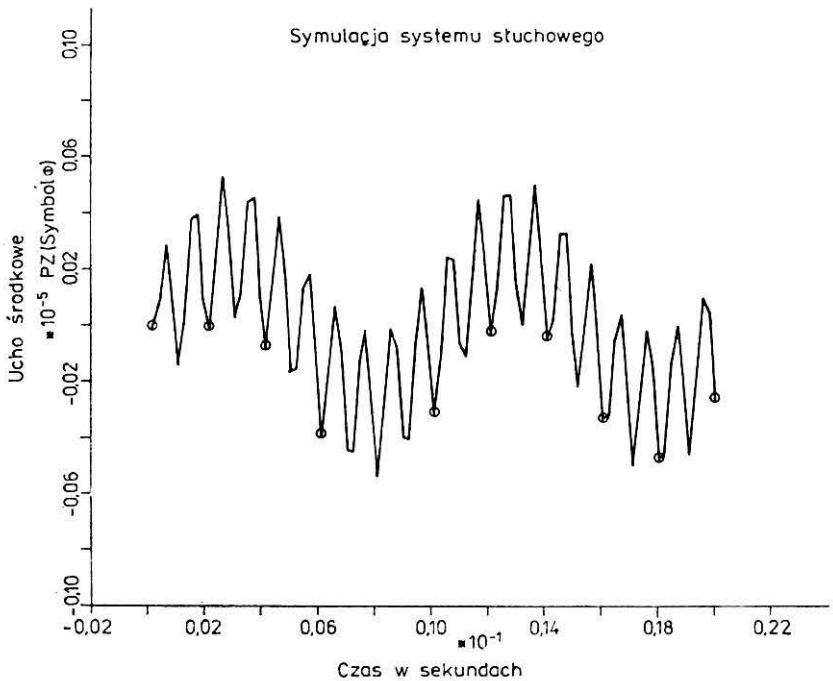




**3-27. Sygnały testowe:**
a — sygnał używany przy badaniu symulowanych komputerowo elementów modelu systemu słuchowego człowieka (przedstawiony przebieg odpowiada czystemu tonowi), b — sygnał używany w modelu systemu słuchowego — akord harmoniczny, zawierający składowe o częstotliwościach będących wielokrotnościami częstości podstawowej, c — sygnał używany w modelu systemu słuchowego; reprezentuje najbardziej złożoną postać sygnału: kombinację przebiegów o częstotliwościach nie będących wielokrotnościami, zatem cały sygnał nie odznacza się regularnością, właściwą tonom harmonicznym

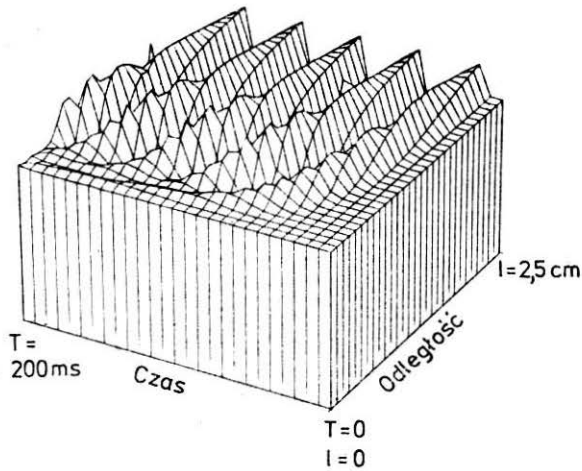


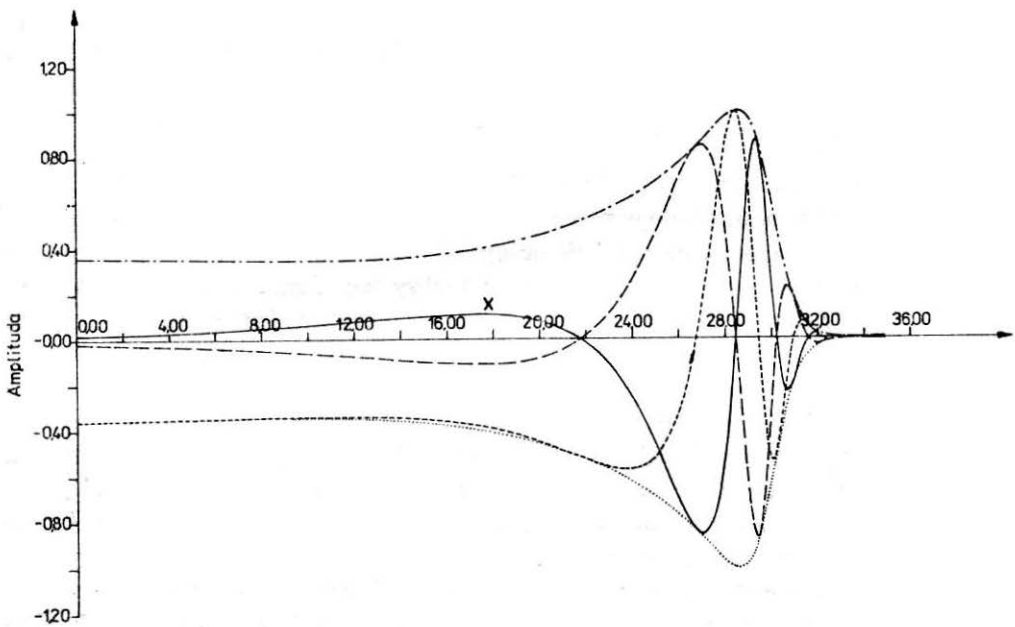

3-28. Odpowiedź ucha zewnętrznego na pobudzenie sygnałem testowym




3-29. Odpowiedź ucha środkowego na pobudzenie sygnałem testowym

3-30. Odpowiedź ucha wewnętrznego na pobudzenie sygnałem testowym. W odróżnieniu od rys. 3-27 i 3-29, prezentujących przebiegi odpowiednich sygnałów w funkcji czasu, na podanym rysunku przedstawiono przebieg sygnału w funkcji czasu oraz w funkcji odległości od helikotremy, gdyż proces zachodzący na błonie podstawnej należy rozważać w kategoriach czasoprzestrzennych

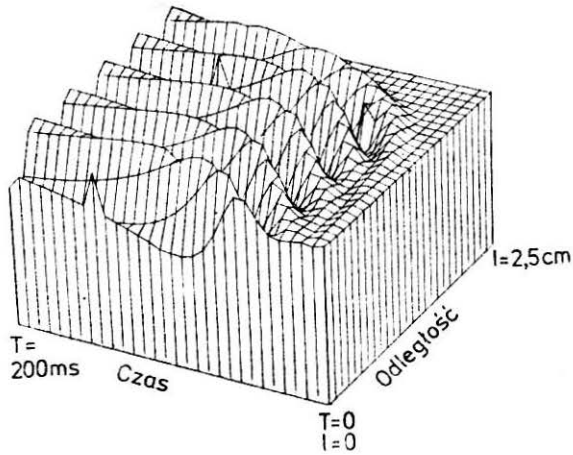



3-31. Obraz ob wiedni drgań błony podstawnej ucha wewnętrznego wraz z obrazem biegnącej fali, w trzech równo odległych momentach czasu, uzyskanej w wyniku symulacji

## 3.2.7. Uwagi końcowe

Skonstruowany model systemu słuchowego ma na celu analizę procesów dynamicznych zachodzących w poszczególnych piętrach systemu słuchowego w czasie percepcji wrażeń dźwiękowych przez człowieka. Dzięki zbudowaniu modelu możliwe stało się zbadanie szeregu hipotetycznych zjawisk zachodzących podczas analizy fali dźwiękowej, a także określenie tych własności sygnału dźwiękowego, które w wyniku procesu redukcji informacji zawar-




3-32. Obraz pobudzeń na wyjściach modeli neuronów jąder ślimakowych, stanowiący odpowiedź symulowanego modelu systemu słuchowego na pobudzenie sygnałem testowym

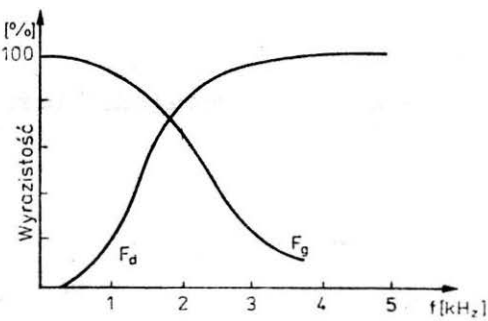
tych w sygnale są wydobywane i przekazywane do mózgu. Wyniki modelowania mają praktyczne zastosowanie przy konstruowaniu urządzeń do automatycznego rozpoznawania sygnałów dźwiękowych, głównie mowy.

## 3.3. Psychologiczne aspekty percepcji mowy

Przytoczony i obszernie przedyskutowany w poprzednim rozdziale model naturalnego, biologicznego systemu percepcji słuchowej pozwala wskazać na kilka aspektów psychologicznych, które mogą być przydatne przy opracowywaniu technicznych systemów analizy sygnałów i próbach automatycznego rozpoznawania mowy. Równocześnie cechy te i parametry wskazują na możliwość bardziej oszczędnego przesyłania sygnału mowy w systemach telekomunikacyjnych, gdyż wykrycie i opisanie tych form zakłóceń i deformacji sygnału mowy, których ucho człowieka nie rejestruje i nie analizuje, może stanowić podstawę do bardziej tolerancyjnego traktowania pewnych mankamentów urządzeń transmitujących sygnały, a w dalszej kolejności może stanowić podstawę do oszczędniejszego projektowania i tańszego realizowania tych systemów.

Jako pierwszą należy odnotować możliwość częstotliwościowego ograniczenia sygnału mowy. Wynika to z jednej strony z dolnoprzepustowych własności ucha zewnętrznego i środkowego, uwidocznionych w przytoczonym wyżej modelu, z drugiej zaś z badań psychologicznych, wskazujących na efekty pogarszania się warunków percepcji mowy przy ograniczaniu jej pasma od góry i od dołu. Przykładowe wyniki takich badań, przytoczone na rys. 3-33, obrazują obniżenie wyrazistości sylab w zależności od częstotliwości granicznej filtracji odpowiednio dolno- i górnoprzepustowej. Łatwo zauważyć, że ograniczenie pasma sygnału polegające na odcięciu fragmentów widma poniżej 300 Hz i powyżej 3500 Hz nie prowadzi (przy braku zakłóceń) do zauważalnego obniżenia zrozumiałości sygnału mowy. Ponieważ widmo sygnału mowy — szczególnie w zakresie głosek szumowych — rozciąga się daleko poza ten obszar (zwłaszcza w kierunku wysokich

częstotliwości, gdzie widmo niektórych głosek ma niezerowe składowe jeszcze przy częstotliwościach powyżej 20 kHz), zatem możliwość zawężenia pasma jest bardzo znaczna. Możliwe jest zresztą, przy dobrych warunkach transmisji sygnału, dalsze zwężenie pasma — jeśli oczywiście godzimy się na ograniczenie wyrazistości, a więc i zrozumiałości mowy. Tego typu akceptacja niepełnej zrozumiałości może być uzasadniona faktem, że treść wy-




3-33. Przybliżony przebieg zależności wskazujących spadek wyrazistości sylab przy badaniach transmisji mowy w funkcji granicznej częstotliwości przy filtracji dolnoprzepustowej ($F_d$) i górnoprzepustowej ($F_g$). Podana zależność ma charakter orientacyjny, gdyż dane doświadczalne prezentowane przez różnych autorów różnią się w szczegółach

powiedzi można rozumieć (kontekstowo uzupełnić) także w przypadku niepełnego zrozumienia jej oddzielnych elementów. Okazuje się, że poziom zrozumiałości wystarczający do tego, aby komunikacja przebiegała sprawnie choć nie bez spornego wysiłku ze strony obydwu porozumiewających się stron, można osiągnąć w przypadku przesyłania pasma o szerokości około 1000 Hz. Pasmo to, w stosunku do wyżej omówionego zakresu tzw. telefonicznego (300—3500 Hz) może być zawężone symetrycznie z obydwu stron, gdyż liczne doświadczenia wykazały, że we wspomnianym przedziale częstotliwości informacje niezbędne do identyfikacji sygnału mowy rozłożone są stosunkowo równomiernie i „wycięcie" jakiejkolwiek części tego pasma powoduje podobne (w sensie miar ilościowych) zmniejszenie wyrazistości zrozumiałości mowy. Warto dodać, że tzw. środek widma sygnału mowy przypada na częstotliwość ok. 1750 Hz, gdyż podobny spadek zrozumiałości jest efektem odcięcia w widmie wszystkich składowych powyżej lub poniżej tej wartości.

Skutki spostrzeżenia, że sygnał mowy może być (bez utraty możliwości prawidłowej jego percepcji) ograniczony częstotliwościowo, są wielorakie i generalnie (z punktu widzenia technika) korzystne. Przy wprowadzaniu informacji do maszyny cyfrowej pozwala to na stosowanie dłuższego kroku dyskretyzacji, a w efekcie prowadzi do oszczędnego gospodarowania pamięcią maszyny. Przy transmisji mowy drogą kablową lub radiową umożliwia to zwielokratnianie transmisji przez stosowanie wielu pasm separowanych częstotliwościowo, do przesyłania wielu rozmów z wykorzystaniem pojedynczego łącza. Naturalnie każde ograniczenie pasma sygnału mowy wpływa niekorzystnie na wrażenie naturalności i subiektywne odczucie jakości sygnału, co doskonale jest znane osobom, które słuchają muzyki

z radioodbiornika lub magnetofonu o niskiej jakości. Jednak zrozumiałość sygnału mowy może nie ulegać istotnym ograniczeniom nawet przy drastycznych ograniczeniach pasma.

Inny wniosek z analizy funkcjonowania systemu słuchowego i z badań wykonywanych przy użyciu jego modelu dotyczy częstotliwościowej rozdzielczości słuchu. Rozróżnianie bliskich (częstotliwościowo) tonów następujących po sobie w dziedzinie czasu dokonywane jest w uchu z bardzo dużą dokładnością, co dawało w przytoczonych wyżej rozważaniach podstawę do określania dobroci analizatora słuchowego na poziomie $Q = \frac{f_0}{\Delta f} \approx 200$.

Jednak równocześnie występujące tony mogą się wzajemnie maskować, przy czym zjawisko to zależy od wielu czynników: warunków eksperymentu (słuchanie jedno- lub dwuuszne), natężeń dźwięku maskowanego i maskującego, ich wysokości i charakteru (ton, szum) itd. Badacze wymienionych zjawisk opisują je wprowadzając zazwyczaj tak zwane pasma krytyczne. Z bardziej znanych definicji pasma krytycznego wymienić warto zaproponowane przez Fletchera pasma określane ma podstawie zagłuszania tonu o częstotliwości $f$ przez szum zawarty w pasmie o częstotliwościach $f \pm \frac{\Delta f}{2}$, przy czym moc akustyczna tonu i szumu są jednakowe. Przyjmując $\Delta f$ jako szerokość pasma krytycznego, możemy narząd słuchu traktować (pod względem zdolności analizy spektralnej sygnałów dźwiękowych) jak zestaw filtrów pasmowych o szerokościach pasm odpowiadających pasmom krytycznym.

Szerokości pasm krytycznych opisywane przez różnych autorów różnią się od siebie znacznie, co ma związek z różnymi warunkami, w jakich były wyznaczone. Jednakże reguła, że szerokość pasma $\Delta f$ wzrasta ze wzrostem częstotliwości środkowej „filtru" $f$ jest niezmienna. W przybliżeniu można przyjąć, że szerokość pasma pozostaje stała w zakresie dolnych częstotliwości do ok. 800 Hz i wynosi — według różnych autorów — od 50 do 100 Hz (najczęściej wymieniana wartość $\Delta f = 60$ Hz), natomiast dla częstości wyższych $\Delta f$ rośnie w przybliżeniu proporcjonalnie do $\log f$ i osiąga przy $f = 8000$ Hz wartości od 500 Hz do 1800 Hz (najczęściej 600 Hz). Należy podkreślić, że podane wartości $\Delta f$ odpowiadają słyszeniu jednousznemu, przy słyszeniu dwuusznym pasma krytyczne są węższe i wynoszą od 30 do 60% podanych wyżej wartości.

Ponadto system słuchowy z bardzo dużą precyzją lokalizuje maksima amplitudowo-częstotliwościowej charakterystyki sygnału, co jest szczególnie ważne przy percepcji mowy, a ma zapewne związek z omawianym przy dyskutowaniu struktury modelu ucha mechanizmem „wyostrzania" charakterystyk częstotliwościowych. Przykładowo można podać, że przy lokalizacji maksimum częstotliwości wynoszącej 700 Hz dostrzegane słuchem przemieszczenia maksimum nie przewyższają 10 Hz, co w porównaniu z omówionymi wyżej szerokościami pasm krytycznych wydaje się wynikiem niewiarygodnym. Podobnie przy częstotliwości 2000 Hz wykrywane słuchem są przemieszczenia maksimum obwiedni widma wynoszące 20 Hz, co jest jeszcze

bardziej zdumiewające. Warto porównać te fakty z omówioną w poprzednim rozdziale własnością systemu artykulacji mowy naturalnej, w którym sygnalizowano zdolność narządów mowy do kształtowania wnęk rezonansowych wywołujących lokalne koncentracje energii w widmie mowy, zwane formantami. Narzuca się wniosek, że formantowa struktura wielu artykułowanych głosek ma związek z własnościami słuchu, ułatwiającymi głównie lokalizację maksimów obwiedni widma.

Rozdzielczość amplitudowa słuchu jest również przedmiotem licznych badań, lecz tu wyniki różnych badaczy mniej różnią się od siebie. Na ogół przyjmuje się, że minimalne odczuwalne słuchem zmiany głośności odpowiadają różnicy poziomów wynoszącej 0,6 dB, przy czym dla szczególnie słabych dźwięków ten próg podwyższa się do $2 \div 3$ dB. Przytoczone wartości dotyczą rozdzielczości amplitudowej słuchu badanej w warunkach laboratoryjnych. Rzeczywiste zdolności rozróżniania amplitudy dźwięków mowy są mniejsze. I tak często cytowane są wyniki badań Flanngana, który stwierdził, że w najistotniejszych z percepcyjnego punktu widzenia rejonach wierzchołków formantów dostrzegalne zmiany poziomu wynoszą około $+3$ dB i $-6$ dB. Ponadto znany jest fakt, że krótkotrwałe zmiany poziomu sygnału nie są przez ucho człowieka wykrywane. Wiąże się to z dużą stałą czasową słuchu, która wynosi $20 \div 30$ ms przy narastaniu i $200 \div 250$ ms przy opadaniu sygnału. W rezultacie ucho nie reaguje także na zmiany szybkości narastania i opadania mocy sygnału mowy — o ile nie przekraczają one wartości 100 dB/s lub nie trwają dłużej niż 20 ms.

Percepcyjne własności i możliwości człowieka co do sygnału mowy są często opisywane z wykorzystaniem pojęć wyrazistości i zrozumiałości elementów mowy w określonych warunkach. Dla potrzeb badań psychoakustycznych rozwinięto obszerną teorię wyrazistości i zrozumiałości, z licznymi wzorami, tabelami i nomogramami. W uproszczeniu można przyjąć, że wyrazistość sygnału mowy przesyłanego lub analizowanego w pasmie częstotliwości o szerokości $F$ i uwzględniającego zakres dynamiki $D$ można wyrazić wzorem

$$A = kDF$$

gdzie współczynnik $k \approx 0{,}95 \cdot 10^{-5}\ \mathrm{dB^{-1}Hz^{-1}}$ normalizuje współczynnik wyrazistości w przedziale (0,1). Podkreślić należy, że przytoczone oszacowanie stanowi pierwsze przybliżenie; dokładniejsze wzory wymagają analizy rozkładu sygnału na osi częstotliwości z uwzględnieniem pasm krytycznych mowy, rozkładu prawdopodobieństwa występowania formantów oraz poziomu sygnału i poziomu szumu. Odpowiednie wzory o większej dokładności i złożoności można, stosownie do potrzeb, znaleźć w literaturze.

Wyrazistość sygnału mowy związana jest z jego zrozumiałością, przy czym zależność ta ma charakter monotonicznie rosnący, ale nieliniowy. Zależna jest od tego, czy rozpatrujemy zrozumiałość głosek sylab, logatomów*),

*) Logatomy są zestawami głosek, które mają podobną budowę jak wyrazy, ale nie są sensownymi (mającymi znaczenie) wyrazami rozważanego języka. Używane są w badaniach fonetycznych i psychologicznych do badań nad percepcją mowy.

wyrazów czy całych wypowiedzi. Odpowiednie tabele i diagramy znaleźć można w podanej na końcu książki literaturze. Z wystarczającym dla praktycznych zastosowań przybliżeniem można przyjąć, że między wartościami wyrazistości wynoszącymi $A = 0{,}1$ (zrozumiałość bliska 0%) a punktem $A = 0{,}6$ (zrozumiałość ok. 80%) ma miejsce proporcjonalna (liniowa) zależność zrozumiałości od wyrazistości. Dla większych wyrazistości odpowiadające im przyrosty zrozumiałości są mniejsze, przy czym oczywiście docelowo przy wyrazistości $A = 1$ zrozumiałość osiąga 100%. Przytoczone dane mają jednak charakter orientacyjny, gdyż różni badacze przytaczają znacznie różniące się charakterystyki, a analizę wydatnie komplikuje wpływ kontekstu. Konieczne jest także uwzględnianie faktu nierównomiernego prawdopodobieństwa występowania różnych fonemów i ich zestaw (diad, triad, sylab), a także całych wyrazów. Wobec powyższego badania zrozumiałości mowy prowadzi się zwykle bezpośrednio: grupa osób słucha nadawanych sygnałów i notuje swoje rozpoznania, co następnie jest porównywane z z wzorcem badanego tekstu. Prace takie są bardzo mozolne, a wnioskowanie prowadzone jest z wykorzystaniem metod statystycznych. Jest to jednak w praktyce jedyna droga, gdyż badania wyrazistości, możliwe do przeprowadzenia na drodze aparaturowych pomiarów, trzeba traktować jedynie jako wstępne, weryfikowane psychologicznie, orientacyjne dane.

# 4

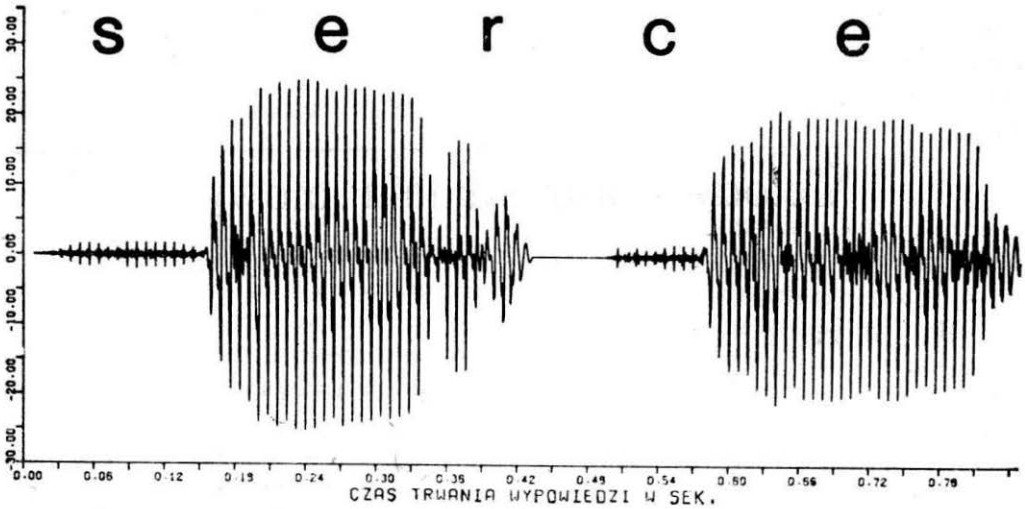
# Metody opisu sygnału mowy

## 4.1. Opis sygnału w dziedzinie czasu

Sygnał mowy może być badany i opisywany na różne sposoby, przy czym każda z omawianych metod ma swoje specyficzne zalety i wady. Dla pełnego obrazu dokonamy więc w tym rozdziale dyskusji różnych metod opisu. Niektóre z nich odgrywają obecnie rolę dominującą i są powszechnie stosowane, inne zaś — w tym opis sygnału w dziedzinie czasu — utraciły obecnie wiele ze swego znaczenia.

Sygnał mowy, traktowany jako przebieg czasowy (rys. 4-1) ma skomplikowany przebieg, będący odzwierciedleniem złożonego charakteru procesu jego artykulacji. Na parametry sygnału ma wpływ jego źródło (którym są albo drgające wiązadła głosowe, albo szum turbulentnego przepływu powietrza przez przewężenia w narządach mowy) i własności dynamiczne kanału głosowego, formującego strukturę sygnału. Operując w dziedzinie czasu sygnał można matematycznie opisać za pomocą splotu przebiegu czasowego sygnału źródła $u_g(t)$ i odpowiedzi impulsowej kanału głosowego $h(t)$:

$$u(t) = \int_0^t h(t-\tau)u_g(\tau)\mathrm{d}\tau \tag{4.1}$$




4-1. Przebieg czasowy sygnału mowy (wyraz *serce*; głos męski)

Interpretacja przytoczonego wzoru wskazuje, że w sygnale czasowym właściwości źródła i właściwości kształtującego dźwięk kanału głosowego są ze sobą ściśle powiązane, nie można zatem rozpatrywać ich oddzielnie, gdyż kształtują obraz przebiegu wspólnie. Tymczasem, jak wiemy, w procesie artykulacji zmieniana jest głównie struktura kanału głosowego, modulującego sygnał, a więc składnik zapisany jako $h(t)$, zaś przebieg czasowy $u_g(t)$ jest zmieniany nieznacznie (szczególnie dla głosek bezdźwięcznych). Ponieważ przebieg czasowy sygnału mowy jest kształtowany przez składniki przypadkowe i zdeterminowane przy równoprawności obydwu, wobec tego obraz przebiegu czasowego różnych wypowiedzi tego samego mówcy może wykazywać więcej wzajemnego podobieństwa niż obraz tej samej wypowiedzi artykułowanej różnymi głosami. Dyskwalifikuje to praktycznie czasową postać sygnału w badaniach nad automatycznym rozpoznawaniem mowy, a także w tych pracach z zakresu głosowej komunikacji pomiędzy ludźmi, które koncentrują uwagę na semantycznej stronie języka i badają — na przykład — skuteczność określonego systemu telekomunikacyjnego z punktu widzenia jego przydatności do przekazywania zrozumiałej mowy. W badaniach nad automatycznym rozpoznawaniem osób mówiących lub nad osobniczymi własnościami sygnału mowy postać czasowa jest również niechętnie stosowana. Przyczyna jest identyczna; silne związanie aspektu osobniczego i semantycznego w tej postaci sygnału.

Reasumując, opis sygnału w postaci czasowej jest na ogół mniej przydatny od innych, omówionych dalej metod jego prezentacji. Jest on jednak ważny, gdyż stanowi punkt wyjścia do wszelkich dalszych metod, ponieważ sygnał mowy jest pierwotnie zawsze dostępny w postaci przebiegu czasowego. Wobec tego jeśli nawet w dalszej analizie będzie używać się przekształconej formy sygnału, to przez etap operowania przebiegiem czasowym trzeba przejść. Sygnał w postaci czasowej może być poddany wielu przekształceniom ułatwiającym dalszą jego analizę i obróbkę. Podany więc będzie

przegląd niektórych spośród tych przekształceń, aby nie wracać do tego tematu wielokrotnie w czasie dalszych rozważań.

Charakterystyki sygnału w dziedzinie czasu mogą dotyczyć jego amplitud i szybkości zmian. Amplituda sygnału może być mierzona w sposób bezwzględny lub przy stosowaniu poziomu odniesienia, którym zgodnie z międzynarodową normalizacją jest sygnał o natężeniu (mocy akustycznej) $10^{-16}$ W/cm$^2$. Mierzony jest również stosunek sygnału użytecznego do szumu. Stosuje się miary logarytmiczne, gdyż rozpiętość między dźwiękami o mocy największej i najmniejszej, które może przyjmować nasze ucho, sięga dwunastu rzędów wielkości i wyrażanie natężeń w skali liniowej wiązałoby się z koniecznością używania bardzo dużych liczb i wielocyfrowych zapisów. Poza tym miara logarytmiczna jest najbardziej naturalna dla skali intensywności dźwięku, gdyż — podobnie jak dla większości zmysłów człowieka — subiektywne wrażenie głośności związane jest raczej z logarytmem wartości bodźca, a nie z samą wartością. W fizjologii wyraża to znane prawo Webera — Fechnera, stwierdzające, że minimalny dostrzegalny przyrost dowolnego bodźca $\Delta p$ jest proporcjonalny do wartości bodźca $p$:

$$\Delta p = kp \tag{4.2}$$

Proste przekształcenie wzoru (4.2) wskazuje na celowość stosowania właśnie logarytmicznych miar przy określaniu związku między wyrażeniem zmysłowym a działającym bodźcem, przy czym reguła ta jest uniwersalna. Prawo Webera ma charakter przybliżony; zależność (4.2) powinna mieć bardziej złożony kształt, jeśli ma dokładnie opisywać wrażenia zmysłowe. Ponadto logarytmiczna zależność nie obowiązuje dla bardzo dużych i dla bardzo małych sygnałów, gdzie na ogół pojawiają się załamania charakterystyki typu nasycenie i próg nieczułości (por. rys. 4-2). Jednak przyjęcie logarytmicznej skali dźwięku można uznać za uzasadnione.

Natężenie dźwięku będzie wyrażane w decybelach. Intensywność dźwięku o natężeniu $I$ wyraża się za pomocą wzoru:

$$i = 10 \log \frac{I}{I_0} \tag{4.3}$$

gdzie $I_0$ jest natężeniem odniesienia (wspomnianym wyżej progiem słyszalności, odpowiadającym natężeniu dźwięku na poziomie $10^{-10}$ $\mu$W/cm$^2$, lub poziomem szumów, wyznaczającym efektywną intensywność dźwięku w danych warunkach transmisji). Natężenia dźwięku ($I$ oraz $I_0$) mierzone w $\mu$W/cm$^2$ są trudne do bezpośredniego pomiaru, często więc określając intensywność dźwięku posługujemy się wartościami ciśnienia akustycznego — odpowiednio $p$ i $p_0$. Ponieważ natężenie dźwięku jest proporcjonalne do kwadratu ciśnienia akustycznego, to wzór (4.3) przyjmuje równoważną postać:

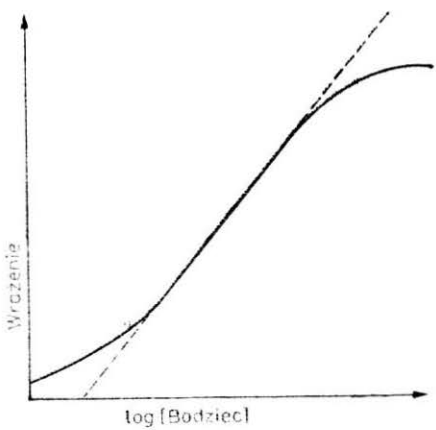
$$i = 20 \log \frac{p}{p_0} \tag{4.4}$$

Progowa wartość $p_0$ w warunkach normalnych wynosi 20,4 Pa (przy tem-

peraturze 20°C i ciśnieniu 1013,25 hPa). Wyliczyć można ją również ze wzoru:

$$p_0 = \sqrt{\frac{1}{t_w}\int_0^{t_w} p_s^2(t)\,dt} \tag{4.5}$$

w przypadku, kiedy brany jest pod uwagę przebieg sygnału zakłócającego $p_s(t)$ rozważanego w czasie $t_w$. Natężenie dźwięku w czasie normalnego,




4-2. Zależność wrażenia zmysłowego (odczuwanego subiektywnie) od obiektywnie mierzonej wartości fizycznej odpowiedniego bodźca ma zwykle charakter logarytmiczny. W układzie współrzędnych „wrażenie-log (bodziec)" odpowiada mu linia prosta o ustalonym nachyleniu, będącym miarą czułości receptora. Rzeczywista zależność odbiega jednak od tej teoretycznej zależności dla bardzo dużych i dla bardzo małych bodźców

równomiernego wypowiadania kolejnych słów i fraz waha się w szerokich granicach, gdyż niektóre fragmenty mowy (zwłaszcza samogłoski) charakteryzują się wielokrotnie wyższym poziomem sygnału niż inne na przykład głoski *f* czy *h*. Fakt ten sprawia, że w czasie artykulacji dowolnej wypowiedzi występują znaczne wahania intensywności sygnału, przy czym istnieje możliwość takiego dobrania zestawu słów (na przykład prostych komend), aby ten „amplitudowo-czasowy profil" wypowiedzi wystarczał do jej jednoznacznej (w rozważanym słowniku) identyfikacji. Zestawienie względnego poziomu poszczególnych grup głosek języka polskiego dla typowego wysiłku głosowego i w miarę równomiernej wymowy przedstawia się następująco:

— samogłoski: 32÷40 dB,
— spółgłoski boczne i samogłoski niesylabiczne: 35 dB,
— spółgłoski nosowe: 30 dB,
— spółgłoski trące dźwięczne oraz drżące: 27 dB,
— spółgłoski zwarto-trące dźwięczne: 26 dB,
— spółgłoski trące bezdźwięczne: 25 dB,
— spółgłoski zwarto-trące bezdźwięczne: 24 dB,
— spółgłoski szumowe wyjątkowo małej energii — *f*, *h* — 20 dB.

Przytoczone dane mają charakter przybliżony, gdyż na moc określonego fonemu mają wpływ indywidualne własności wymowy określonej osoby, tempo mowy, a także cechy prozodyczne wypowiedzi (inaczej kształtuje się

amplituda tych samych głosek na początku, w środku i na końcu zdania, inaczej w sylabach akcentowanych, a inaczej w nie akcentowanych, wreszcie różnice mogą wynikać z kontekstu wypowiedzi i wpływu głosek otaczających daną, rozpatrywaną w badaniach).

Skala amplitudowa sygnału mowy może być przekazywana i przetwarzana bez żadnych zmian. Może również podlegać transformacjom, gdyż spostrzegane przez człowieka różnice intensywności sygnału mają charakter względny (por. wzór (4.2)), a również stosunek sygnału do szumu jest zupełnie inny dla sygnałów o małej amplitudzie, a inny (znacznie korzystniejszy) dla sygnałów o dużej amplitudzie. W rezultacie możliwe jest dokonywanie kompresji amplitudy sygnału przed jego przetwarzaniem lub przesyłaniem, a następnie — po przesłaniu lub przetworzeniu — możliwe jest proste odtworzenie pierwotnej postaci sygnału przez poddanie go operacji odwrotnej do kompresji. Zabiegi te ogólnie można opisać wrowadzając nieliniową f u n k c j ę k o m p r e s j i $F(x)$, która przy wprowadzaniu sygnału mowy $u(t)$ do systemu przetwarzającego lub przesyłającego wykorzystywana jest wprost:

$$u'(t) = F[u(t)] \tag{4.6}$$

zaś przy odtwarzaniu sygnału z przetworzonego wzorca stosuje się to odwzorowanie odwrotnie:

$$u(t) = F^{-1}[u'(t)] \tag{4.7}$$

Zależność $F(x)$ może być w zasadzie dowolnej postaci, powinna jedynie spełniać warunek malejącego nachylenia charakterystyki ze wzrostem argumentu $x$:

$$\frac{\mathrm{d}F}{\mathrm{d}x} = f(x), \quad \frac{\mathrm{d}F}{\mathrm{d}x} > 0, \quad \frac{\mathrm{d}f}{\mathrm{d}x} < 0 \text{ dla } x > 0 \tag{4.8}$$

Przykładem jest funkcja:

$$F(x) = \ln x \tag{4.9}$$

której odpowiada

$$f(x) = \frac{1}{x} \tag{4.10}$$

przy czym ze względu na własności funkcji logarytmicznej trzeba rygorystycznie przestrzegać warunku $x > 0$, co dla realnego sygnału mowy oznacza konieczność rozpatrywania wartości bezwzględnej z przebiegu i korygowania tej wartości w pobliżu zera. Z tych względów jako międzynarodowy standard przyjęto stosowanie funkcji $F(x)$ w postaci zależności liniowo-logarytmicznej. Istnieją dwie odmiany tego standardu:

— amerykańska, wyrażająca się tak zwaną charakterystyką typu $\mu$:

$$F(x) = \operatorname{sgn}(x) \frac{\ln(1+\mu|x|)}{\ln(1+\mu)} \tag{4.11}$$

— europejska (opracowana przez Niemiecki Urząd Poczt), wyrażająca się tak zwaną charakterystyką typu A:

$$F(x) = \begin{cases} \operatorname{sgn}(x)\dfrac{1+\ln(A\cdot|x|)}{1+\ln A} & \text{dla } 1 \geqslant x \geqslant \dfrac{1}{A} \\ \operatorname{sgn}(x)\dfrac{A|x|}{1+\ln A} & \text{dla } \dfrac{1}{A} \geqslant x \geqslant 0 \end{cases} \tag{4.12}$$

Stosowanie obu wymienionych charakterystyk wymaga, aby sygnał $u(t)$, reprezentowany we wzorach (4.11) i (4.12) przez argument $x$, był unormowany, przy czym dla wzoru (4.11) wymagane jest, aby:

$$-1 \leqslant u(t) \leqslant 1 \tag{4.13}$$

zaś dla wzoru (4.12) wymagane jest dodatkowo $x \geqslant 0$, co zresztą wynika jednoznacznie z zapisu tego wzoru.

Występujące we wzorach (4.11) i (4.12) parametry $\mu$ oraz $A$ (od których zresztą pochodzą nazwy odpowiednich charakterystyk) mogą być dobierane tak, aby optymalnie dopasować charakterystykę do aktualnych potrzeb. W szczególności parametr $\mu$ we wzorze (4.11) umożliwia kształtowanie stosunku sygnału do szumu w sygnale wynikowym. Gdy $\mu$ przybiera większe wartości, stosunek ten pozostaje stały w dość dużym zakresie amplitud sygnału (co jest zaletą), ale ma równocześnie mniejszą wartość (co jest wadą). Wybór $\mu$ musi więc być kompromisem między wymaganiami dokładności i stopnia kompresji. Interpretacja $\mu$ jest przy tym dość oczywista: określa on poziom wejściowego sygnału, przy którym charakterystyka zmienia się z liniowej w logarytmiczną. Interpretacja $A$ we wzorze (4.12) jest podobna, przy czym zalecane jest przyjmowanie wartości $A$ równej 87,7. Zysk kompresji (wyrażający się różnicą wzmocnienia dla małych i dla dużych sygnałów) wynosi przy tym 24 dB.

Szybkość zmian sygnału zależna jest od jego amplitudy i granicznej częstotliwości, zgodnie z oszacowaniem Bersteina:

$$\sup\left|\frac{\mathrm{d}^k x}{\mathrm{d}t^k}\right| = \omega_g^k \sup|x(t)| \tag{4.14}$$

gdzie:

$x(t)$ — przebieg sygnału mowy,

$\omega_g$ — pulsacja odpowiadająca częstości granicznej $f_g$ sygnału ($\omega_g = 2\pi f_g$)

Widać więc, że czasowe parametry sygnału są w tym zakresie determinowane przez jego własności widmowe, omawiane w dalszym podrozdziale. Warto jedynie zwrócić uwagę na fakt, że emitowany sygnał mowy ma bardzo bogate widmo i rozciąga się bardzo daleko w kierunku wysokich częstotliwości (zwłaszcza dla głosek szumowych). Natomiast dla jego identyfikacji i poprawnej percepcji wystarczające jest posługiwanie się sygnałem odfiltrowanym, którego częstotliwość graniczna $f_g$ jest ustalona przez parametry użytego filtru.

Szybkość zmian sygnału w dziedzinie czasu może być dodatkowo zwiększana w procesie preemfazy, polegającym na uwydatnieniu w sygnale jego składowych o wysokich częstotliwościach. Potrzeba stosowania preemfazy wynika z tego, że w naturalnym sygnale mowy składowe o dużych często-

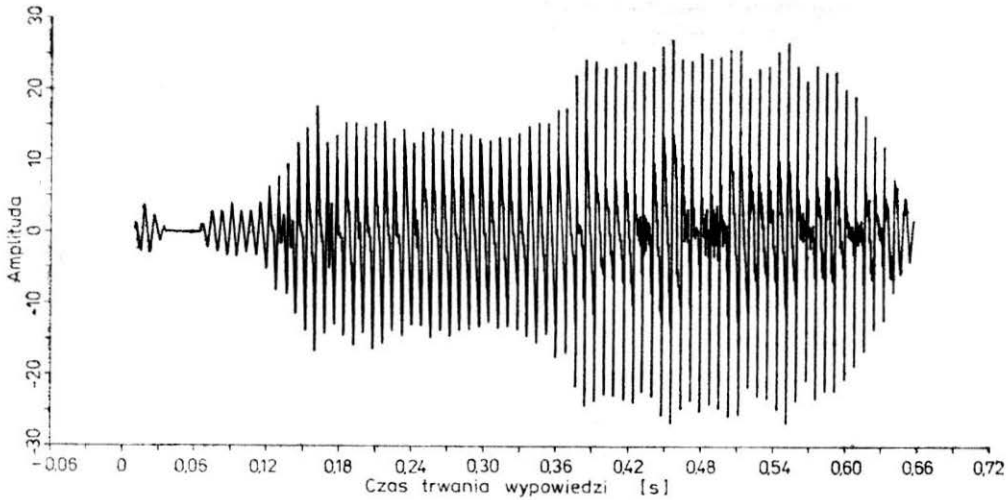
tliwościach mają mniejszą moc i w wyniku tego stosunek sygnału do szumu, korzystny w zakresie składników małoczęstotliwościowych osiąga mniejsze wartości dla sygnałów wielkoczęstotliwościowych. Ujmując to samo w jeszcze inny sposób można powiedzieć, że różnice dynamiki sygnału w zakresach mało- i wielkoczęstotliwościowych sięgająca 50 dB, co utrudnia znalezienie poprawnej wartości wzmocnienia sygnału. Jeśli dokona się tak dużego wzmocnienia, aby sygnał dla wielkich częstotliwości był „czytelny", to nastąpi przesterowanie aparatury dla małych częstotliwości. Jeśli zaś unormuje się wzmocnienie biorąc pod uwagę poprawne przenoszenie składowych małoczęstotliwościowych, to składniki wielkoczęstotliwościowe znikną całkowicie.

Zabieg, który częściowo usuwa wskazane niedogodności — preemfaza — może być traktowany jako filtracja formująca, osłabiająca generalnie składowe sygnału o małych częstotliwościach i relatywnie wzmacniająca składowe o częstotliwościach dużych. Obraz takiego wzmocnienia może być różny. Na przykład, rozpatrywane bywa prawo filtracji określające zasadę liniowego stłumienia (odwrotnie proporcjonalnego do częstotliwości) składowych widma o częstościach niższych niż 5 kHz i przenoszenia bez zmian pozostałej części widma, rozważane bywają także i inne reguły. Każda z nich ma w istocie arbitralny charakter, gdyż poszczególne elementy mowy charakteryzują się różnym stopniem stłumienia składowych wielkoczęstotliwościowych i wymagają indywidualnej odmiennej korekty, a ponadto charakterystyki, o których mowa, są silnie uzależnione osobniczo, zatem każda przyjęta reguła będzie poprawnie funkcjonowała jedynie dla pewnego podzbioru głosów i sposobów wymowy. Najczęściej przyjmuje się, że preemfaza jest — z matematycznego punktu widzenia — różniczkowaniem sygnału:

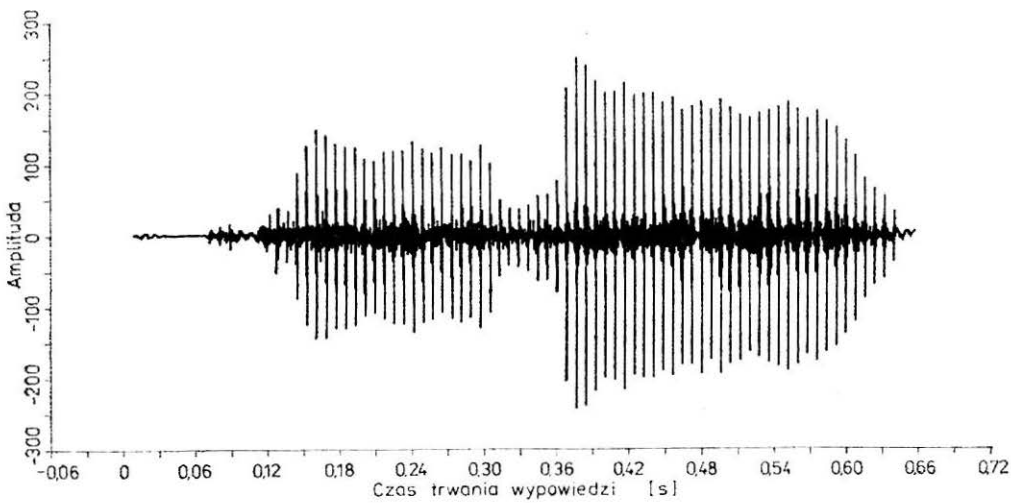
$$x_p(t) = \frac{\mathrm{d}}{\mathrm{d}t}\,[x(t)] \tag{4.15}$$

Podejście takie ma wiele zalet: jest proste w rozważaniach teoretycznych, daje proporcjonalne do częstotliwości wzmocnienie sygnału, co prawie idealnie odpowiada stopniowi stłumienia tegoż sygnału dla głosek dźwięcznych. Jest także stosunkowo proste do realizacji zarówno w układach analogowych, jak i w cyfrowych. Skutki takiej preemfazy sygnału zobaczyć można na rys. 4-3 i 4-4. Istotnie, po zastosowaniu preemfazy liczne składowe sygnału, uprzednio niewidoczne, stają się czytelne i możliwe do analizy.

Zabiegiem realizowanym na sygnale mowy w dziedzinie czasu, a mającym istotne znaczenie przy jego przesyłaniu, przetwarzaniu i rozpoznawaniu, jest ograniczenie sygnału w dziedzinie amplitud. Wspomniano już wyżej o możliwościach i korzyściach, jakie wynikają z zastosowania do sygnału mowy technik kompresji amplitudy. Będzie także mowa o możliwościach, jakie wiążą się z przetworzeniem sygnału do postaci cyfrowej, zanim to jednak nastąpi, należy wspomnieć o udanych próbach przetworzenia sygnału do postaci fali prostokątnej o wartościach wynoszących wyłącznie $+1$ lub $-1$ (rys. 4-5). Taki skrajnie ograniczony amplitudowo sygnał mowy może



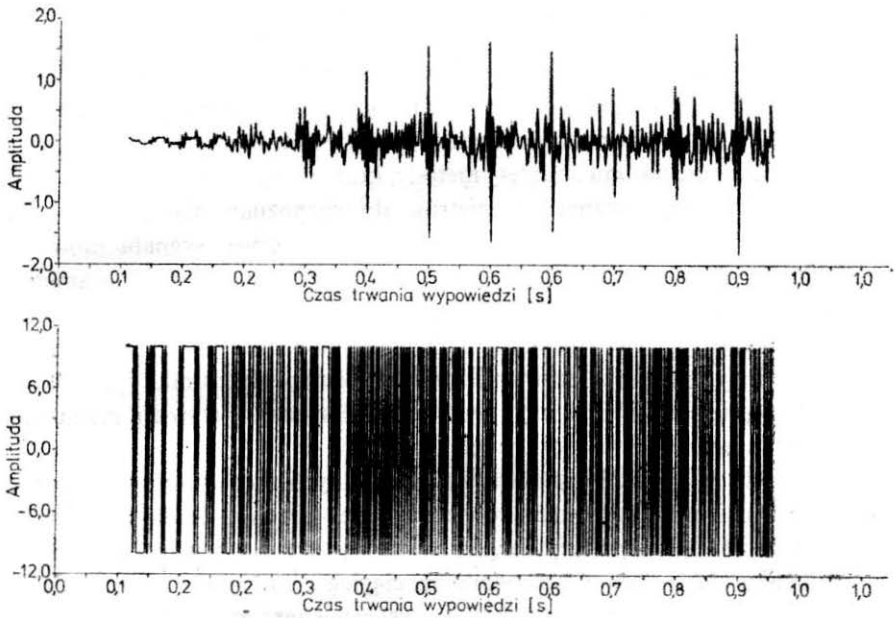

**4-3. Przebieg czasowy sygnału mowy rejestrowanego bez żadnych dodatkowych zabiegów wskazuje na dominację małych częstotliwości. Składowe o dużych częstotliwościach są mało widoczne ze względu na malejącą ze wzrostem częstotliwości wydajność energetyczną źródła tonu krtaniowego**




**4-4. Przebieg czasowy tego samego sygnału mowy z rys. 4-3, poddanego zabiegowi preemfazy. Na ogół taki przebieg znacznie lepiej nadaje się do analizy i rozpoznawania**

być poprawnie rozpoznawany przez człowieka — a o ileż mniej informacji zawiera w stosunku do sygnału oryginalnego! Fakt ten powodował przez wiele lat duże ożywienie wśród specjalistów zajmujących się sygnałem mowy, przy czym inżynierowie telekomunikacji wcześniej zrezygnowali z wykorzystania „przyciętego" sygnału mowy w celu oszczędzania łączy telefonicznych przy przesyłaniu mowy, gdyż sygnał w tej postaci jest wyjątkowo nieprzyjemny do słuchania i jego rozumienie wiąże się z dużym wysiłkiem, a ponadto transmisja fali prostokątnej w łączach fonicznych napotyka trudności. Natomiast z punktu widzenia automatycznego rozpoznawania mowy ogra-

niczony amplitudowo sygnał jest interesujący ze względu na to, że jest to sygnał cyfrowy, wygodny do wprowadzenia do maszyny cyfrowej, a w dodatku umożliwiał on — pozornie, jak się wkrótce okazało — łatwe rozpoznawanie na podstawie jednego tylko parametru: częstości przejść przez zero. Istotnie, w sygnale, który został tak krańcowo zubożony, częstość zmiany znaków sygnału (częstość przejść przez zero) była jedynym zacho-




4-5. Skrajnie ograniczony amplitudowo obraz sygnału mowy (u dołu) zachowuje wystarczającą ilość informacji, aby człowiek mógł go prawidłowo rozpoznać, pomimo że w stosunku do przebiegu oryginalnego sygnału (u góry) zachowano zgodność tylko jednego parametru: gęstości przejść przez zero. Wykorzystano początkowy odcinek wyrazu *serce* (por. rys. 4-1)

wanym parametrem. Parametr ten, oznaczany zazwyczaj $\varrho_0$, był możliwy do określenia stosunkowo prostymi środkami, łatwy do wyrażenia w postaci cyfrowej (wystarczało na przykład zliczać przejścia przez zero w ustalonym przedziale czasu i wprowadzać wyniki okresowo do maszyny cyfrowej), a ponadto zmieniał się stosunkowo powoli, co oszczędzało pamięć komputera i pozwalało wygodnie tworzyć i wykorzystywać wzorce przebiegów $\varrho_0$ dla wybranych wypowiedzi. W dodatku parametr ten miał stosunkowo prostą interpretację — reprezentował mianowicie (w przybliżeniu, gdyż pełna teoria na temat parametru $\varrho_0$ przypisuje mu znacznie więcej własności) uśrednioną częstotliwość sygnału w krótkich interwałach czasowych. Przydatność parametru $\varrho_0$ w badaniach nad rozpoznawaniem mowy wydawała się przesądzona — wszak przy wszystkich wymienionych zaletach gwarantował on ponadto prawidłowe rozpoznanie mowy, gdyż wskazywały na to pomyślne eksperymenty z odsłuchowym rozpoznawaniem „przyciętego" sygnału mowy przez ludzi!

Niestety, kolejne badania i usiłowania skonstruowania systemu rozpoznawania mowy opierające się na częstości przejść przez zero nie przynosiły

rezultatów. Parametr ten zawierał wystarczająco dużo informacji, aby ludzie mogli tak zdeformowany sygnał poprawnie interpretować, jednak nie zawierał informacji wystarczającej do tego, aby rozpoznawanie mógł przeprowadzić automat. Po raz kolejny okazało się, że możliwości mózgu są znacznie większe niż możliwości techniki.

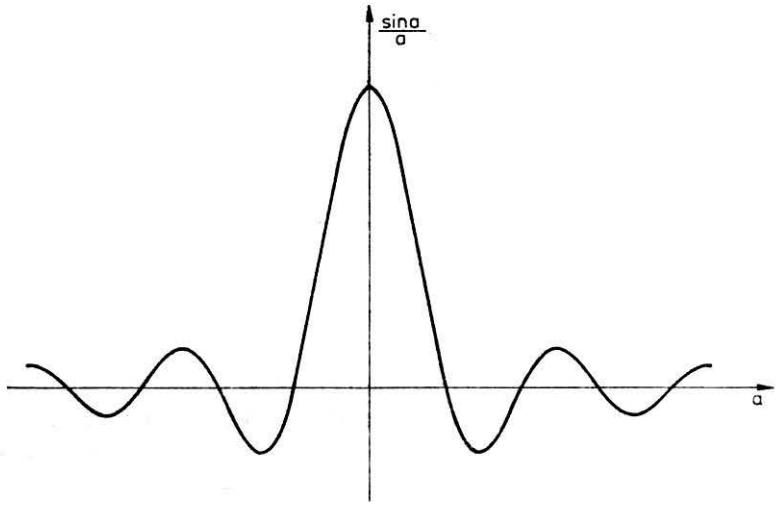
Usiłowano metodę tę „reanimować", wzbogacając ją o dodatkowe elementy. Obok $\varrho_0$, będącego częstością przejść przez zero sygnału oryginalnego, pojawiły się parametry: częstość przejść przez zero sygnału pochodnej sygnału mowy $\varrho_1$, częstość drugiej pochodnej $\varrho_2$ itd. Usiłowano także wykorzystać częstość przejść przez zero sygnału scałkowanego $\varrho_{-1}$, $\varrho_{-2}$, ... Liczba parametrów rosła, znacznie wolniej rosła jednak jakość rozpoznawania, tracono natomiast zalety metody, które miały opierać się na prostocie i łatwości pozyskiwania parametrów do rozpoznawania. Równocześnie powstały atrakcyjne „konkurencyjne" techniki opisu sygnału mowy dla potrzeb telekomunikacji i cybernetyki, w następstwie czego parametr $\varrho_0$ stracił na znaczeniu. Dziś niewielu badaczy i niewiele ośrodków prowadzi poszukiwania metod rozpoznawania opartych na tak przetworzonym sygnale mowy, zwłaszcza że w międzyczasie postęp elektroniki spowodował, że dostępne pamięci i techniki przetwarzania sygnałów stawiają przed badaczami zupełnie nowe możliwości.

Droga do większości spośród wzmiankowanych możliwości wiedzie przez sygnały cyfrowe i komputerowe metody przetwarzania. Zatem na zakończenie tego rozdziału, poświęconego analizie sygnału mowy w dziedzinie czasu, zajmiemy się metodami przetworzenia ciągłego, analogowego (naturalnego) sygnału mowy na postać cyfrową — dyskretną zarówno w dziedzinie czasu, jak i w dziedzinie amplitud, a w dodatku z reguły kodowaną.

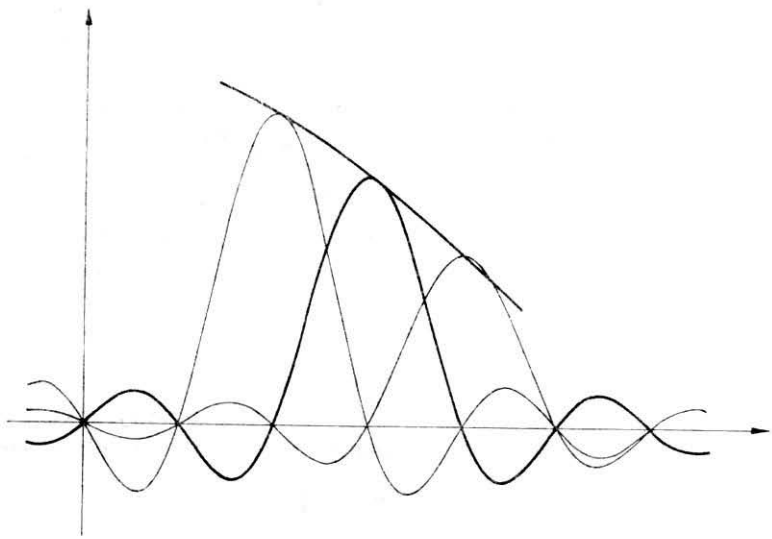
Możliwość zamiany ciągłego sygnału mowy na serię dyskretnych próbek, pobieranych (najczęściej) w równoodległych dyskretnych momentach czasu, wynika ze znanego twierdzenia Kotielnikowa-Shannona. Na podstawie tego twierdzenia można odtworzyć ciągły sygnał mowy $x(t)$ ze zbioru próbek tego sygnału, danych w dyskretnych momentach czasu $t = n\,T_p$ ($n = \ldots, -2, -1, 0, 1, 2, 3, \ldots$), gdzie $T_p$ jest czasem upływającym między pobraniem kolejnych próbek (kwantem próbkowania). Można opisać to wzorem:

$$x(t) = \sum_{n=-\infty}^{\infty} x(nT_p)\,\frac{\sin\omega_g(t-nT_p)}{\omega_g(t-nT_p)} \tag{4.16}$$

Jak łatwo zauważyć, we wzorze funkcja $\frac{\sin a}{a}$ „moduluje" impulsowe wartości $x(nT_p)$ dla momentów $t \neq nT_p$. Przebieg tej funkcji pokazano na rys. 4-6. Widać z niego, że wpływ określonego składnika we wzorze (4.16) będzie malał wraz ze wzrostem różnicy $|t-nT_p|$ przyjmując wartość równą dokładnie 1 dla $t = nT_p$ $\left(\text{gdyż } \lim_{a\to 0} \frac{\sin a}{a} = 1\right)$ oraz dążąc — niestety niemonotonicznie — do zera dla $|t-nT_p| \to \infty$. Obrazowo można więc sobie przedstawić proces odtwarzania przebiegu $x(t)$ jako składanie funkcji




4-6. Przebieg funkcji sin $(a)/a$ odgrywającej bardzo istotną rolę we wszystkich obliczeniach związanych z próbkowaniem sygnałów



4-7. Przybliżona ilustracja funkcjonowania tezy twierdzenia o próbkowaniu. Próbkowany sygnał (gruba linia u góry wykresu) jest odtwarzany zgodnie ze wzorem (4.16) jako superpozycja wartości próbek sygnału w dyskretnych punktach $nT_p$ mnożonych przez funkcje $\sin(a)/a$

typu $\frac{\sin a}{a}$ umieszczonych w momentach czasu $t = nT_p$ i przemnożonych przez składowe $x(nT_p)$. Proces ten zobrazowano na rys. 4-7, gdzie górna, pogrubiona linia odtwarzana jest przez sumowanie przebiegów typu $\frac{\sin a}{a}$. Odtworzenie sygnału $x(t)$ na podstawie próbek $x(nT_p)$ możliwe jest w sposób prawidłowy jedynie pod warunkiem, że

$$T_p \leqslant \frac{\pi}{\omega_g} = \frac{1}{2f_g} \tag{4.17}$$

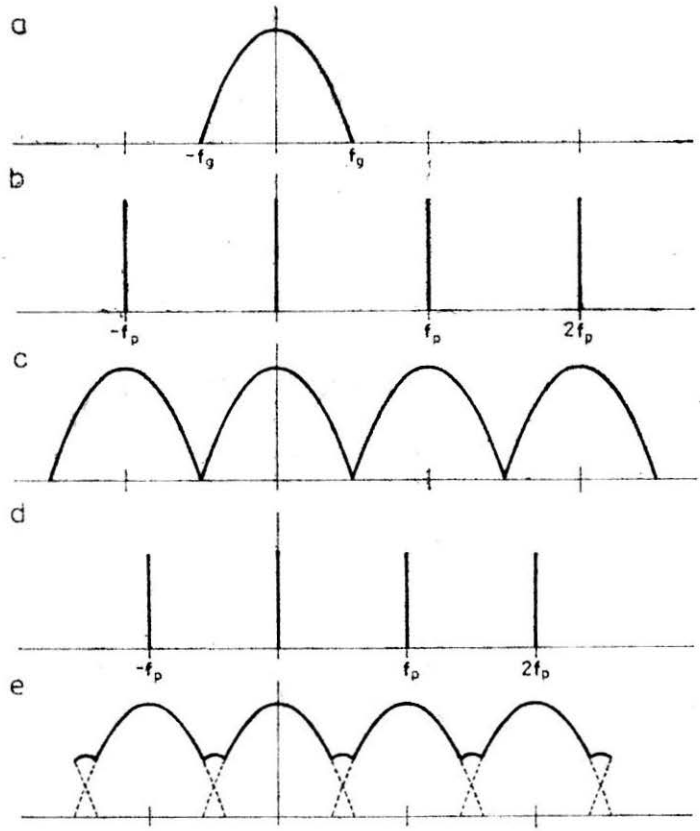
gdzie:

$f_g$ — graniczna częstotliwość sygnału $x(t)$,

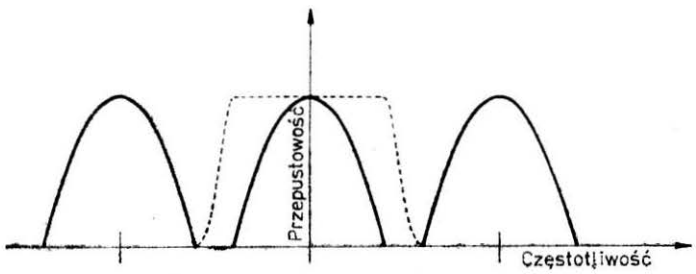
$\omega_g = 2\pi f_g$ — odpowiadająca tej częstotliwości pulsacja kątowa.

Wzór (4.17) wskazuje, że minimalna częstotliwość, z jaką należy próbkować sygnał, aby go wiernie odtworzyć z postaci cyfrowej, musi być dwukrotnie większa niż maksymalna częstotliwość składowej występującej w sygnale. Warunek zawarty we wzorze (4.17) jest bardzo kategoryczny. Jeśli nie dotrzymamy warunku (4.17) i w próbkowanym sygnale znajdą się częstotliwości większe od podwojonej częstotliwości próbkowania, to w sygnale odtworzonym zgodnie ze wzorem (4.16) pojawią się tak zwane zniekształcenia zwierciadlane. Nie wnikając w nazbyt wiele szczegółów można stwierdzić, że widmo sygnału poddanego próbkowaniu różni się od widma sygnału oryginalnego, gdyż sygnał zmieniając swoją formę z ciągłego na dyskretny wzbogaca swoje widmo o elementy wynikające z faktu próbkowania. Wyliczenia, które prowadzą do tego wniosku — pomimo ich w istocie elementarnego charakteru — są dość uciążliwe. Zamiast więc opierać się na wywodach formalnych lepiej posłużyć się intuicją. Jest faktem powszechnie znanym, że widmo sygnału okresowego ma charakter dyskretny. Przykładowo widmo czystego tonu (idealnej fali sinusoidalnej) jest złożone z pojedynczego prążka w punkcie odpowiadającym częstotliwości tej fali. Z symetrii prostego i odwrotnego przekształcenia Fouriera wynika także i odwrotna — właśnie tu potrzebna — prawidłowość: widmo sygnału dyskretnego w dziedzinie czasu (próbkowanego) będzie miało charakter okresowy. Bliższa analiza pokazuje, że okresowość ta polega na „powieleniu" widma sygnału oryginalnego $x(t)$ w odstępach (na osi częstotliwości) wynoszących $\frac{1}{T_p}$. Zilustrowano to na rys. 4-8, na którym u góry pokazano przykładowe widmo sygnału, niżej prążki pojawiające się przy prawidłowym próbkowaniu sygnału $\left(T_p = \frac{1}{2f_g}\right)$ oraz widmo sygnału poddanego próbkowaniu. Widać, że przy poprawnym próbkowaniu poszczególne części widma nie zachodzą na siebie i można je prawidłowo odtworzyć. Na rysunku 4-8d i e pokazano sytuację powstającą przy niewłaściwie wybranym okresie próbkowania. Zbyt małe wartości częstotliwości próbkowania $f_p = \frac{1}{T_p}$ powodują nakładanie się kolejnych części widma sygnału próbkowanego, w wyniku czego widmo ulega zniekształceniu. Oczywiste jest, że przy odtwarzaniu własności sygnału z takich — wadliwie dobranych — próbek czasowych dojdzie do deformacji sygnału — szczególnie w zakresie jego wysokoczęstotliwościowych składników.

Na marginesie tych rozważań można odnotować jeszcze jeden — dość oczywisty — fakt. Otóż przy przejściu od sygnału dyskretnego (cyfrowego) do analogowego konieczna będzie filtracja dolnoprzepustowa dla „wycięcia" większych częstotliwości z widma sygnału (rys. 4-9). Ponadto odtwarzając sygnał z postaci dyskretnej do postaci ciągłej, używa się tak zwanych interpolatorów, a nie korzysta się ze wzoru (4.16), którego użycie wiąże się z uciążliwymi i czasochłonnymi przeliczeniami. Zamiast więc obliczać wartość odtwarzanego sygnału dla punktów $t \neq nT_p$ można aproksymować przebieg funkcją schodkową (to znaczy przyjmować $x(t) = x(nT_p)$ dla wszyst-




4-8. Ilustracja zjawiska aliasingu: oryginalny (nie próbkowany) sygnał (a) ma widmo ograniczone ($-f_g$, $+f_g$). Widmo sygnału próbkującego (b) składa się z prążków rozmieszczonych w odległościach $f_p = 1/T_p$ wzdłuż całej osi częstotliwości. W wyniku nałożenia sygnału próbkowania na przebieg sygnału otrzymuje się sygnał dyskretny (spróbkowany), którego widmo ma charakter okresowy i składa się z widm oryginalnego procesu (a) powtarzanych w odstępach $f_p$. Jeśli $f_p \geq 2f_g$, to nie dochodzi do nakładania się widm i ich zniekształcenia (c). Wybór za małej częstości próbkowania $f_p$ powoduje zniekształcenie widma sygnału próbkowanego (e). Zjawisko to nazywane jest nakładaniem się widm lub aliasingiem

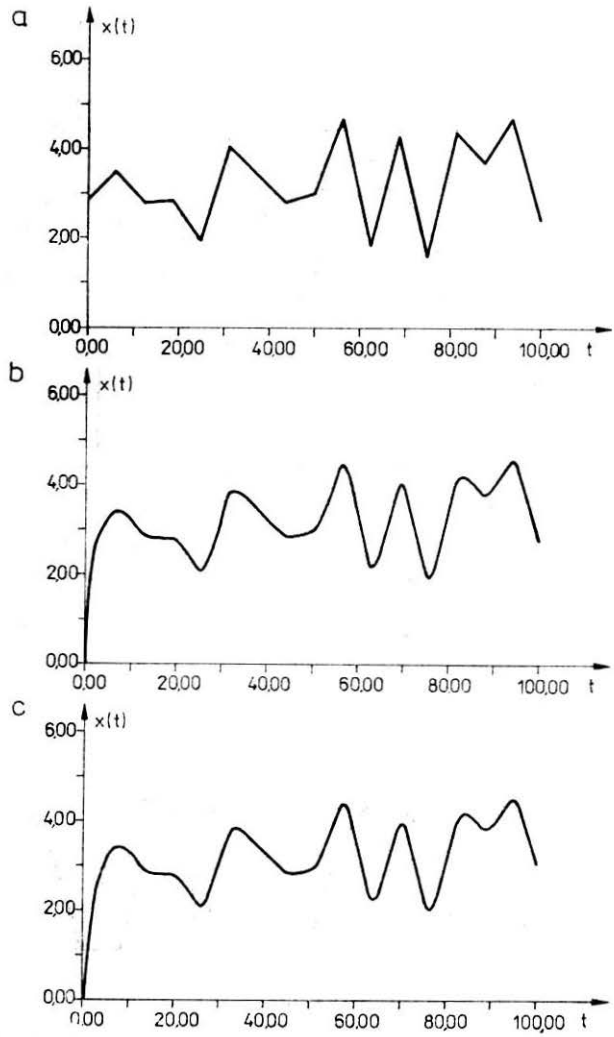



4-9. Filtr dolnoprzepustowy (charakterystyka dana linią przerywaną) powoduje wycięcie jednej części periodycznego widma sygnału próbkowanego i pozwala na jego odtworzenie w procesie przetwarzania cyfrowo-analogowego

kich $t \in [nT_p, (n+1)T_p]$). Takie założenie upraszcza obliczenia i konstrukcję przetwornika, ale powoduje, że odtworzony sygnał ma nieciągły przebieg, co w wielu przypadkach utrudnia jego prawidłowe wykorzystanie. Z tego powodu używa się także interpolatorów wyższych rzędów, które odtwarzają sygnał w przedziale $[nT_p, (n+1)T_p]$ aproksymując go liniowo między wartościami $x(nT_p)$ oraz $x((n+1)T_p)$, względnie wykorzystując aproksymację kwadratową, czyli tworząc łuki paraboli opartej na punktach $x((n-1)T_p)$, $x(nT_p)$ oraz $x((n+1)T_p)$. Na rysunku 4-10 pokazano przebieg czasowy odtworzony na podstawie sygnału poddanego próbkowaniu przy przyjęciu aproksymacji liniowej, kwadratowej oraz wielomianem trzeciego stopnia.

Widać, że przy aproksymacji parabolicznej dokładność odtworzenia sygnału jest zadowalająca i wyższe stopnie interpolatorów niewiele polepszają jakość odtworzonego sygnału.

Tak obszerna dyskusja sposobów odtwarzania wygnału analogowego z postaci cyfrowej wydaje się na pozór zbyteczna w kontekście głównego (i trud-




4-10. Przy przetwarzaniu cyfrowo-analogowym „wygładza" się odtworzony przebieg za pomocą interpolatorów. Rząd interpolatora wpływa na dokładność i gładkość odtworzonej krzywej, jednak główne korzyści odnosi się przy zastosowaniu interpolacji pierwszego rzędu (a), która zmienia rwaną, schodkową odpowiedź przetwornika na ciągły sygnał, oraz drugiego rzędu (aproksymacja z wykorzystaniem parabol) — (b). Wprowadzenie wyższego rzędu interpolacji (c) nie daje dalszych, zauważalnych korzyści

niejszego do praktycznej realizacji) zagadnienia przetwarzania sygnału analogowego na postać cyfrową. Tymczasem sposób wykorzystania sygnału cyfrowego po konwersji może decydować o pożądanym sposobie konwersji. Praktycznie nie jest możliwe do zrealizowania dokonanie sumowania wg wzoru (4.16) nieskończonej liczby składników. Zatem dokładne odtworzenie analogowego sygnału za pomocą sygnału cyfrowego jest niemożliwe. Zastępując we wzorze (4.16) nieskończoną sumę przez sumę skończonej liczby składników popełnia się błąd tym większy, im mniej składników

wlicza się do sumy. Zakładając, że przetwarzana funkcja ma postać $x(t) = 1$ można nawet podać analityczną zależność, określającą wielkość błędu $\Delta$ w funkcji liczby branych pod uwagę elementów sumy $M$. Zależność ta wyraża się wzorem:

$$\Delta(M) = 1 - \frac{4}{\pi} \sum_{k=0}^{M} \frac{(-1)^k}{2k+1} \tag{4.18}$$

z którego między innymi można wyliczyć niezbędną liczbę wyrazów sumy, koniecznych do zapewnienia założonej dokładności odtworzenia sygnału. Przykładowo dla dokładności $\Delta \leqslant 0{,}01$ konieczne jest uwzględnienie przynajmniej $M = 31$ próbek, a dla osiągnięcia $\Delta \leqslant 0{,}001$ konieczne jest użycie ponad dwustu składników sumy! Pomijając złożoność obliczeniową i koszt takich obliczeń, używanie wzoru (4.16) z dużą liczbą uwzględnianych składników wprowadza kolejny niekorzystny element — opóźnienie. Istotnie, posługując się wzorem (4.16) (w zmodyfikowanej postaci, uwzględniającej skończony zakres sumowania) musimy oczekiwać przynajmniej $M\,T_p$ sekund na odtworzenie prawidłowej wartości sygnału. Jest to w większości przypadków niedopuszczalna strata czasu — zbyteczna, kiedy rutynowo posługuje się wspomnianymi wyżej technikami interpolacji. Ale skoro nie zamierza się opierać na wzorze (4.16) w zakresie odtwarzania sygnału, to nie trzeba także brać pod uwagę wynikającej z niego zależności (4.17). Innymi słowy, decydując się na odtwarzanie sygnału z pewnym błędem można oprzeć się przy doborze częstości próbkowania na wielkości tego błędu (przy założonym sposobie interpolacji). Warto podkreślić, że otrzymana tą drogą częstotliwość próbkowania jest większa od wynikającej ze wzoru (4.17), co jest korzystne z punktu widzenia dokładności odtworzenia sygnału — i niekorzystne z punktu widzenia ilości informacji cyfrowych, które trzeba przetwarzać. Oszacowanie częstotliwości próbkowania na podstawie wielkości dopuszczalnego błędu $\Delta$ może być dane wzorem:

$$\Delta = T_p \sup \left| \frac{dx}{dt} \right| = T_p \omega_g [\sup |x(t)|] \tag{4.19}$$

To nowe oszacowanie także zależy od częstości granicznej $\omega_g$, ale łatwo się przekonać, że przy rozsądnych wymaganiach odnośnie dokładności $\Delta$ oszacowania $T_p$ są ostrzejsze niż wynikające ze wzoru (4.17).

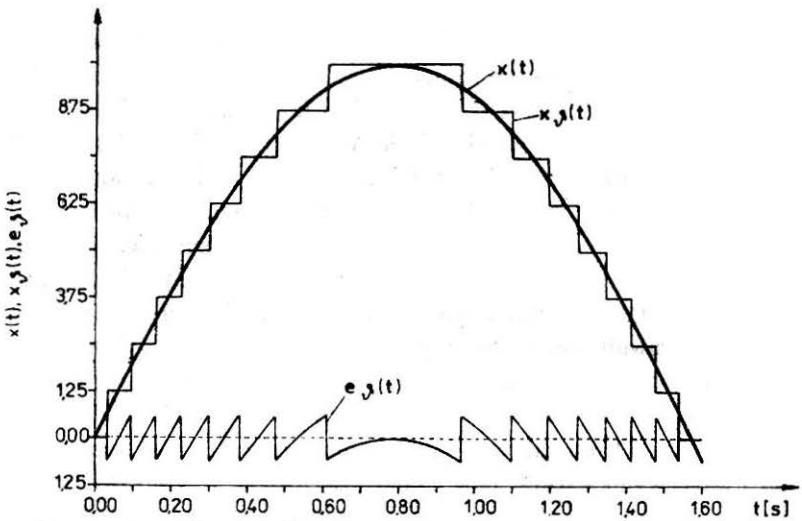
Dyskretyzacja amplitudy sygnału (tzw. kwantyzacja) wprowadza naturalnie błędy, związane z tym, że przebieg cyfrowy $x_v(t)$ przyjmuje jedynie ustalone dyskretne wartości, identyczne z wartościami przebiegu analogowego $x(t)$ jedynie w pewnych momentach czasu. W rezultacie błąd kwantowania $e_v(t)$, wyrażający się wzorem

$$e_v(t) = x(t) - x_v(t) \tag{4.20}$$

jest różny od zera w niemal wszystkich momentach czasu. Zilustrowano to na rys. 4-11.

Podstawowe znaczenie dla procesu przetwarzania analogowo-cyfrowego na omawianym tu etapie kwantowania amplitudy ma wybór liczby poziomów

kwantowania. Ponieważ amplituda szumu kwantowania $e_v(t)$ wynosi połowę wielkości odstępu pomiędzy sąsiednimi poziomami kwantowanego sygnału, wobec tego wybór większej liczby poziomów kwantowania gwarantuje większą dokładność i mniejsze szumy, a co za tym idzie — powinien być preferowany. Niestety wybór taki wiąże się z kosztem: im więcej wyróżnia się poziomów, tym więcej bitów będzie musiał mieć kod cyfrowy



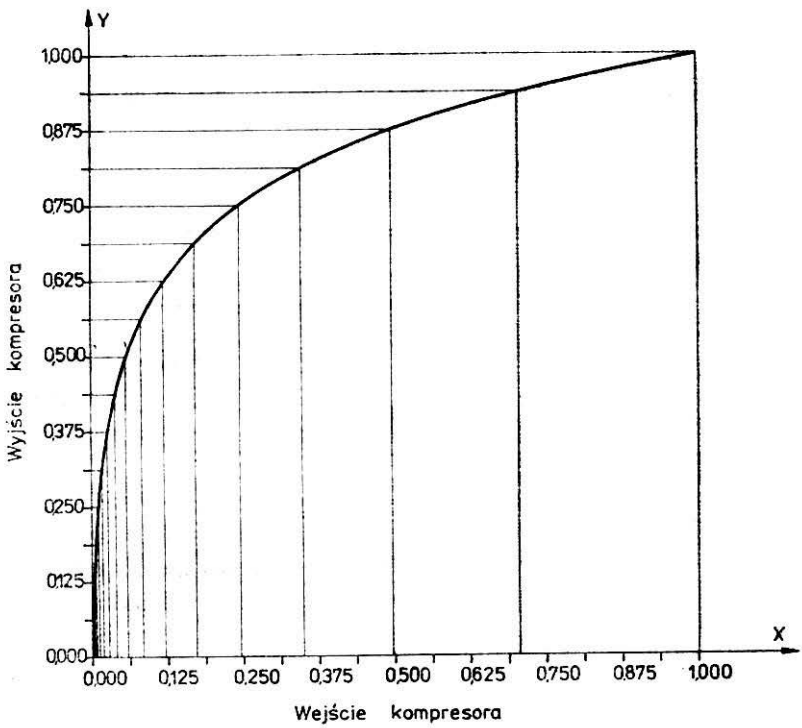

4-11. Kwantowanie amplitudy: Przebieg skwantowany $x_\vartheta(t)$ (połówka sinusoidy) różni się od przebiegu oryginalnego $x(t)$, a pokazany niżej przebieg błędu $e_\vartheta(t)$ nazywany jest zwykle szumem kwantyzacji. Wielkość tego szumu zależy głównie od kroku (przyrostu) kwantyzacji

reprezentujący dane w pamięci systemu przetwarzającego, a także odpowiednio większy i kosztowniejszy będzie przetwornik. Przyjmuje się niekiedy, że liczba bitów przetwornika $n$ powinna być związana z potrzebnym stosunkiem poziomu sygnału do szumu SNR wyrażonym w decybelach. Przybliżony wzór ujmujący tę zależność przytaczany jest często w postaci

$$n = \frac{\text{SNR}}{6} \tag{4.21}$$

Oznacza to, że przy wymaganym zakresie dynamiki wynoszącym 90 dB trzeba posłużyć się przetwornikiem 15-bitowym, wyróżniającym w próbkowanym sygnale 32 768 poziomów. Taki przetwornik (a raczej jego bardziej typowy odpowiednik 16-bitowy) jest preferowany z uwagi na fakt, że istnieje bardzo wiele gotowych systemów cyfrowych, pracujących przy długości słowa wynoszącej 16 bitów. Jest on jednak bardzo drogi. Znacznie tańszy jest przetwornik 12-bitowy, zapewniający dynamikę powyżej siedemdziesięciu decybeli (wyróżniający 4096 poziomów amplitudy sygnału). Dla celów technicznych 12 bitów okazuje się jednak nadal zbyt wysokim kosztem i w powszechnym użyciu (na przykład w telefonii cyfrowej) są przetworniki 8-bitowe. Nominalnie zapewniają one zaledwie niespełna 50 dB zakres dynamiki, w rzeczywistości jednak dokładność przetwarzania i jakość

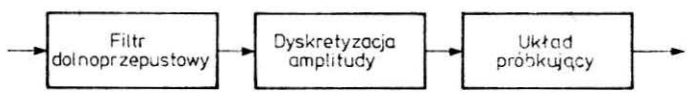
przetworzonego sygnału może być w tych przetwornikach nie gorsza niż w 12-bitowych, dzięki zastosowaniu omówionej kompresji amplitudy. Na rysunku 4-12 pokazano, że dzięki zastosowaniu kompresora amplitudy równomierne kwantowanie amplitudy sygnału na wyjściu kompresora odpowiada nierównomiernemu kwantowaniu sygnału wejściowego, rozumianemu w ten sposób, że małe wartości sygnału są kwantowane przy użyciu „gęściej" rozmieszczonych poziomów dyskretyzacji.




4-12. Charakterystyka typowego kompresora amplitudy. Widać, że równym odstępom poziomów kwantyzacji na wyjściu kompresora odpowiadają nierówne (rosnące ze wzrostem wartości $X$) przedziały wartości wejściowej. W ten sposób poziomy kwantyzacji mogą być rozłożone równomiernie (co upraszcza budowę przetwornika analogowo-cyfrowego), a przedziały dyskretyzacji są korzystnie zagęszczone dla małych wartości wejściowego sygnału, co polepsza stosunek sygnał/szum

Proces kwantowania amplitudy sygnału mowy może być także źródłem subtelnych błędów o dość nieoczekiwanym pochodzeniu. Jak wspomniano, proces kwantowania zamienia gładki przebieg funkcji $x(t)$ na schodkowy w kształcie przebieg $x_p(t)$ (por. rys. 4-11). Taki przebieg schodkowy zawiera wyższe harmoniczne, nieobecne w oryginalnym sygnale. Harmoniczne te przypadają na większe częstotliwości od uwzględnianej przy projektowaniu systemu próbkującego czasowo sygnał $f_g$ i mogą ulegać zdudnieniu z częstotliwością próbkowania $f_p$. W konsekwencji powstawać mogą składowe dodatkowe o częstotliwościach mieszczących się w granicach $0 \div f_g$, zniekształcające przetwarzany sygnał. Błąd ten, zwany szumem granulacji albo „ćwierkaniem", bywa bardzo uciążliwy, gdyż pojawia się szczególnie przy małych amplitudach przetwarzanego sygnału. Jest to logiczne: dla

słabych sygnałów różnica między funkcją schodkową a ciągłą jest bardziej istotna. Sposobem zwalczania omówionego zjawiska jest dodawanie do przetwarzanego sygnału składowych szumowych (tzw. *dither noise*) lub celowe wprowadzenie niestacjonarności pracy przetwornika. W przypadku sygnału mowy, kiedy użyty zakres amplitud sygnału wejściowego $x(t)$ będzie przekraczać rozpiętość poziomów przetwarzania, to sygnał ulegnie „przycięciu" i jego charakterystyka wzbogaci się o wyższe częstotliwości, przekraczające przyjętą wartość $f_g$ — ze wszystkimi wyżej omówionymi konsekwencjami. Zjawisko takie zagraża szczególnie wtedy, gdy filtr dolnoprzepustowy, ograniczający pasmo do założonej wartości $f_g$, poprzedza w przetworniku (rys. 4-13) układ kwantujący amplitudę sygnału. Zakres przetwarzanych



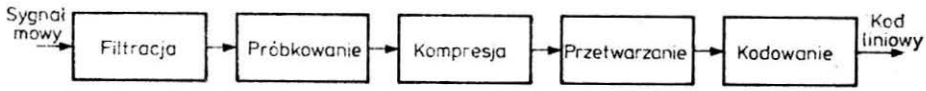

4-13. Układ przetwornika, w którym może zajść zniekształcenie sygnału na skutek dyskryminacji amplitudy sygnału filtrowanego

amplitud powinien być wówczas przynajmniej o 3 dB większy od wartości szczytowych sygnału wejściowego, gdyż proces filtracji dolnoprzepustowej na przykład fali prostokątnej dostarcza przebiegu o wartości szczytowej większej niż sygnał wejściowy.

Po przetworzeniu amplitudy sygnału na postać dyskretną z użyciem omówionych wyżej metod kwantowania oraz po jego próbkowaniu w dziedzinie czasu sygnał ma postać cyfrową i może być przetwarzany metodami cyfrowymi — nie przestając być sygnałem w dziedzinie czasu. Przed jakimkolwiek przetworzeniem sygnał musi, na ogół, być zakodowany. Przy cyfrowym przetwarzaniu sygnału mowy można stosować dowolne formy kodowania cyfrowej postaci sygnału, przy czym najczęściej wykorzystuje się prosty kod binarny, w którym poszczególnym wartościom skwantowanego sygnału odpowiadają wprost liczby dwójkowe, wynikające z zamiany odpowiedniej wartości z systemu dziesiętnego na dwójkowy. Typowym zabiegiem, jaki się tu stosuje dla wygody reprezentacji odpowiednich wartości, jest „przesunięcie" sygnału w dziedzinie amplitud o wartość odpowiadającą maksymalnej amplitudzie sygnału. Stosuje się to w celu uniknięcia konieczności kodowania liczb ujemnych. W rezultacie sygnał przyjmujący (po skwantowaniu) wartości od $-127$ do $+128$ zamieniony zostaje na sygnał zmieniający się w granicach od 0 do 255 (dla 8-bitowego przetwornika).

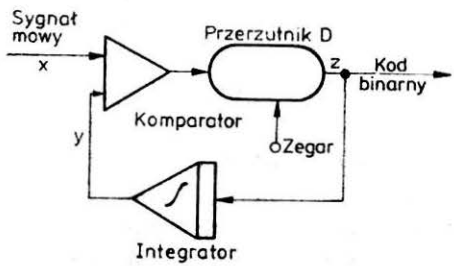
Nieco bardziej złożona sytuacja występuje w przypadku wykorzystywania cyfrowej postaci sygnału mowy w telekomunikacji. O ile bowiem badacz lub automatyk wykorzystujący sygnał mowy w swoim komputerze ma pełną swobodę wyboru sposobu kodowania, o tyle inżynier łączności musi tak przedstawiać sygnał w nadajniku, aby był on jednoznacznie łatwo interpretowany w odbiorniku. Wynika z tego między innymi konieczność podporządkowania się ścisłym wymaganiom norm międzynarodowych, a to oznacza, że z niezliczonej mnogości różnych form przedstawienia sygnału wy-

brane zostają niektóre — i tylko te mogą być stosowane. W przypadku telekomunikacji wykorzystywane są więc dwie jedynie formy kodowania sygnału: modulacja delta DM i modulacja impulsowo-kodowa PCM. Używane także „mieszane" techniki: adaptacyjna modulacja delta ADM i różnicowa modulacja impulsowo-kodowa DPCM mają mniejsze znaczenie. Modulacja impulsowo-kodowa jest w kontekście wszystkiego, co powiedziano wyżej, prostsza do opisania, choć bardziej złożona w realizacji. Nadajnik w systemie PCM dokonuje (rys. 4-14) filtracji dolnoprzepustowej sygnału




4-14. Struktura nadajnika sygnału w systemie PCM

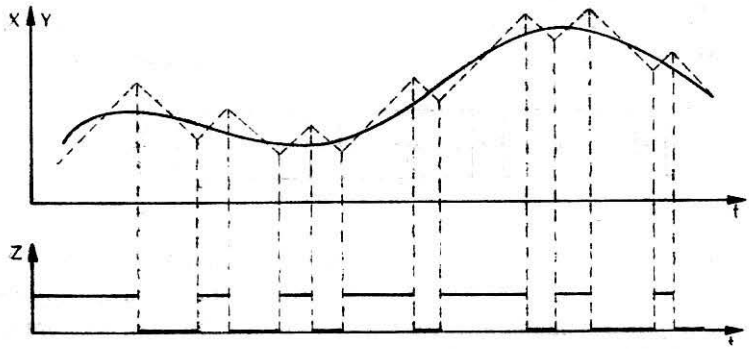
mowy (najczęściej stosuje się filtr o częstotliwości granicznej 3400 Hz), próbkowania sygnału (zwykle z częstością 8 kHz, co oznacza, że odstęp zabezpieczający przed nakładaniem się widm wynosi 1200 Hz), kompresji amplitud sygnału (według prawa $A$ lub $\mu$) oraz przetwarzania analogowo-cyfrowego (zwykle wyróżnia się 256 poziomów sygnału i wyraża się je za pomocą 8-bitowego słowa binarnego, co w połączeniu ze wspomnianą kompresją amplitud daje w przybliżeniu ten sam efekt, jak przetwarzanie 12-bitowe). Kod binarny na wyjściu przetwornika jest przekształcany na kod liniowy o własnościach zależnych od kanału, w którym będzie dokonywana transmisja sygnału. Kody takie mogą mieć w ogólnym przypadku złożony charakter, w szczególności mogą być zabezpieczane przed zniekształceniem przekazywanej wiadomości odpowiednimi bitami nadmiarowymi, co pozwala zarówno wykrywać ewentualne błędy transmisji, a także w niektórych przypadkach poprawiać zniekształcone kody. Problematyka kodów redundancyjnych i sposobów transmisji sygnałów daleko wykracza jednak poza zamierzone ramy tej książki i dlatego będzie tu pominięta.




4-15. Przetwornik (nadajnik sygnału cyfrowego) w układzie z modulacją delta

Alternatywny sposób kodowania, wspomniana już modulacja delta, stosowany jest głównie ze względu na prostą realizację układową. W nadajniku sygnału (rys. 4-15) pracującym przy użyciu tej metody modulacji nie występują de facto wspomniane wyżej procesy próbkowania, kwantyzacji i kodowania, a i odbiornik jest tu bardzo uproszczony. Zadanie nadajnika sprowadza się bowiem do porównywania liniowo rosnącej lub malejącej ze stałym nachyleniem (rys. 4-16) aproksymacji sygnału z jego rzeczywistą wartością. Na wyjściu nadajnika pojawia się przy tym sygnał 1, jeśli rzeczy-

wista wartość sygnału jest większa od aproksymowanej i 0, jeśli wartość aproksymowana przewyższa rzeczywistą. Realizacja układowa tej zasady działania wymaga posiadania komparatora, przerzutnika i integratora, a odtwarzanie sygnału — samego integratora (por. rys. 4-15), jest więc, ze sprzętowego punktu widzenia bez porównania tańsza niż PCM. Metoda




4-16. Przebiegi wybranych sygnałów czasowych w przetworniku z rys. 4-17. Istota działania modulatora delta polega, jak wynika z analizy przebiegów, na porównaniu aktualnej wartości sygnału z jego aproksymacją za pomocą sygnału z integratora (narastającego lub opadającego w czasie ze stałym nachyleniem). Gdy sygnał wejściowy ma wartość większą niż aproksymowany wysyłany jest sygnał $z = 1$, w przeciwnym przypadku $z = 0$. Pewna strefa nieczułości komparatora porównującego sygnały $X$ i $Y$ jest korzystna, gdyż zmniejsza częstotliwość wysyłania sygnałów $z$ do linii — chociaż odbywa się to kosztem zwiększonego „myszkowania" sygnału

modulacji delta zapewnia też dużą wartość stosunku sygnału do szumu (bez trudu osiąga się SNR = 65 dB). Na tym jednak jej zalety się kończą. Do niewątpliwych wad modulacji delta należą wprowadzane przez nią zakłócenia — dwojakiego rodzaju, w zależności od charakterystyki przetwarzanego sygnału. Pierwsza ewentualność pogorszenia jakości sygnału występuje dla sygnałów o dużych amplitudach i wysokich częstotliwościach. Przetwornik nie nadąża wówczas za zmianami sygnału i rozbieżność pomiędzy sygnałem rzeczywistym a aproksymowanym może wówczas osiągać duże wartości. Błąd ten, jakkolwiek łatwo zauważalny przy porównywaniu przebiegów czasowych sygnału oryginalnego i aproksymowanego, jest maskowany przez dużą energię sygnału i może być uznany za mniej uciążliwy niż błąd przeciwny, polegający na pojawianiu się oscylacji sygnału wokół wartości rzeczywistej dla sygnałów mniejszych niż pojedynczy dodatni lub ujemny przyrost sygnału aproksymującego. Obok wymienionych szumów kwantyzacji, towarzyszących modulacji delta, ma ona dodatkowo tę niekorzystną własność, że wymaga na ogół znacznie większej szybkości przesyłania bitów niż w systemach PCM o tej samej jakości transmisji.

## 4.2. Opis sygnału mowy w dziedzinie częstotliwości

Opis sygnału w dziedzinie częstotliwości jest podstawową, rutynowo stosowaną i w istocie najbardziej przydatną formą jego opisu. Użyteczność widmowej prezentacji sygnału — w szczególności w odniesieniu do sygnału

mowy — wynika z kilku faktów, których wyliczenie ułatwi skoncentrowanie uwagi przy śledzeniu dalszego tekstu na zagadnieniach najważniejszych dla całości problemu. Początkowe przesłanki wiążą się z wiadomościami podanymi w rozdz. 2 i 3. Jak łatwo było zauważyć, wynika z nich między innymi fakt, że w procesie artykulacji mowy kształtowana jest głównie obwiednia amplitudowo-częstotliwościowa sygnału (poprzez odpowiednio formowaną strukturę rezonansową traktu głosowego), w procesie percepcji zaś przed etapem analizy sygnału w sieciach nerwowych mózgu następuje etap wydzielenia składowych o poszczególnych częstotliwościach przez wyspecjalizowane struktury ucha wewnętrznego (błona podstawna, komórki rzęskowe, spiro- i ortoneurony zwoju spiralnego). Zatem biologiczny nadajnik formuje, a biologiczny odbiornik analizuje — głównie widmo sygnału. Fakt ten przemawia — obok innych, przytoczonych dalej argumentów — za stosowaniem również metod częstotliwościowych do analizy sygnału mowy.

Analiza widmowa może być realizowana wieloma metodami i może służyć do różnych celów; w rozdziale tym będziemy w stanie przedyskutować jedynie niektóre spośród możliwych metod i wskazać kilka bardziej typowych celów analizy widmowej sygnału mowy. Obszerniejsze omówienia poruszonych tu tematów można znaleźć w literaturze wymienionej na końcu książki.

W odniesieniu do wielu sygnałów, wśród których jest też i sygnał mowy, prawdziwe jest twierdzenie, iż świadomie kształtowane składowe sygnału mieszczą się głównie w jego amplitudowo-częstotliwościowej charakterystyce, podczas gdy wpływ czynników losowych determinuje w pierwszym rzędzie strukturę charakterystyk fazowo-częstotliwościowych. Rozważając sygnał w dziedzinie czasu stwierdza się równoczesny wpływ zarówno stosunków amplitudowych, jak i fazowych, na wypadkowy przebieg sygnału. Po dokonaniu analizy widmowej rozdzielenie wymienionych składników staje się banalnie proste. Podobnie, choć przy użyciu nieco bardziej złożonych metod, analiza częstotliwościowa pozwala rozróżnić te własności sygnału, za które odpowiedzialne jest źródło tonu, od tych, które są wynikiem procesu modulacji sygnału w narządach mowy. Przydatność takiej analizy do rozpoznawania mowy jest w świetle wszystkich wcześniej przedstawionych rozważań bezdyskusyjna, zaś samą technikę, wykorzystującą pojęcie tzw. cepstrum sygnału, omówimy szczegółowo nieco dalej.

Punktem wyjścia we wszystkich metodach wykorzystujących analizę spektralną jest para transformacji przekształcenia Fouriera:

— przekształcenie proste

$$G(f) = \mathscr{F}[g(t)] \tag{4.23}$$

— przekształcenie odwrotne

$$g(t) = \mathscr{F}^{-1}[G(f)] \tag{4.24}$$

We wzorach (4.23) i (4.24) funkcja $g(t)$ oznacza czasowy przebieg sygnału mowy, a $G(f)$ oznacza jego widmo. W odniesieniu do sygnału mowy można przyjąć, że $g(t)$ jest funkcją przyjmującą wartości rzeczywiste; wówczas

$G(f)$ jest funkcją przyjmującą wartości zespolone oraz jest to funkcja parzyście sprzężona:

$$G(f) = G^*(-f) \tag{4.25}$$

Gdzie $G^*$ oznacza liczbę zespoloną sprzężoną w stosunku do $G$. Wobec tego funkcję $G(f)$ można zapisać:

$$G(f) = |G(f)|e^{j \arg G(f)} \tag{4.26}$$

Wówczas moduł $|G(f)|$ odpowiada amplitudzie składowej o częstotliwości $f$ w wejściowym sygnale $g(t)$, zaś argument arg $G(f)$ jest kątem przesunięcia fazowego składowej o częstotliwości $f$ w sygnale $g(t)$ w chwili $t = 0$. Ważną dla dalszej analizy własnością przekształcenia Fouriera jest zachowywanie niezmienniczości energetycznej. Własność ta, wiązana w literaturze z nazwiskiem Persevala, umożliwia obliczanie mocy sygnału na dwa sposoby: w dziedzinie czasu: przez całkowanie kwadratu wartości sygnału w ustalonym przedziale czasu $T$:

$$P = \frac{1}{T}\int_0^T [\, g(t)]^2 dt \tag{4.27}$$

lub w dziedzinie częstotliwości — przez całkowanie charakterystyki sygnału w pełnym pasmie częstotliwości $F_d \div F_g$:

$$P = \int_{F_d}^{F_g} G(f) \cdot G^*(t) df \tag{4.28}$$

Postać wzorów (4.23) i (4.24) zależy od własności sygnału poddawanego analizie. Z punktu widzenia analizy sygnału mowy interesujące będzie rozważenie czterech przypadków szczególnych:

— sygnału okresowego, ciągłego — na przykład stany ustalone samogłosek;

— sygnału szumowego, ciągłego — na przykład spółgłoski szumowe;

— sygnału impulsowego (próbkowanego);

— sygnału dyskretnego (cyfrowego) zarówno w dziedzinie czasu, jak i w dziedzinie częstotliwości.

Dla sygnału okresowego i ciągłego o okresie wynoszącym $T$ widmo $G(f)$ jest funkcją dyskretną, przyjmującą wartości różne od zera jedynie dla $f$ będących całkowitymi wielokrotnościami częstotliwości podstawowej $f_1 = \frac{1}{T}$. Oznaczając przez $f_k = k\, f_1 = \frac{k}{T}$ możemy w tym przypadku zależność (4.23) zapisać w postaci:

$$G(f_k) = \frac{1}{T}\int_{-T/2}^{T/2} g(t) e^{-j2\pi f_k t}\, dt \tag{4.29}$$

Przekształcenie odwrotne (4.24) ma w tym przypadku postać:

$$g(t) = \sum_{k=-\infty}^{\infty} G(f_k) e^{j2\pi f_k t} \tag{4.30}$$

i daje w wyniku ponownie ciągłą, okresową funkcję czasu. Warto zwrócić uwagę, że we wzorze (4.30), w rzeczywistych przypadkach, zawsze można wskazać taką częstość graniczną $F_g$, że dla wszystkich $f_k > F_g$ zachodzi $G(f_k) = 0$. Wobec czego nie jest konieczne sumowanie dla $k \to \infty$. Podobnie problem ujemnych wartości $k$ nie nastręcza trudności praktycznych w świetle zależności (4.25).

Sygnał, który nie jest periodyczny, w szczególności sygnał szumowy, może być traktowany jako sygnał o okresie zmierzającym do nieskończoności. W takim przypadku odstępy pomiędzy prążkami dyskretnego widma zmierzają do zera ($\lim_{T\to\infty} 1/T = f_k - f_{k-1} = 0$) i widmo staje się ciągłe. Przekształcenia (4.23) i (4.24) mają w tym przypadku postać

$$G(f) = \int_{-\infty}^{\infty} g(t)\mathrm{e}^{-\mathrm{j}2\pi ft}\mathrm{d}t \tag{4.31}$$

$$g(t) = \int_{-\infty}^{\infty} G(f)\mathrm{e}^{\mathrm{j}2\pi ft}\mathrm{d}f \tag{4.32}$$

i są — jeśli pominąć znak w wykładniku — identyczne, co upraszcza wiele rozważań praktycznych. Często przyjmuje się, że para wzorów (4.31) i (4.32) stanowi właściwą definicję przekształcenia Fouriera, pozostałe zaś wzory szczegółowe są przypadkami szczególnymi. W istocie przydatność wzorów (4.31) i (4.32) organicza się do rozważań formalnych z uwagi na nieskończone granice całek w nich występujących. W warunkach rzeczywistych nawet mając do czynienia z funkcjami ciągłymi zarówno w dziedzinie czasu, jak i w dziedzinie częstotliwości, zmuszeni jesteśmy stosować warianty przytoczonych wzorów o określonych granicach całkowania:

$$G(t) = \int_{-T}^{T} g(t)\mathrm{e}^{-\mathrm{j}2\pi ft}\mathrm{d}t \tag{4.33}$$

$$g(t) = \int_{-F}^{F} G(f)\mathrm{e}^{\mathrm{j}2\pi ft}\mathrm{d}f \tag{4.34}$$

co jest równoważne użyciu wzorów definicyjnych (4.31) i (4.32), w których funkcje podcałkowe $g(t)$ i $G(f)$ odpowiednio przemnożone zostały przez odpowiednie „funkcje okna": czasowe $h(t)$ i częstotliwościowe $H(f)$. Mają one tę własność, że przyjmują wartość zero poza przedziałem $(-T, T)$ lub $(-F, F)$ odpowiednio. Możemy więc zapisać:

$$\tilde{G}(f) = \int_{-\infty}^{\infty} g(t)\mathrm{e}^{-\mathrm{j}2\pi ft}h(t)\mathrm{d}t \tag{4.35}$$

$$\tilde{g}(t) = \int_{-\infty}^{\infty} G(f)\mathrm{e}^{\mathrm{j}2\pi ft}H(f)\mathrm{d}f \tag{4.36}$$

Przejście od przekształcenia (4.31), (4.32) do przekształcenia (4.33), (4.34) lub równoważnego mu przekształcenia (4.35), (4.36) nie odbywa się bez-

karnie. Można udowodnić, że mnożeniu w całkach (4.35) i (4.36) odpowiada operacja splotu po transformacji. W rezultacie funkcja $\tilde{G}(t)$ obliczona według wzoru (4.35) jest splotem w dziedzinie częstotliwości rzeczywistego widma $G(f)$ oraz widma funkcji okna $h(t)$. Dla funkcji okna w postaci:

$$h(t) = \begin{cases} 1 \text{ dla } -T \leqslant t \leqslant T \\ 0 \text{ poza wskazanym przedziałem czasu} \end{cases} \tag{4.37}$$

widmo ma postać

$$\tilde{h}(f) = T\frac{\sin(\pi fT)}{(\pi fT)} \tag{4.38}$$

W wyniku splotu tej funkcji z rzeczywistym widmem $G(f)$ jest uzyskiwany przebieg widma przybliżonego $\tilde{G}(f)$ zgodnie ze wzorem

$$\tilde{G}(f) = \int_{-\infty}^{\infty} G(f-q)\tilde{h}(q)\mathrm{d}q \tag{4.39}$$

Łatwo zauważyć, że w ogólnym przypadku funkcje $G(f)$ i $\tilde{G}(f)$ mogą się znacznie różnić, co ogranicza praktyczną stosowalność wzorów (4.33) i (4.34), jako aproksymacji zależności (4.31) i (4.32), a także całego omówionego podejścia. Wprawdzie dobierając odpowiednio funkcje okna $h(t)$ i $H(f)$ można wskazane niekorzystne wpływy minimalizować, stosując np. okno typu $\cos^2(x)$. Przytoczona dyskusja miała na celu uzasadnienie celowości rozważania przez nas licznych przypadków szczegółowych i praktycznych postaci transformacji dla tych przypadków, gdyż — jak wspomniano — z praktycznego punktu widzenia nieuzasadniony jest pogląd, iż są to jedynie przypadki szczególne transformacji (4.31) i (4.32).

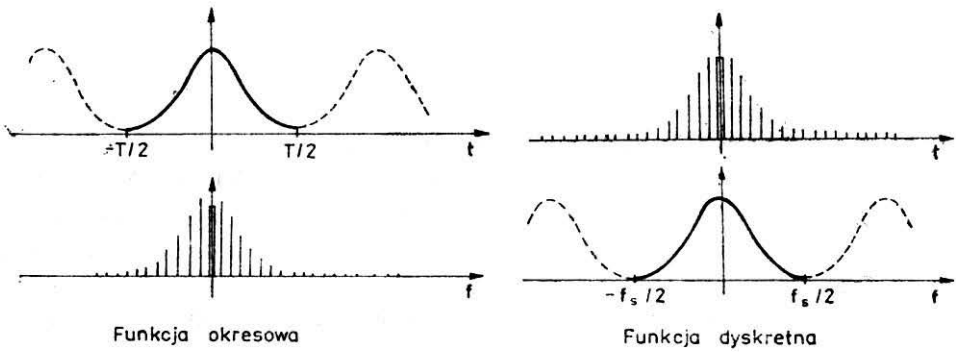
Kolejny szczegółowy rozważany przez nas przypadek dotyczy funkcji $g(t)$ próbkowanej (równomiernie) w dziedzinie czasu. Kolejne próbki czasowe sygnału $g(t)$ brane w odstępach czasu $\Delta t$ (czyli w momentach czasu $t_n = n\Delta t$) oznaczone być mogą przez $g(t_n)$ lub jeszcze prościej $g(n)$. Widmo takiej próbkowanej funkcji ma charakter okresowy — częstotliwość powtarzania wynosi $f_s = 1/\Delta t$. Warto zwrócić uwagę na charakterystyczną symetrię, jaka zachodzi między rozważanym tu przypadkiem a uprzednio dyskutowanym (por. wzory (4.29) i (4.30)) przypadkiem funkcji okresowej w dziedzinie czasu (rys. 4-17). Widmo funkcji okresowej ma charakter dyskretny, natomiast widmo funkcji dyskretnej ma charakter okresowy. Z własności okresowości widma funkcji dyskretnej (próbkowanej) korzystano już wcześniej przy dyskutowaniu twierdzenia o próbkowaniu przy przetwarzaniu analogowo-cyfrowym. Powracając do transformacji Fouriera funkcji próbkowanej możemy zapisać

$$G(f) = \sum_{n=-\infty}^{\infty} g(t_n)\mathrm{e}^{-\mathrm{j}2\pi f t_n} \tag{4.40}$$

oraz, wykorzystując okresowość widma,

$$g(t_n) = \frac{1}{f_s}\int_{-f_s/2}^{f_s/2} G(f)\mathrm{e}^{\mathrm{j}2\pi f t_n}\mathrm{d}f \tag{4.41}$$

Rozwój metod cyfrowej analizy sygnału mowy spowodował wzrost zainteresowania przypadkiem, kiedy zarówno w dziedzinie czasu, jak i w dziedzinie częstotliwości mamy do czynienia z funkcjami dyskretnymi. Dyskretna transformacja Fouriera przeprowadzana bywa zwykle z wykorzystaniem algorytmu FFT (szybkiej transformaty Fouriera, zaproponowanego w 1965 roku przez J. W. Cooleya i J. W. Tukeya). Dyskretne przekształcenie Fouriera, realizowane z zasady na ograniczonym zbiorze próbek w dziedzinie czasu i w takim samym co do liczebności zbiorze dyskretnych wartości




4-17. Porównanie dyskretności i okresowości w dziedzinie czasu i w dziedzinie amplitudy. Symetria prostego i odwrotnego przekształcenia Fouriera powoduje, że sygnał okresowy w dziedzinie czasu ma dyskretne widmo (harmoniczne), funkcja dyskretna zaś w dziedzinie czasu (próbkowany sygnał) ma widmo okresowe

częstotliwości w dziedzinie amplitud, zakłada okresowość zarówno funkcji czasu, jak i widma. Wynika to z dotychczasowych rozważań. Dyskretna funkcja czasu ma okresowe widmo, dyskretne widmo odpowiada zaś okresowej funkcji w dziedzinie czasu — wniosek jest więc oczywisty. Oznaczając wartości funkcji czasu w dyskretnych momentach $t_n$ przez $g(n)$ oraz oznaczając dyskretne wartości widma w punktach odpowiadających częstotliwości $f_k$ przez $G(k)$ możemy obecnie parę odwzorowań (4.23) i (4.24) zapisać w postaci:

$$G(k) = \frac{1}{N}\sum_{n=0}^{N-1} g(n)\mathrm{e}^{-\mathrm{j}\frac{2\pi kn}{N}} \tag{4.42}$$

$$g(n) = \frac{1}{N}\sum_{n=0}^{N-1} G(k)\mathrm{e}^{\mathrm{j}\frac{2\pi kn}{N}} \tag{4.43}$$

gdzie bardzo ważny parametr $N$ oznacza łączną liczbę dyskretnych odczytów funkcji czasu $g(n)$ lub liczbę dyskretnych prążków w dziedzinie częstotliwości.

W badaniach nad sygnałem mowy rozpatrywana jest zwykle charakterystyka amplitudowo-częstotliwościowa, czyli moduł $G(f)$. Jak wiadomo, formalnie można go wyznaczyć ze wzoru:

$$|G(f)| = \sqrt{G(f)G^*(f)} \tag{4.44}$$

Charakterystyka fazowa jest mniej przydatna i będzie rozważana dalej.

Przytaczane wzory, pozwalające wyznaczać $G(f)$, a następnie potrzebną charakterystykę amplitudowo-częstotliwościową, są bez wyjątku nieprzydatne praktycznie. Nawet pozornie łatwe do użycia przy wykorzystaniu techniki cyfrowej wzory (4.42) i (4.33) są niepraktyczne, gdyż wyliczenie widma za ich pomocą wymaga zastosowania $N^2$ mnożeń zmiennych zespolonych, podczas gdy wzmiankowany już algorytm FFT zakłada konieczność wykonania jedynie $N \log_2 N$ mnożeń, co oznacza (przy typowych zbiorach danych, gdzie $N > 1000$) ponad stukrotne przyspieszenie obliczeń.

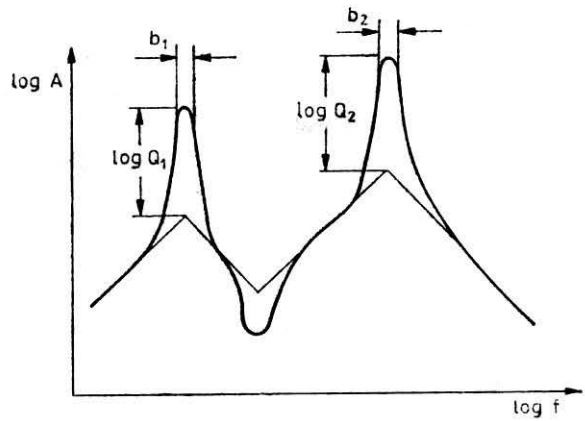
Praktyczne metody wyznaczania widma sygnału mowy mogą być więc zebrane w trzy grupy:

— wydzielanie składowych o różnych częstotliwościach za pomocą filtrów analogowych;
— wydzielanie pasm częstotliwości za pomocą filtrów cyfrowych;
— analiza cyfrowa z wykorzystaniem algorytmów FFT oraz Zoom-FFT.

Decydując się na analizę analogową musimy dokonać wyboru pasma stosowanego zestawu filtrów. Istnieją dwie możliwości: można wybrać pasma poszczególnych filtrów o jednakowej szerokości (otrzymując liniową skalę częstotliwości) lub określić stały stosunek szerokości pasma używanego filtru do jego częstotliwości środkowej (otrzymując logarytmiczną skalę częstotliwości). Skala logarytmiczna jest korzystniejsza ze względu na to, że lepiej koresponduje z naturalnymi własnościami słuchu człowieka, który jak wiadomo rozróżnia wysokości dźwięku (subiektywne odczucia częstotliwości), zgodnie z prawem Webera, a więc w sposób (w przybliżeniu) logarytmiczny. Ponadto, przy analizie obejmującej ponad 3 dekady stała procentowa szerokość filtru jest korzystniejsza ze względu na możliwość efektywnego prowadzenia pomiaru. W przeciwnym przypadku przyjęte stałe pasmo używanych filtrów albo będzie wymuszało użycie w następnej dekadzie ogromnej (liczącej setki pozycji) liczby filtrów, albo dokładność analizy w niższej dekadzie będzie niezadowalająca (cała dekada będzie pokryta przez jeden lub dwa filtry). Dalszymi zaletami skali logarytmicznej (wyznaczanej przez filtry o stałej procentowej szerokości) są: łatwe wykrywanie różnych zależności przy użyciu skali logarytmicznej oraz stałość dobroci $Q$ (rys. 4-18). Za przyjęciem takiej skali przemawia także tradycja metrologii akustycznej, która zazwyczaj opierała się na pomiarach wykonywanych filtrami o stałej procentowej szerokości. Natomiast filtry o stałej szerokości ułatwiają graficzną prezentację wyników przetwarzania sygnału (w szczególności wszystkie przytaczane rysunki wykonywane są z reguły w skali liniowej, czyli stałej szerokości pasm analizy), a także dobrze korespondują z nowoczesnymi, cyfrowymi metodami analizy, w których uzyskiwana skala częstotliwości jest z reguły liniowa.

Rozpatrując stałą procentową szerokość pasma filtrów analizujących widmo można wyróżnić filtry o szerokości pasma wynoszącej jedną oktawę, filtry 1/3-oktawowe oraz filtry o szerokości 1/10 oktawy. Filtry oktawowe mają pasmo o szerokości 70,7%, gdyż zgodnie z nazwą ich pasmo rozciąga się od pewnej ustalonej częstości dolnej $F_d$ do częstości górnej $F_g = 2 F_d$. Wyliczając częstotliwość środkową jako $F_0 = \sqrt{F_d F_g}$ oraz pasmo jako $\Delta F =$

$= F_g - F_d$ bez trudu wyznaczamy również stosunek $\Delta F/F_0 = 1/\sqrt{2} =$ $= 70{,}7\%$. Zestaw filtrów oktawowych konstruowany jest zazwyczaj przy ustaleniu centralnej częstości wynoszącej 1000 Hz i obejmuje typowo 10 filtrów pokrywających łącznie 3 dekady: od 22,5 Hz (dolna graniczna częstotliwość pierwszego filtru, którego częstotliwość środkowa wynosi 31,5 Hz) do 22,5 kHz (górna graniczna częstotliwość filtru o częstotliwości środkowej 16 kHz).



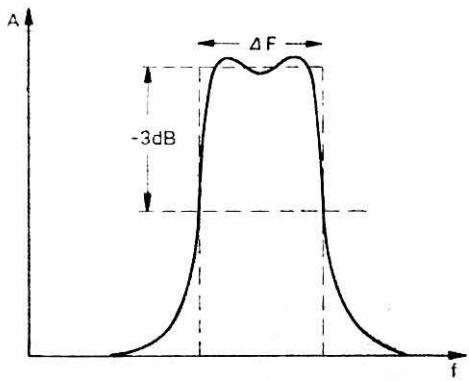

4-18. Zaletą logarytmicznej prezentacji osi częstotliwości sygnału jest między innymi łatwość określania związków pomiędzy wysokością i szerokością krzywych rezonansowych: $b_1 = 1/Q_1$ oraz $b_2 = 1/Q_2$

Do dokładniejszej analizy wykorzystywane są filtry o szerokości 1/3 oktawy (dziesięć filtrów na dekadę). Przy budowie takich filtrów bierze się pod uwagę zależność górnej częstości granicznej $F_g$ i częstości dolnej $F_d$ zgodnie ze wzorem $F_g = 2^{1/3} F_d$, skąd łatwo wyliczyć, że $\Delta F/F_0 = 23{,}1\%$. Filtry o mniejszych szerokościach pasma, na przykład wspomniane filtry o szerokości 1/10 oktawy, pozwalają na dokładniejsze rozróżnianie drobnych szczegółów w widmie, jednak ich użycie związane jest z długim czasem ustalania się odpowiedzi na wyjściu filtru i dlatego ich użyteczność w analizie sygnału mowy jest ograniczona. Są one natomiast przydatne do analizy sygnałów, które pozostają nie zmienione przez długi czas i mogą być analizowane przez zestaw filtrów o dużej rozdzielczości. W celu pokonania tej niedogodności stosuje się niekiedy specjalne metody, na przykład specjalną kompresję czasową w analizatorze typu 3348 firmy Brüel and Kjaer, jednak problem jako taki pozostaje i jest jednym z trudniejszych w problematyce analizy spektralnej. Przyjmując za podstawę dyskusji przybliżoną relację

$$bn = 1 \tag{4.45}$$

w której $b$ oznacza procentową szerokość pasma, a $n$ — liczbę okresów fali dźwiękowej, niezbędnych do ustalenia się przebiegów na wyjściu filtru, możemy łatwo stwierdzić, że dla filtru oktawowego ($b = 0{,}707$) przebiegi ustalają się praktycznie po pojedynczym okresie fali, dla filtru 1/3-oktawowego wymaganych jest 5 okresów, a dla filtru o $b = 0{,}01$ konieczne jest odczekanie aż 100 okresów — podczas gdy w sygnale mowy nader rzadko występują odcinki o tak długim czasie trwania ustalonego przebiegu. Warto

przy tym ustosunkować się krótko do stosowanej w praktyce metody „wydłużania" krótkotrwałych odcinków sygnału mowy przez nagrywanie ich na taśmę i odtwarzanie w pętli wielokrotnie podczas analizy. Metoda taka, obok niedogodności związanej z wydłużeniem czasu trwania analizy, której wyniki nie mogą niestety być wykorzystywane na bieżąco, ma dodatkową niedogodność. Wynika ona z faktu, że w analizowanym przebiegu pojawią się składniki wywołane sztucznie wprowadzoną okresowością sygnału, które mogą zniekształcać rzeczywisty obraz analizowanych zjawisk. Wykorzystując „w pętli" sygnał o czasie trwania $T$ wprowadzamy sztucznie do analizowanego widma prążki o częstościach $1/T$, $2/T$, ... co bywa zaniedbywane przy opracowywaniu wyników pomiarów i powoduje istotne zniekształcenie widma głosek krótkotrwałych.

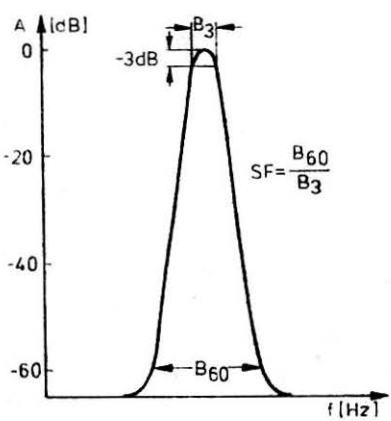



4-19. Szerokość pasma filtru może być określana jako szerokość szumowa (szerokość pasma filtru, który wycina, z białego szumu sygnał o tej samej energii, a ma prostokątną charakterystykę — patrz linia przerywana na rysunku), względnie częściej — jako szerokość rzeczywistej charakterystyki filtru na poziomie — 3 dB w stosunku do wierzchołka charakterystyki. Zwykle obie wartości są praktycznie identyczne, a druga jest łatwiejsza do określenia

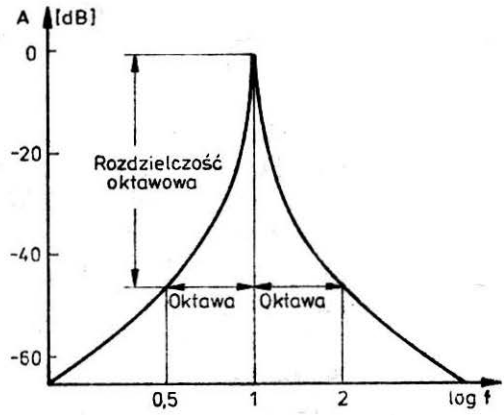
Rozważając filtry analogowe jako urządzenia do uzyskania informacji o widmie sygnału mowy należy przeanalizować wybrane charakterystyki tych filtrów i wprowadzić kryteria pozwalające określać ich jakość. Podstawowym parametrem filtru pasmowego nastrojonego na określoną częstotliwość $F_0$ jest jego pasmo przepustowe $\Delta F$ (rys. 4-19). Definiować można różne pasma, przy czym najczęściej są używane dwie definicje. Pierwsza określa tak zwane p a s m o s z u m o w e jako szerokość pasma idealnego filtru (tzn. mającego prostokątną charakterystykę amplitudowo-częstotliwościową) o identycznej częstotliwości środkowej, który wydobywa tę samą moc z sygnału będącego białym szumem, co rozważany filtr rzeczywisty. Tłumacząc to na język praktyczny można powiedzieć, że wspomniana definicja określa pasmo jako szerokość prostokąta mającego tę samą wysokość i tę samą powierzchnię, co charakterystyka amplitudowo-częstotliwościowa rozważanego filtru (por. rys. 4-19). Z samego opisu przedstawionej definicji można wywnioskować, że jest ona mało praktyczna. Istotnie,

określenie dla rzeczywistego filtru jego pasma we wspomniany sposób jest pracochłonne i niewygodne. Z tego powodu używa się drugiej definicji, mającej nieco arbitralny charakter (wybór 3 dB ma charakter umowny — patrz dalej), ale bardzo wygodnej i łatwej w stosowaniu, a ponadto — co bardzo ważne — dającej dla większości praktycznie realizowanych filtrów prawie identyczne wartości szerokości pasma, jak wspomniana wyżej „szumowa" definicja. Ustala się mianowicie szerokość pasma jako szerokość charakterystyki amplitudowo-częstotliwościowej filtru na poziomie −3 dB w stosunku do wysokości wierzchołka obwiedni. Definicja ta jest tak wygodna i rozpowszechniona, że określenie „szerokości pasma" podane bez dodatkowych wyjaśnień zawsze odnosi się do tak właśnie zdefiniowanej szerokości. W definicji tej, pomimo jej prostoty, kryje się pewna niejednoznaczność, wynikająca z faktu, że charakterystyka rzeczywistego filtru nigdy nie ma idealnie płaskiego wierzchołka, lecz pojawiają się na niej zafalowania. Punkt, od którego odmierza się trzydecybelowy odstęp jest więc w pewnym stopniu umowny. Zafalowania obwiedni dla dobrych filtrów powinny być minimalne, a przebieg rozważanej charakterystyki wyznaczany jest na drodze pomiarowej z ograniczoną dokładnością, zatem wpływ wspomnianej arbitralności na końcowy rezultat i jego precyzję może być uznany za mało znaczący.

Oprócz częstotliwości środkowej i szerokości pasma (dowolnie rozumianego) do opisu własności filtru potrzebna jest dodatkowo ocena stromości zboczy jego charakterystyki. Używane są w tym celu dwa parametry. Pierwszy z nich, używany głównie do opisu charakterystyki filtrów liniowo rozłożonych w skali częstotliwości, nazywany bywa współczynnikiem kształtu (ang. *shape factor*) i definiowany jest jako stosunek szerokości pasma przepustowego na poziomie −60 dB do szerokości pasma na poziomie −3 dB, przyjętego jako podstawa określenia szerokości filtru (rys. 4-20). Drugi parametr nazywany jest rozdzielczością (selektyw-




4-20. Określenie współczynnika kształtu filtru (SF), będącego stosunkiem szerokości charakterystyki na poziomie −60 dB w stosunku do wierzchołka $B_{60}$ do szerokości charakterystyki na poziomie −3 dB —$B_3$

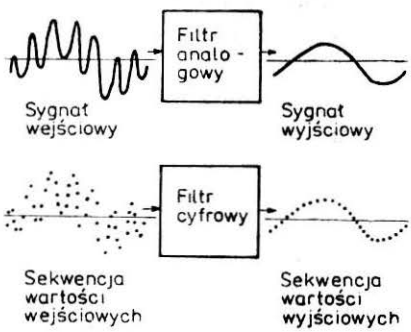



4-21. Stosując logarytmiczną skalę częstotliwości można zdefiniować jakość filtru podając jego rozdzielczość oktawową, czyli stopień tłumienia sygnałów o częstotliwości większej lub mniejszej od $f_0$ filtru o jedną oktawę

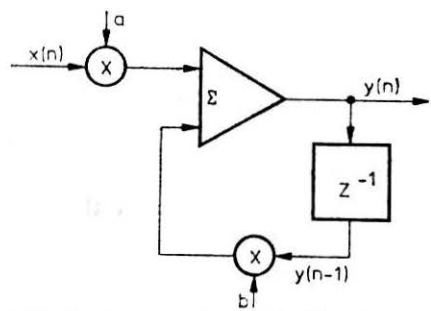
nością) oktawową i określany jest jako wielkość tłumienia filtru dla częstotliwości odległych o jedną oktawę od częstotliwości środkowej filtru (rys. 4-21). Oczywiście, drugi z wymienionych parametrów odnosi się głównie do filtrów tworzących zestawy o stałej (względnej) szerokości pasma, czyli filtrów nawiązujących do logarytmicznej skali częstotliwości.

Jak wspomniano uprzednio, obok wciąż popularnych i chętnie stosowanych w praktyce filtrów analogowych, służących do wydzielania różnych częstotliwości z wejściowego sygnału, coraz popularniejsze stają się filtry cyfrowe i technika szybkiej transformaty Fouriera. Sygnał wejściowy w obydwu wymienionych technikach musi być poddany przekształceniu do postaci cyfrowej, co generalnie utrudnia użycie tych technik, jednak liczne zalety metod cyfrowej analizy częstotliwościowej przeważają nad uciążliwością wstępnego przetwarzania sygnału i technika cyfrowa w coraz szerszym zakresie wykorzystywana jest także w dziedzinie analizy częstotliwościowej sygnału mowy.

Technika filtracji cyfrowej, bo od niej rozpoczniemy dyskusję, stanowi koncepcyjnie kontynuację filtracji analogowej (por. schemat na rys. 4-22),




4-22. Zasada działania filtru cyfrowego (u dołu) jest identyczna, jak w przypadku filtru analogowego (u góry), jednak filtr cyfrowy działa na dyskretnych (próbkowanych) wartościach sygnału wejściowego




4-23. Struktura prostego, jednobiegunowego filtru cyfrowego. Oznaczenie ⊗ używane jest dla operacji mnożenia, Σ dla dodawania, zaś $z^{-1}$ dla opóźnienia obiegu sygnału o jeden takt. Oznaczenia te będą używane w dalszych rysunkach bez objaśniania

z tą jednak różnicą, że sygnał przed filtracją musi być doprowadzony do postaci cyfrowej, a po filtracji ma także postać cyfrową — tylko jego skład spektralny uległ odpowiedniej modyfikacji, zależnej od własnego użytego filtru. Przefiltrowany sygnał cyfrowy może być przesłany wprost do dalszych, cyfrowych z reguły, systemów analizujących, przetwarzających lub przesyłających sygnał na duże odległości. Może być też oczywiście, za pomocą przetwornika cyfrowo-analogowego, przekształcony na powrót do postaci analogowej i w tej formie wykorzystany. Współczesna technika cyfrowa stawia do dyspozycji projektanta i wykonawcy filtru cyfrowego wyjątkowo bogaty zestaw możliwości technicznych i teoretycznych. Z jednej strony bowiem znane metody syntezy filtrów cyfrowych (rekursywnych i nierekursywnych) pozwalają projektować filtry o całkowicie dowolnie wybieranych charakterystykach. W szczególności możliwe jest projektowa-

nie filtrów dolno-, górno- i środkowoprzepustowych; typu Czebyszewa, Butterwortha i dowolnego innego — przy czym w odróżnieniu od realizacji analogowej na parametry tych filtrów i ich charakterystyki nie nakłada się praktycznie żadnych ograniczeń. Z drugiej strony powstają wciąż nowe, doskonalsze i coraz tańsze podzespoły cyfrowe o rosnącym stopniu scalenia, co ułatwia wykonanie zaprojektowanych układów cyfrowych i ich miniaturyzację.

Zasadę działania filtru cyfrowego wygodnie jest zilustrować na przykładzie prostego jednobiegunowego filtru dolnoprzepustowego (rys. 4-23). Widać, że zasada działania filtru polega na wyliczaniu sygnału wyjściowego w kolejnych momentach czasu $y(n)$ jako sumy ważonej sygnału wejściowego w danej chwili czasowej $x(n)$ oraz sygnału wyjściowego w poprzedniej) chwili $y(n-1)$

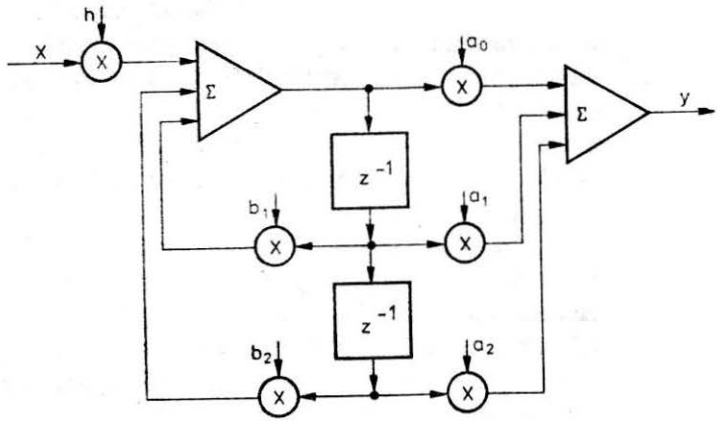
$$y(n) = ax(n)+by(n-1) \tag{4.46}$$

Współczynniki $a$ i $b$ określają własności filtru. W szczególności dla $a = 0{,}1$ oraz $b = 0{,}9$ powstaje filtr o zastępczej stałej czasowej wynoszącej 10 okresów próbkowania wejściowego sygnału, co jest równoważne przedziałowi uśredniania wynoszącemu 20 okresów. Filtr taki bardzo efektywnie wygładza wejściowy sygnał. Na przykład, w skrajnie niekorzystnym przypadku filtrowania sygnału sinusoidalnego o częstotliwości będącej połową częstości próbkowania fluktuacje sygnału wyjściowego nie przekraczają 0,3 dB. Inne parametry filtru można natychmiast uzyskać zmieniając wartości parametrów $a$ i $b$*). Warto te informacje uzupełnić jedynie tym, że oznaczenie $z^{-1}$, użyte na rys. 4-23 do oznaczenia opóźnienia sygnału o jeden takt, wiąże się z tzw. transformacją $\mathscr{Z}$ używaną do projektowania filtrów cyfrowych. Transformacja ta pełni dla układów cyfrowych analogiczną rolę, jak transformacja Fouriera lub Laplace'a dla układów ciągłych. Jej wzór definicyjny wiąże dyskretną (cyfrową) funkcję czasu $g(n)$ z funkcją argumentu zespolonego $G(z)$, przy czym argument zespolony $z$ związany jest z okresem próbkowania $\Delta t$ wzorem $z = \exp(j2\pi f\Delta t)$. Odwzorowanie $G(z) = \mathscr{Z}[g(n)]$ zapisywane jest zwykle w postaci

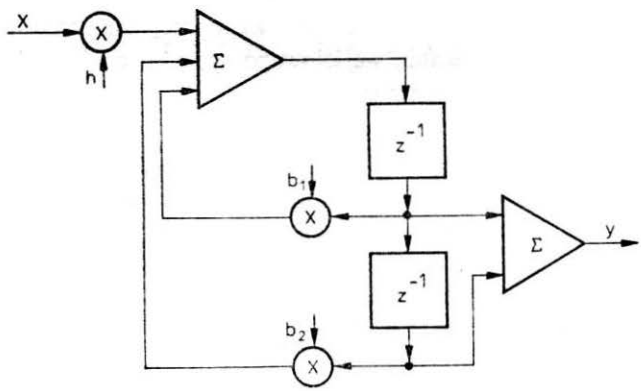
$$G(z) = \sum_{n=-\infty}^{\infty} g(n)z^{-n} \tag{4.47}$$

i pozwala wygodnie opisywać i projektować najrozmaitsze systemy cyfrowe, w tym także filtry cyfrowe. Filtry te najczęściej mają postać dwubiegunowych rekursywnych układów o ogólnej postaci podanej na rys. 4-24. Praktyczne realizacje rzeczywistych filtrów bywają uproszczone w stosunku do schematu podanego na rys. 4-24. Przykładowo, na rys. 4-25 podano schemat filtrów używanych w analizatorze 2131 firmy Brüel and Kjaer, przy czym przez odpowiedni dobór parametrów filtry te mogą być używane bądź jako dolnoprzepustowe filtry Buttervortha, bądź jako pasmowe filtry Czebyszewa o dowolnie dobieranej szerokości pasma: 1 oktawa, 1/3 oktawy lub 1/12 oktawy.

*) W istocie wybór ogranicza się do jednego parametru, gdyż zakłada się $a+b = 1$.




**4-24. Ogólna struktura filtru dwubiegunowego. Zależnie od doboru parametrów filtr ten może mieć rozmaite charakterystyki**




**4-25. Struktura filtrów używanych w analizatorze firmy Brüel and Kjaer. Filtry te pracują zarówno jako pasmowe, jak i jako dolnoprzepustowe**

Całkowicie odmienną, czysto cyfrową techniką uzyskiwania widma sygnału mowy jest algorytm FFT (szybkiej transformaty Fouriera) wspominany już wcześniej przy wprowadzaniu wzoru (4.42). Ze względu na istnienie bogatej i łatwo dostępnej literatury tego tematu nie wydaje się celowe dyskutowanie szczegółów w tej książce, warto jedynie odnotować kilka podstawowych własności FFT, istotnych z punktu widzenia analizy sygnału mowy. Najbardziej rozpowszechniona i chyba najwygodniejsza jest wersja algorytmu FFT, przy której liczba próbek czasowych sygnału $N$ (równa oczywiście liczbie wyliczonych pasm częstotliwości w widmie) jest potęgą liczby 2. Zakładając zatem $N = 2^m$ możemy ponumerować próbki czasowe i wydzielone pasma częstotliwości liczbami binarnymi $m$-bitowymi, co znajduje zastosowanie przy realizacji algorytmu. Aby go zilustrować, przyjmijmy przykładowo $N = 8$ ($m = 3$) i zapiszmy równanie (4.42) w postaci macierzowej. Zauważmy przy tym, że mnożniki zespolone $e^{-j\frac{2\pi kn}{N}}$ przyjmować będą wyłącznie wartości: $e^{-j0} = 1$, $e^{-j\pi/4}$ (co odpowiada obrotowi o kąt 45° i oznaczane będzie skrótowo przez $A$), $e^{-j\pi/2}$ (obrót o 90°, oznaczenie $B$), $e^{-j3\pi/4}$ (symbolicznie $C$), $e^{-j\pi} = -1$, $e^{-j5\pi/4}$ (symbolicznie $D$),

$e^{-j3\pi/2}$ (symbolicznie $E$) oraz $e^{-j7\pi/4}$ (symbolicznie $F$). Oczywiście — nie należy o tym zapominać, gdyż istotnie wpływa to na złożoność obliczeń — wartości oznaczone jako $A$, $B$, $C$, $D$, $E$ i $F$ są wielkościami zespolonymi, z wyjątkiem dwu przypadków szczególnych $e^{-j0}$ oraz $e^{-j\pi}$. Używając wprowadzonych oznaczeń wypiszemy teraz w formie macierzowej transformację (4.42) dla rozważanego przypadku $N = 8$. Na prawym marginesie wzoru, dla dalszej analizy, ponumerowano równanie kolejnymi liczbami binarnymi.

$$\begin{bmatrix} G(0) \\ G(1) \\ G(2) \\ G(3) \\ G(4) \\ G(5) \\ G(6) \\ G(7) \end{bmatrix} = \frac{1}{8} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & A & B & C & -1 & D & E & F \\ 1 & B & -1 & E & 1 & B & -1 & E \\ 1 & C & E & A & -1 & F & B & D \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & D & B & F & -1 & A & E & C \\ 1 & E & -1 & B & 1 & E & -1 & B \\ 1 & F & E & D & -1 & C & B & A \end{bmatrix} \begin{bmatrix} g(0) \\ g(1) \\ g(2) \\ g(3) \\ g(4) \\ g(5) \\ g(6) \\ g(7) \end{bmatrix} \begin{matrix} (000) \\ (001) \\ (010) \\ (011) \\ (100) \\ (101) \\ (110) \\ (111) \end{matrix} \tag{4.48}$$

Przedstawiony zapis macierzowy układu równań nie wykazuje niezbędnych symetrii prowadzących do uproszczeń skracających obliczenia i dlatego trzeba we wzorze (4.48) poprzestawiać wiersze, aby uwypuklić regularność budowy macierzy wiążącej ze sobą wartości zespolone $G(0)$, $G(1)$, ..., $G(7)$ oraz wartości próbek sygnału $g(0)$, $g(1)$, ..., $g(7)$. Przydatna przy tym będzie wprowadzona binarna numeracja równań. Otóż odczytując numery binarne od tyłu (w odwrotnej kolejności bitów) otrzymujemy nowe numery, wskazujące na sposób przestawienia odpowiednich wierszy. Warto zauważyć, że wiersze o numerach symetrycznych w układzie dwójkowym (0, 2, 5, 7) nie będą przestawiane. W wyniku przestawień powstaje macierz o wyraźnej regularności:

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & B & -1 & E & 1 & B & -1 & E \\ 1 & E & -1 & B & 1 & E & -1 & B \\ 1 & A & B & C & -1 & D & E & F \\ 1 & D & B & F & -1 & A & E & C \\ 1 & C & E & A & -1 & F & B & D \\ 1 & F & E & D & -1 & C & B & A \end{bmatrix} \tag{4.49}$$

Wykorzystując tę regularność rozkłada się macierz daną wzorem (4.49) na trzy macierze o budowie umożliwiającej zminimalizowanie liczby mnożeń. Można wykazać, że macierz opisana wzorem (4.49) jest iloczynem trzech następujących (kolejnych!) macierzy:

$$\begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & B & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & E & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & A & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & D & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & C \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & F \end{bmatrix} \tag{4.50}$$

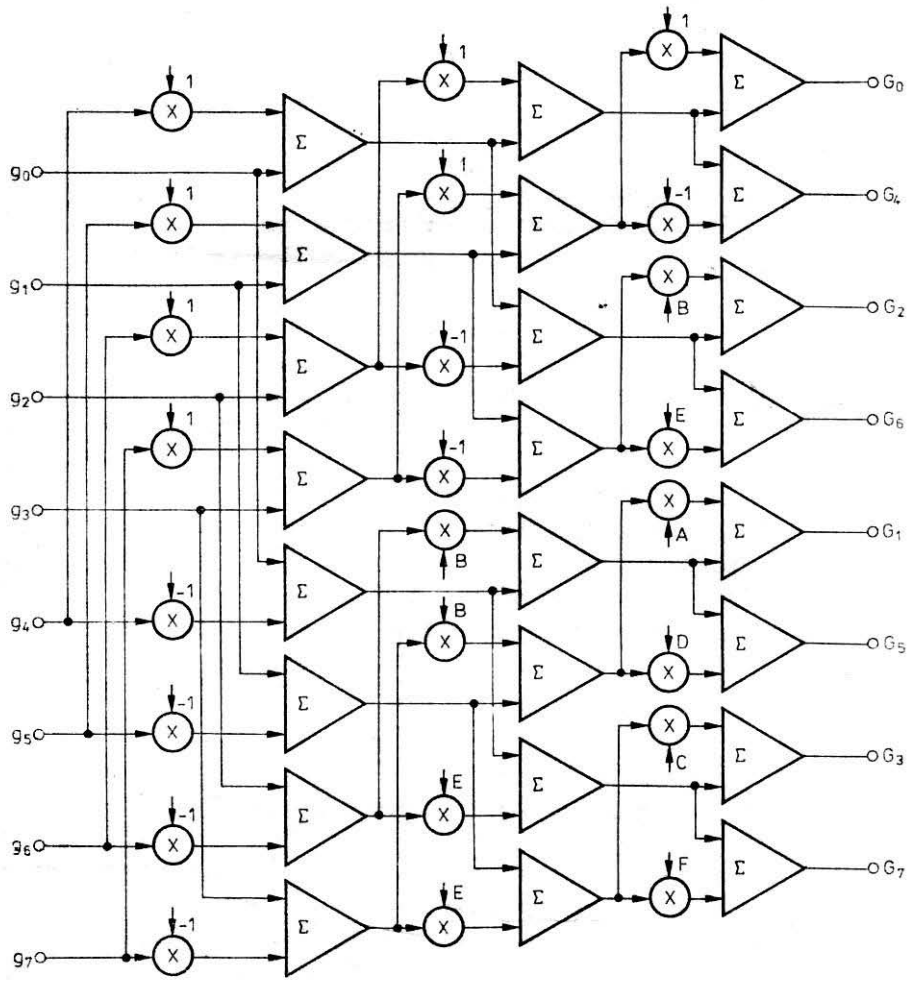
$$\begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & B & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & B \\ 0 & 0 & 0 & 0 & 1 & 0 & E & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & E \end{bmatrix} \tag{4.51}$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \end{bmatrix} \tag{4.52}$$

Bardzo regularna, blokowa struktura macierzy danych wzorami (4.50), (4.51) i (4.52) pozwala przypuszczać, że dokonanie transformacji drogą wymnożenia wejściowego sygnału (w postaci dyskretnego wektora o składowych $g(0), g(1), g(2), \ldots, g(7)$) kolejno przez wymienione macierze powinno być w sumie prostsze niż dokonanie transformacji według wzoru (4.48). Jest to przypuszczenie słuszne i głęboko uzasadnione, jako że algorytm FFT, którego istotę oddają macierze (4.50), (4.51) i (4.52) wraz z przenumerowaniem tworzącym z macierzy we wzorze (4.48) macierz (4.49) — dostarcza sposobu obliczenia transformacji Fouriera $N/\log_2 N$ razy szybciej niż postępowanie według wzoru (4.48). Wartość $N/\log_2 N$ może być w ogólnym przypadku (dla dużych $N$) bardzo duża — rzędu setek czy nawet tysięcy. Oznacza to możliwość wykonania w kilkadziesiąt sekund obliczeń, które realizowane na tym samym sprzęcie cyfrowym trwałyby kilka godzin — i jest to bez wątpienia jeden z bardziej znaczących wyników w zakresie algorytmów cyfrowego przetwarzania sygnałów.

Schemat obliczeń, wykorzystujący używaną wcześniej symbolikę, przedstawiono na rys. 4-26 (dla rozważanego przykładu $N = 8$). Łatwo zauważalna własność algorytmu FFT, widoczna także ze schematu podanego na rys. 4-26, polega na możliwości wykonania operacji „in situ", to znaczy odpowiednie składniki transformaty wynikowej $G(k)$ zajmują te same miejsca

w pamięci urządzenia liczącego, które uprzednio zajmowały elementy $g(n)$, co prowadzi w rezultacie do bardzo ekonomicznej gospodarki pamięcią. Wprawdzie ze względu na fakt, że wartości $G(k)$ są zespolone, potrzebne jest dla nich dwukrotnie więcej miejsca niż dla próbek czasowych sygnału $g(n)$, które są liczbami rzeczywistymi. Ponadto, pewna nieekonomiczność




4-26. Struktura algorytmu FFT. Oznaczenia jak we wzorze (4.48) i na rys. 4-23÷4-25. Algorytm FFT jest obecnie najwygodniejszą metodą uzyskiwania widma sygnału

w rozmieszczeniu wektora wynikowego transformaty $G(k)$ wynika z faktu, że transformata sygnału $g(n)$ będącego ciągiem wartości rzeczywistych jest symetryczna, a dokładniej mówiąc — parzyście sprzężona (por. wzór (4.25)), w wyniku czego wartości transformaty dla częstotliwości ujemnych mogą być jednoznacznie wyliczone na podstawie wartości transformaty dla odpowiednich wartości dodatnich, a niestety zajmują miejsce w wektorze G(k).

Sprawa rozmieszczenia w pamięci wartości będących wynikiem algorytmu

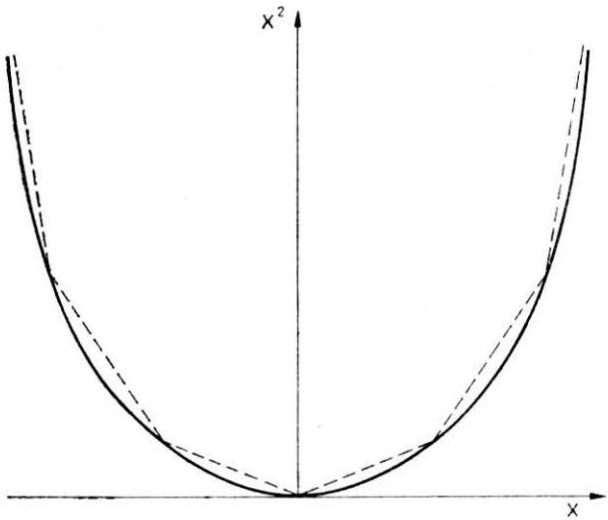
FFT wymaga nieco więcej uwagi, gdyż jest to niezbędne do właściwej interpretacji wyliczonych wartości $G(k)$. Łatwo się przekonać (na przykład ze wzoru (4.48)), że wartość $G(0)$ jest średnią wartości $g(0)$, $g(1)$, ..., $g(N)$, czyli reprezentuje składową stałą. Kolejne dalsze wartości $G(1)$,$G(2)$) odpowiadają równomiernie wzrastającym wyższym częstotliwościom aż do wartości $G(N/2)$, która odpowiada częstotliwości granicznej sygnału $f_g$, a dokładniej — połowie częstotliwości próbkowania $f_s/2$. Dalsze wartości $G(N/2+1)$, $G(N/2+2)$, ..., $G(N)$ odpowiadają tak samo rozmieszczonym kolejnym dalszym częstotliwościom, jednak ze względu na fakt okresowości widma dyskretnego wartości te są identyczne z wartościami dla odpowiednich częstotliwości ujemnych i ze względu na wspomnianą symetrię nie wnoszą nowych informacji. Ostatnia wartość $G(N)$ odpowiada częstości próbkowania $f_s$. Trzeba przy tym mieć na względzie fakt, że składowe sygnału w pobliżu częstości $f_g$ są z reguły zniekształcone przez zjawisko nakładania się widm i pomimo ich wyliczenia nie można traktować ich wartości jako wiarogodnych. Aby temu niekorzystnemu zjawisku przynajmniej częściowo przeciwdziałać, sygnał mowy przed poddaniem procesowi próbkowania jest ograniczony częstotliwościowo do przedziału $(0, f_g)$ za pomocą filtrów o wyjątkowo dużych nachyleniach charakterystyki dolnoprzepustowej. Przykładowo używając filtru o nachyleniu ponad 120 dB/oktawę (!) otrzymuje się próbki czasowe umożliwiające wykorzystanie ponad 80% wyliczonych składników transformaty $G(k)$ przy zakresie dynamiki ponad 72 dB. Należy jednak zwrócić uwagę, że nawet przy tak skrajnym ograniczeniu sygnału nie da się uznać za prawidłowe wszystkich obliczonych składowych widma. Tak więc stwierdzenie, że pierwszym $N/2$ elementom wektora $G(k)$ odpowiadają częstotliwości od 0 do $f_g = f_s/2$ służy głównie do tego, aby zdefiniować pojęcie szerokości pasma w przypadku analizy FFT. Szerokość ta jest stała i wynosi $\Delta f = 2 f_g/N$. Szerokość ta może ulegać poszerzeniu (a więc — pogorszeniu) na skutek wpływu „okna czasowego" uwzględniającego ograniczony czas trwania próbki przetwarzanego sygnału mowy. Podana wartość szerokości pasma, wynosząca $\Delta f = 2f_g/N$, odpowiada przypadkowi okna prostokątnego, natomiast dla okna Hanninga wartość ta musi być zwiększona o 50%, dla okna zaś gaussowskiego wzrost szerokości pasma wynosi aż 90%. Problem stosowania i wpływu okien o różnej długości na postać sygnału i jego widma będzie jednak bardziej szczegółowo rozważany w kolejnym podrozdziale.

Niekorzystną własnością analizy widmowej prowadzonej z wykorzystaniem algorytmu FFT jest limitowanie przez liczbę próbek $N$ rozdzielczości częstotliwościowej sygnału mowy, szczególnie w kontekście wcześniej wspomnianych dodatkowych ograniczeń: konieczności odrzucenia połowy wyliczonych prążków widmowych jako nie wnoszących nowej informacji oraz małej wiarygodności części wyliczonych wartości widmowych w otoczeniu częstości granicznej $f_g$. Częściową rekompensatą za te niedogodności jest pewna dodatkowa możliwość, chętnie wykorzystywana w większych systemach komputerowych, w których jest do dyspozycji odpowiednio duża pojemność pamięci. Jak wskazano wyżej, rozdzielczość częstotliwościowa zależy wyłącz-

nie od częstości próbkowania (lub inaczej — częstości granicznej, gdyż są to wartości związane) oraz od liczby próbek. Zawartość próbek jest przy tym mniej ważna — o ile tylko nie wprowadza nowych częstotliwości, fałszujących widmo rzeczywistego sygnału. Jeśli więc uzupełni się $N$ rzeczywistych próbek rozważanego fragmentu ciągłego sygnału mowy serią „próbek" mających zerową amplitudę — powiedzmy przykładowo jeśli doda się po $N$ próbkach sygnału $N$ zer, to wówczas z punktu widzenia rozważanych algorytmów analizy widmowej dwukrotnie zwiększy się liczba określanych obliczeniowo linii w badanym widmie i w tym samym stosunku polepszy się rozdzielczość. „Sztuczkę" opisaną tu można stosować wydłużając w razie potrzeby próbkę sygnału (przez dopisanie zer) więcej niż dwukrotnie i jedynym ograniczeniem w tym zakresie jest pamięć używanego komputera. Dokładność amplitudowa analizy częstotliwościowej prowadzonej z wykorzystaniem szybkiego przekształcenia Fouriera jest głównie limitowana dokładnością (wyrażaną liczbą bitów użytego przetwornika) wejściowych danych o przebiegu czasowym $g(n)$. Przykładowo 12-bitowy przetwornik gwarantuje 72 dB zakres dynamiki — pod warunkiem, że dokładność nie zostanie utracona w trakcie zaokrągleń w obliczeniach prowadzonych według zadanego algorytmu. Aby uniknąć utraty dokładności, obliczenia algorytmu FFT prowadzi się z wykorzystaniem większej liczby bitów niż używane do przedstawienia wejściowego sygnału $g(n)$ i używane do reprezentacji wyniku analizy $G(k)$. Typowo stosuje się przy obliczeniach 16 bitów, co zabezpiecza przed skutkami błędów biorących swoje źródło w zaokrągleniach w działaniach matematycznych. Dokładność i szybkość działań matematycznych związanych z algorytmem FFT w istotny sposób zależą także od sposobu realizacji mnożeń przez zespolone czynniki $e^{-j\frac{2\pi k}{N}}$. Wykorzystując tożsamość $e^{jx} = \cos x + j \sin x$ możemy stwierdzić, że do obliczeń wymagane są wartości funkcji trygonometrycznych $\sin x$ oraz $\cos x$, które typowo w systemach komputerowych obliczane są ze wzorów odpowiadających rozwinięciom na szeregi — a więc w sposób czasochłonny. Ponieważ do realizacji przekształcenia Fouriera z wykorzystaniem algorytmu FFT potrzebne są jedynie niektóre, regularnie rozmieszczone wartości wskazanych funkcji (a właściwie jednej z nich, gdyż mając wartość $\sin x$ można bez trudu wyznaczyć $\cos x$ i na odwrót), przeto jest rzeczą korzystną i celową posługiwanie się zapamiętanymi, stabelaryzowanymi wartościami funkcji trygonometrycznych, a nie programami ich sukcesywnego wyliczania.

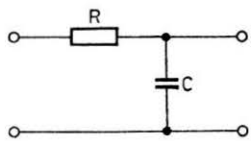
Algorytm FFT dokonuje przekształcenia Fouriera w sposób bliski definicyjnemu wzorowi, nie wymaga przeto żadnych uzupełnień. Analiza częstotliwościowa wykonywana z wykorzystaniem filtrów — analogowych bądź cyfrowych — dostarcza w istocie przebiegów czasowych o częstotliwościach leżących w przedziale wynikającym z parametrów filtru. Aby na tej podstawie określić widmo mocy rozważanego sygnału mowy, należy sygnały wejściowe z filtrów poddać dwu zabiegom: podnoszenia do kwadratu oraz uśredniania. Oba wspomniane zabiegi są łatwiejsze do przeprowadzenia

w przypadku sygnałów mających postać cyfrową. W przypadku sygnałów analogowych realizacja odpowiednich procesów dokonywana jest w sposób przybliżony, wnoszący dodatkowy błąd do metody. Operacja podnoszenia do kwadratu, w znacznej części analogowych systemów widmowej analizy sygnałów, jest wykonywana w sposób przybliżony z wykorzystaniem układów nieliniowych zrealizowanych z wykorzystaniem wzmacniacza operacyjnego i zespołu odpowiednio spolaryzowanych diod. Diody te formują charakterystykę statyczną układu, aproksymującą przebieg paraboliczny (czyli potrzebną funkcję $x^2$) za pomocą linii łamanej złożonej z odpowiednio rozmieszczonych odcinków charakterystyk prostoliniowych (rys. 4-27).




4-27. Aproksymacja funkcji kwadratowej za pomocą linii łamanej, wykorzystywana przy wyznaczaniu widma mocy sygnału

Taka realizacja w niektórych systemach uznawana jest za zbyt prymitywną i zastępowana jest analogowymi członami mnożącymi, działającymi na hallotronach lub układach półprzewodnikowych o charakterystyce logarytmicznej. W innych, użytkowych układach nawet realizacja oparta na linii łamanej uznana jest za zbyt kosztowną i zastąpiona jest przez prostowanie dwupołówkowe przebiegów na wyjściu poszczególnych filtrów. Rozbieżności między wartościami poprawnymi a uzyskanymi z uproszczonej analizy mogą jednak wówczas być dość znaczne.




4-28. Prosty czwórnik RC używany do uśredniania sygnału metodą „ważenia wykładniczego"

Kolejna czynność to uśrednianie wyniku w czasie. Pojawia się przy tym problem sposobu uśredniania, gdyż używane są, między innymi, dwa podejścia: wyliczanie wartości średniej w ustalonym czasie („prostokątne" ważenie sygnału) oraz uśrednianie w prostym filtrze dolnoprzepustowym *RC* (rys. 4-28), nazywane czasami „ważeniem wykładniczym". Oba typy uśred-

niania wprowadzają pewne błędy do widma analizowanego sygnału i ich stosowanie jest kwestią wyboru tego wariantu, którego niedoskonałości w kontekście konkretnego zastosowania wydają się mniej istotne. Wiele typów aparatury analitycznej dostarcza użytkownikowi swobodnych możliwości wyboru między wymienionymi ewentualnościami. Generalnie, dysponując próbką sygnału o czasie trwania $T$ korzystniej jest stosować uśrednianie równomierne*) (z „prostokątnym" oknem), natomiast uśrednianie wykładnicze jest bardziej wygodne dla sygnałów wolnozmiennych, dla których widmo powinno „nadążać" za zmianami sygnału i uśrednianie powinno mieć charakter „kroczący". Uśrednianie za pomocą dolnoprzepustowego filtru RC ma również zastosowanie w tych przypadkach, kiedy wymagany jest równomierny (w skali częstotliwości) rozkład błędów statystycznych (zależnych od iloczynu szerokości pasma $B$ i czasu uśredniania $T_A$). Osobnym zagadnieniem, związanym z rozważanym tu problemem uśredniania sygnału na wyjściach filtrów, jest wybór czasu uśredniania. Jeśli analizowany sygnał ma określoną dominantę o częstotliwości $f$, wówczas dla uzyskania poprawnych rezultatów wybiera się czas uśredniania zgodnie ze wzorem:

$$T_A > 3/f \tag{4.53}$$

W przypadku sygnałów bez dominującej składowej lub sygnałów szumowych można oprzeć się na innym kryterium. Przy ustalonym pasmie częstotliwości dokonującego analizy filtru $B$ oraz przy założonym poziomie odchylenia standardowego wyniku analizy $E$ (w dB) czas uśredniania można wyznaczyć z zależności

$$T_A > \frac{18{,}84}{BE^2} \tag{4.54}$$

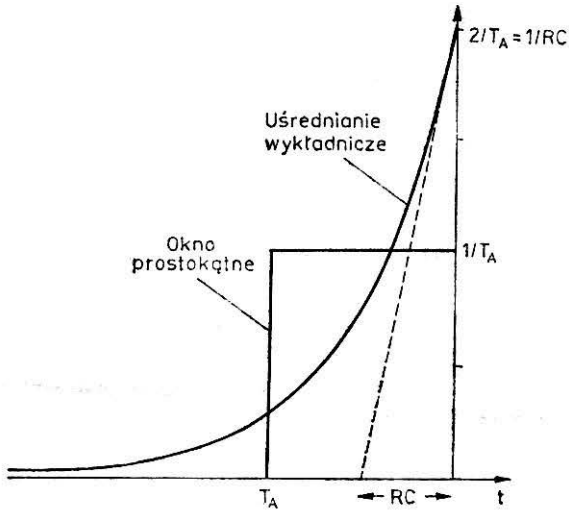
Wartość czasu uśredniania dobiera się także niekiedy empirycznie lub ustala na poziomie ustalonym zwyczajowo dla danego typu sygnału. Przykładowo dla sygnału mowy przyjęto tradycyjnie $T_A = 10$ ms. Taka wartość czasu uśredniania wynika z dynamiki sygnału mowy i jego widma.

Dla uśredniania liniowego interpretacja czasu $T_A$ jest oczywista (rys. 4-29), natomiast otwarta pozostaje interpretacja czasu uśredniania dla ważenia wykładniczego, gdyż przebieg zanikający wykładniczo trwa — teoretycznie — nieskończenie długo, zanim zaniknie do zera. Przyjmuje się jednak, dla uśredniania za pomocą dolnoprzepustowego filtru RC czas uśredniania określony jako (rys. 4-29)

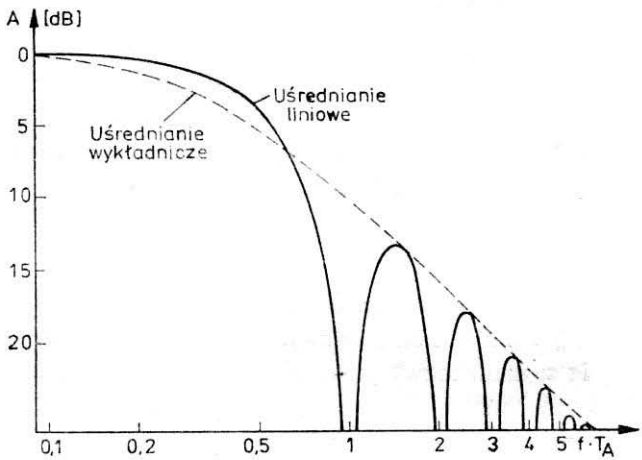
$$T_A = 2RC \tag{4.55}$$

W dziedzinie częstotliwości kształty charakterystyk częstotliwościowych ważenia prostokątnego i wykładniczego różnią się (rys. 4-30), co należy mieć na uwadze dokonując oceny wyników analizy.

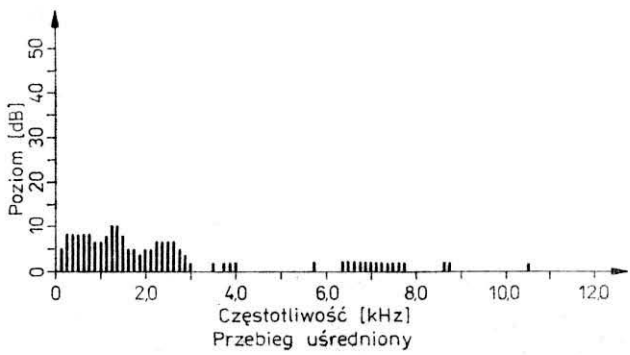
*) Uśrednianie równomierne przebiega zgodnie ze wzorem $x_{śr} = \frac{1}{T_A}\int_0^{T_A} x(t)\,dt$ i dlatego używana jest również nazwa uśrednianie liniowe lub całkowanie liniowe. W zastosowaniach całka zastępowana bywa sumą — szczególnie dla sygnałów cyfrowych.




4-29. Porównanie czasowej formy okna prostokątnego i ważenia wykładniczego. Czas uśredniania $T_A$ jest równy długości okna prostokątnego lub podwójnej wartości stałej czasowej $RC$

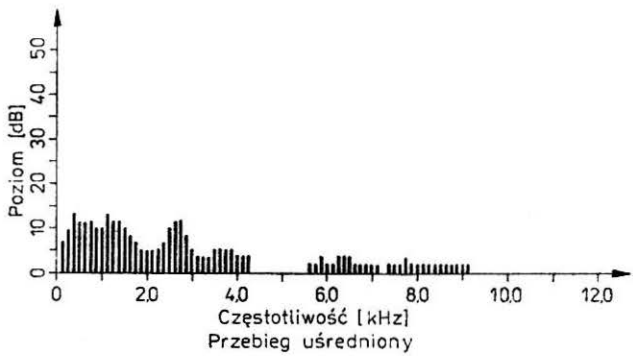



4-30. Porównanie charakterystyki widmowej (amplitudowo-częstotliwościowej) okna prostokątnego (linia ciągła) i uśredniania wykładniczego (linia przerywana). Częstotliwość w jednostkach względnych $fT_A$

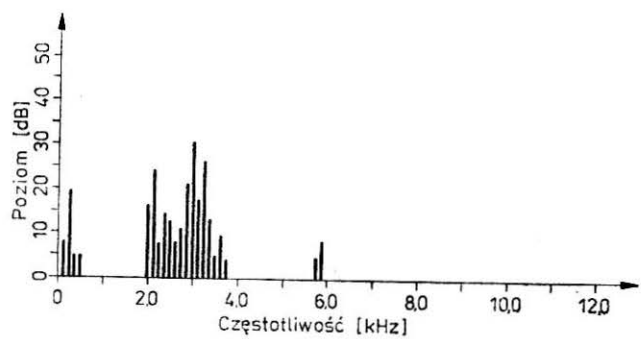



4-31. Przykład uśrednionego widma sygnału mowy; wypowiedź: *brat Zygmunta*

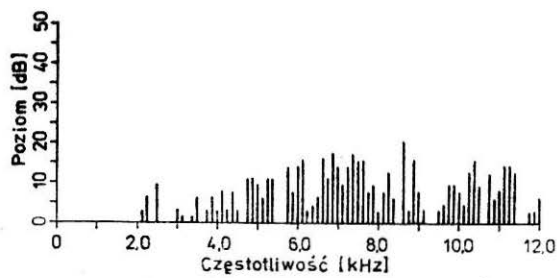
Analiza częstotliwościowa ma zastosowanie w odniesieniu do sygnału mowy jedynie w kontekście tzw. długoterminowego widma sygnału mowy, uwzględniającego obraz widmowy sygnału w czasie wielokrotnie większym od okresów quasi-stacjonarności widma, lub w odniesieniu do tych fragmentów sygnału, dla których można przyjąć, że w czasie ich generacji widmo sygnału nie podlega istotnym zmianom. Przykładowo na rys. 4-31 pokazano długoterminowe widmo sygnału mowy, obejmujące wypowiedź *brat Zygmunta* (głos męski), zaś na rys. 4-32 pokazano podobne widmo dla wypowiedzi *stos drewna*. Widma te uzyskano za pomocą uśredniania kilkuset widm chwilowych z wykorzystaniem maszyny cyfrowej o dużej pamięci. Możliwe jest jednak uzyskiwanie podobnych widm metodami aparaturowymi, w szczególności analogowymi. Na rysunkach 4-33 i 4-34 pokazano przykładowo widma chwilowe sygnału mowy w wybranych odcinkach czasu, odpowiadających artykulacji określonych głosek.




**4-32. Inny przykład uśrednionego widma sygnału mowy; wypowiedź: *stos drewna***

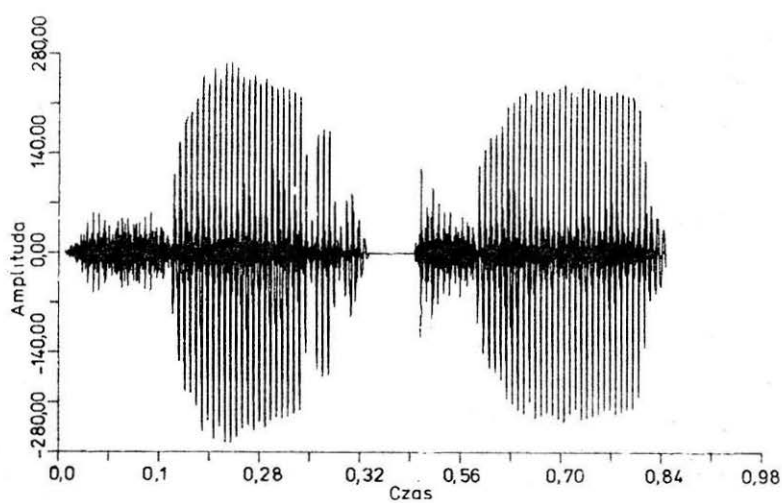



**4-33. Krótkookresowe widmo sygnału mowy (czas uśredniania: 9 ms) przedstawiające stan ustalony samogłoski *i*. Widoczne „skupienie" widma w zakresie małych częstotliwości oraz wyraźne formanty**




**4-34. Krótkookresowe widmo sygnału mowy — spółgłoska szumowa *s*. Widmo jest odmienne od przedstawionego na rys. 4-33: szerokie, zlokalizowane w zakresie dużych częstotliwości i pozbawione wyraźnych struktur**

Widma podobne do przytoczonych są bardzo przydatne w wielu badaniach, przykładowo widmo długookresowe może być wykorzystane do określania globalnych charakterystyk sygnału, potrzebnych przy projektowaniu aparatury wzmacniającej (nagłaśnianie), linii transmisji, a także układów testujących wymienione urządzenia „sygnałem mowopodobnym". Do rozpoznawania mowy lub w celu szczegółowych badań nad jej artykulacją, percepcją i transmisją widma takie są mało przydatne. Dotyczy to także widm chwilowych wybranych, quasi-ustalonych fragmentów sygnału mowy. Ich przydatność do badań nad mechanizmami artykulacji i percepcji jest bezsporna, mogą one także być przydatne przy analizie i rozpoznawaniu izolowanych fonemów, sylab, diad, triad, logatomów. Jednak zarówno przy transmisji mowy, jak i przy jej rozpoznawaniu bardzo istotne znaczenie mają czasowe zmiany widma sygnału. Wynika to zresztą ze wszystkiego, co zostało powiedziane na temat analizy sygnału mowy w systemie słuchowym człowieka, a także z opisanego wyżej modelu procesu artykulacji mowy. Ruchy narządów artykulacyjnych dynamicznie kształtują widmo mowy, a zmiany objętości poszczególnych wnęk rezonansowych tworzących się w trakcie głosowym formują płynne przejścia od jednej postaci widma do drugiej, przy czym zmiany te bywają bardziej istotne z punktu widzenia procesu rozpoznawania mowy niż stany ustalone głosek! Istotnie, wykazano w licznych badaniach, iż wiele głosek można poprawnie rozpoznawać wyłącznie przy słyszeniu poprzedzających je i następujących po nich głosek, natomiast stan ustalony sygnału odpowiadający stricte rozważanej głosce nie niesie wystarczających informacji — przedstawiony słuchaczom w izolacji nie jest poprawnie interpretowany. Sygnał spreparowany w sposób polegający na usunięciu głoski przy pozostawieniu „stanów przejściowych" przed i po rozważanej głosce — rozpoznawany jest bez trudu w całości — przy czym słuchacze nie dostrzegają braku stanu ustalonego badanego fonemu.



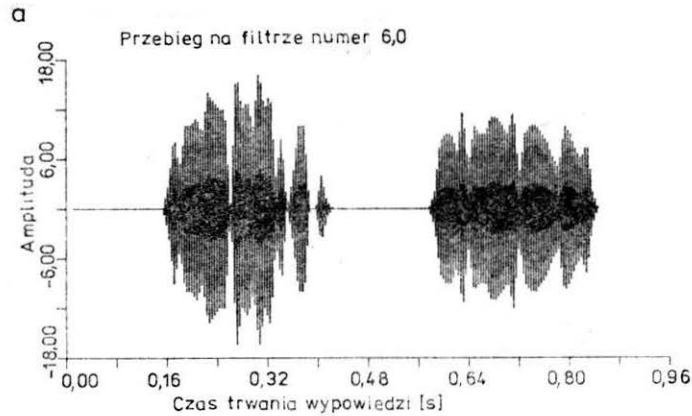

4-35. Przebieg sygnału mowy, który może podlegać filtrowaniu (wypowiedź *serce* głos męski)

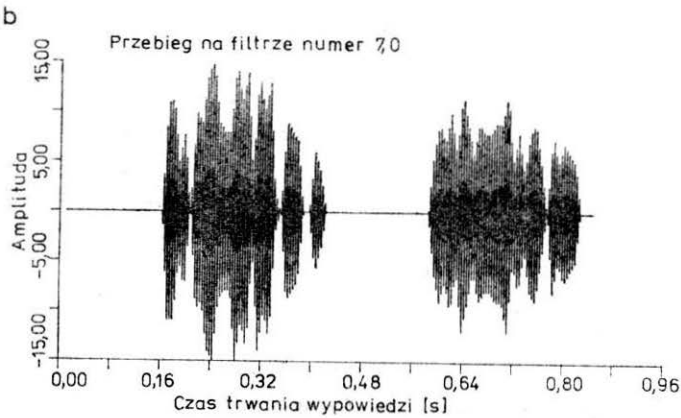
Przytoczone argumenty sprawiają, że kończąc rozważania na temat analizy częstotliwościowej trzeba od razu rekomendować jako najwłaściwszą formę prezentacji sygnału mowy do analizy, rozpoznawania i badań nad transmisją analizę czasowo-częstotliwościową, będącą przedmiotem rozważań w kolejnym podrozdziale.

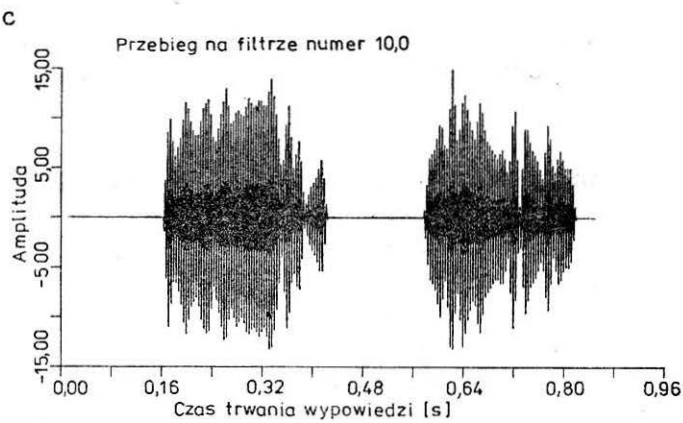
## 4.3. Czasowo-częstotliwościowa zmienność sygnału mowy

Poprzedzające dwa podrozdziały stanowiły opis typowych technik wykorzystywanych w analizie sygnałów: analizę w dziedzinie czasu i analizę częstotliwościową. Stwierdzono w nich, że przebieg czasowy sygnału mowy przykładowo pokazany na rys. 4-35 zawiera w istocie wszystkie niezbędne do analizy i rozpoznania elementy, ale w niedogodnej formie. Można więc dobrać zestaw filtrów o pożądanych własnościach i badać przebiegi sygnału w wybranych pasmach częstotliwości. Przykładowo na rys. 4-36 pokazano przebieg tego samego sygnału, co prezentowany na rys. 4-35 w wybranych pasmach. Każdy z tych przebiegów (a można ich sporządzić znacznie więcej — w typowej analizie wykorzystuje się od kilkudziesięciu do kilkuset pasm częstotliwości) dostarcza informacji o dynamice sygnału w wybranym fragmencie widma. Przykładowo dla wypowiedzi z rys. 4-35 widoczne są w pasmach odpowiadających małym częstotliwościom segmenty samogłoskowe *e*, a w pasmach dużej częstotliwości segmenty odpowiadające narastaniu i zanikaniu szumowej głoski *s* (występują one dwukrotnie, gdyż dźwięk zapisywany ortograficznie jako *c* jest w istocie złożeniem plozyjnej głoski *t* i szumowej głoski *s*).

W sygnale rozłożonym na pasma częstotliwościowe i analizowanym w formie przebiegów czasowych w tych pasmach zawarta jest cała niezbędna informacja, jednak objętość koniecznego do analizy materiału wzrasta do rozmiarów trudnych do zaakceptowania. Można oczywiście wybrać określony odcinek czasu i przedstawić widmo sygnału w tym momencie w postaci przedstawionej na rys. 4-33 i 4-34, jednak takie wyrywkowe analizowanie sygnału jest również mało użyteczne. W celu uchwycenia równocześnie wymiaru częstotliwościowego sygnału i jego czasowej zmienności trzeba zastosować reprezentację trójwymiarową. Ustawienie widm chwilowych podobnych do pokazanych na rys. 4-33 i 4-34 kolejno, jedne za drugimi, niewiele daje, gdyż widma się wzajemnie zasłaniają. Konieczne jest spojrzenie z góry — komputer daje możliwość takiej prezentacji sygnału. Na rysunku 4-37 a÷c pokazano kolejne obrazy zmienności częstotliwościowo-czasowej sygnału pod coraz większym kątem. Na rysunku 4-38 a÷e pokazano ten trójwymiarowy obraz z różnych kierunków — zależnie od potrzeb i wymagań obserwatora. Przy niektórych ustawieniach głównie obserwuje się czasową zmienność sygnału w poszczególnych pasmach, w wyniku innego ustawienia otrzymuje się obraz eksponujący głównie strukturę widma w poszczególnych momentach czasu i służyć on może śledzeniu drobnych detali i zmian obrazu sygnału w sąsiednich widmach. Wszystkie te obrazy przedstawiają jedną i tę samą wypowiedź *serce* — ale jakże inaczej zaprezentowaną.



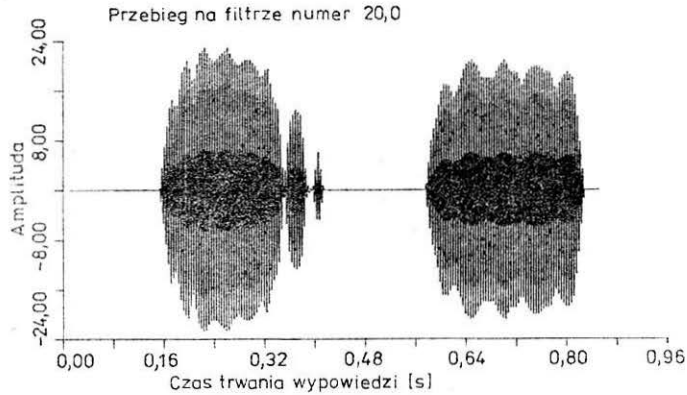



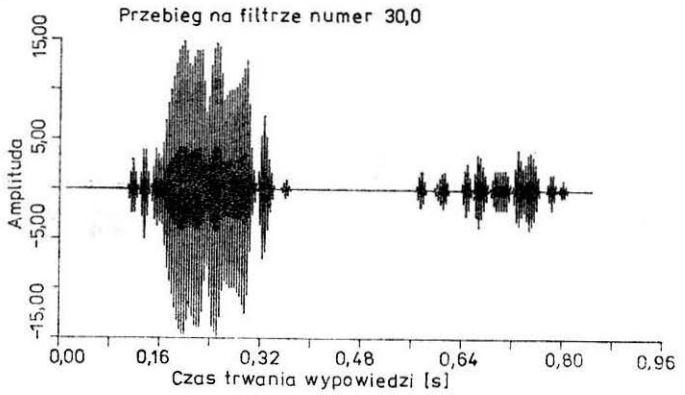

**4-36. Przebieg będący wynikiem filtracji sygnału z rys. 4-35 za pomocą filtra o szerokości pasma 125 Hz i częstotliwości środkowej:**

**a — 750 Hz (widoczne głównie przebiegi związane z samogłoskami), b — 875 Hz, c — 1250 Hz, d — 2500 Hz, e — 3750 Hz, f — 5000 Hz (znikły praktycznie składowe pochodzące od samogłosek, widoczne są wyłącznie przebiegi spółgłoskowe — głoski szumowe *s*), g — 6250 Hz, h — 7500 Hz i — 10 000 Hz**

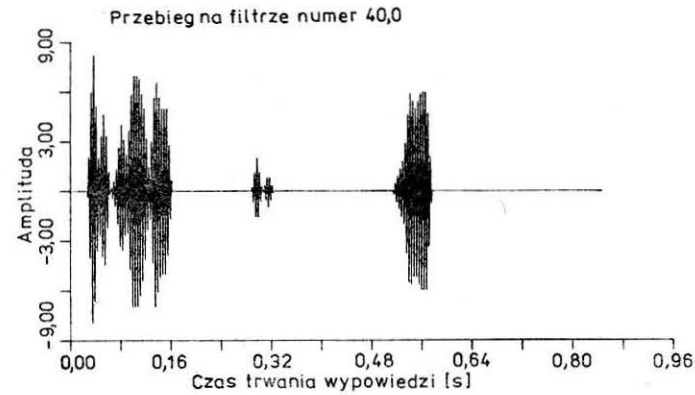
d



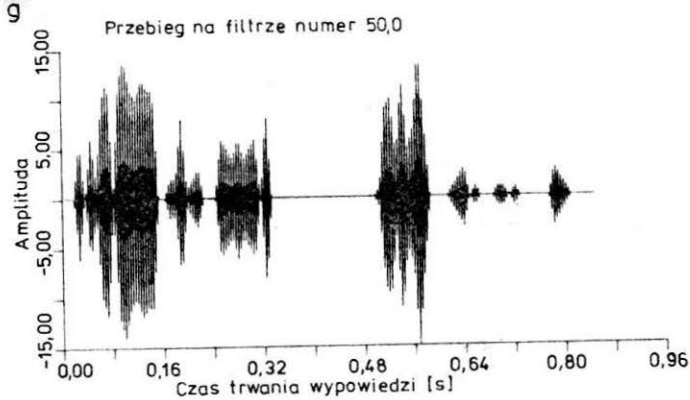

e



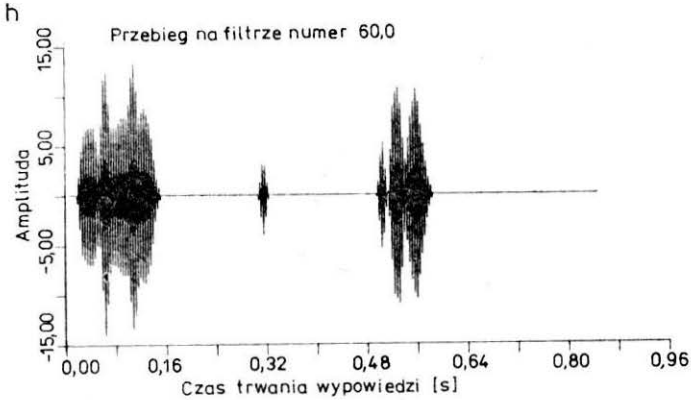

f



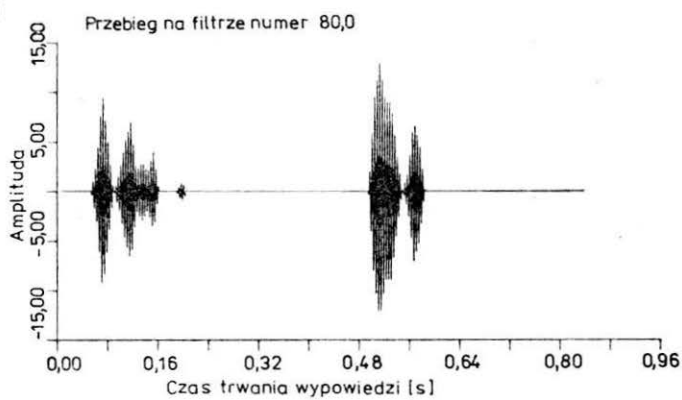
g

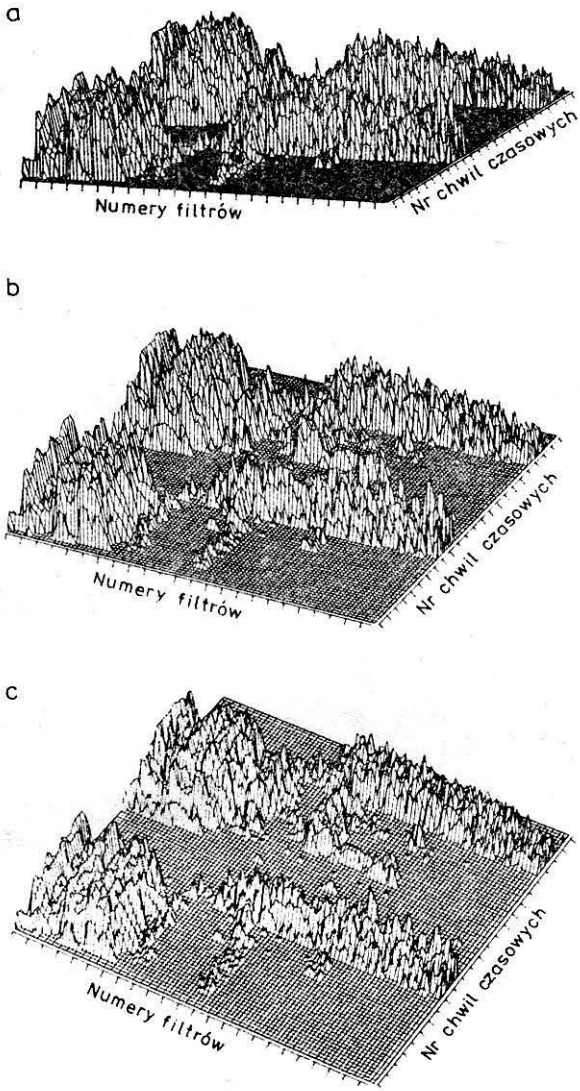



h




i



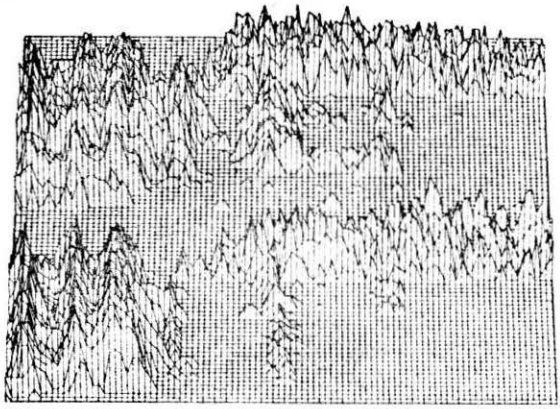



4-37. Próba spojrzenia na widma chwilowe (wypowiedź *serce*, głos męski), ustawione jedno za drugim, według kolejnych chwil czasowych — z góry: a — kąt widzenia niewielki — 15° w stosunku do podstawy (wiele szczegółów widm ulega zatarciu), b — kąt widzenia 30°, c — kąt widzenia 45° — optymalne warunki do obserwacji szczegółów widma

Zalety czasowo-częstotliwościowej prezentacji sygnału docenić można porównując obrazy tak przedstawionego sygnału mowy dla różnych wypowiedzi. Na rysunkach 4-39 a ÷ e przedstawiono przykładowo widmo kilku prostych wyrazów, na rys. 4-40 zaś przedstawiono czasowo-częstotliwościową zmienność sygnału mowy w wypowiedziach *stos drewna* (a) i *Brat Zygmunta* (b). Są to te same wypowiedzi, dla których uśrednione widma przedstawiono na rys. 4-31 i 4-32. Łatwo porównać, o ile bardziej szczegółowy i pełen treści jest obraz czasowo-częstotliwościowej zmienności sygnału niż obraz całościowy, uśredniony.

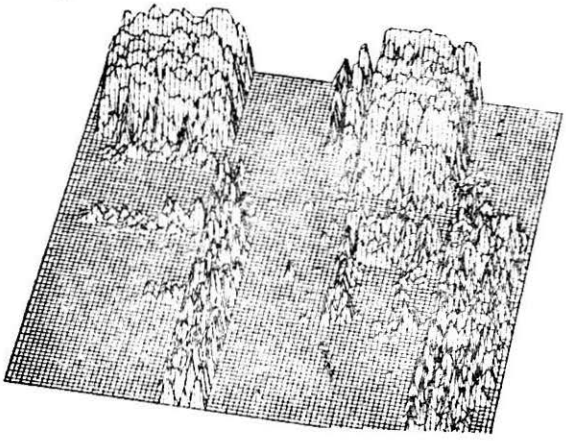
Nie zawsze badacz zajmujący się analizą sygnału mowy ma do dyspozycji komputer o dużych możliwościach graficznych i dlatego nie zawsze możliwe jest prezentowanie „trójwymiarowych" dynamicznych widm w postaci

a

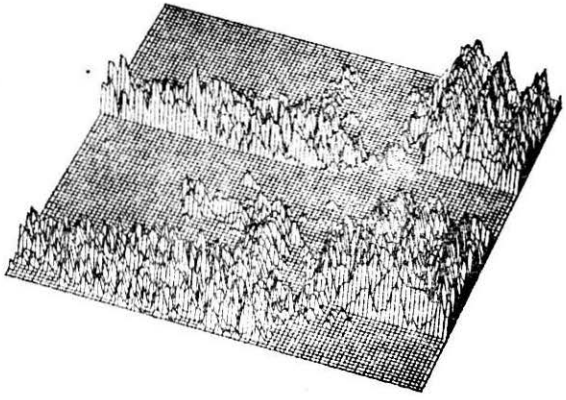


**4-38.** „Trójwymiarowe" widmo może być oglądane pod dowolnym kątem (wypowiedź *serce*):
a — 0°, b — 100°, c — 150°, d — 200°, e — 250°

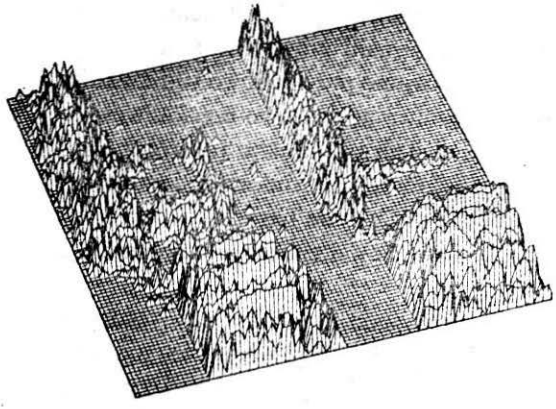
b



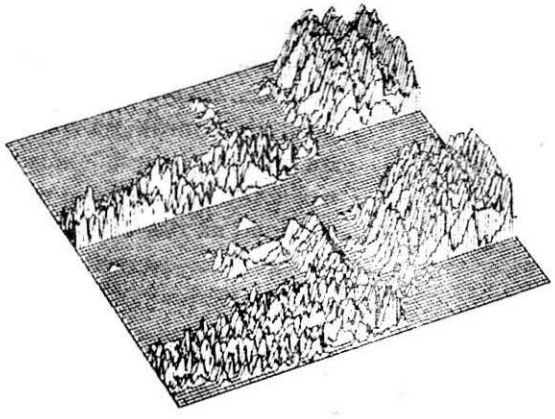
c

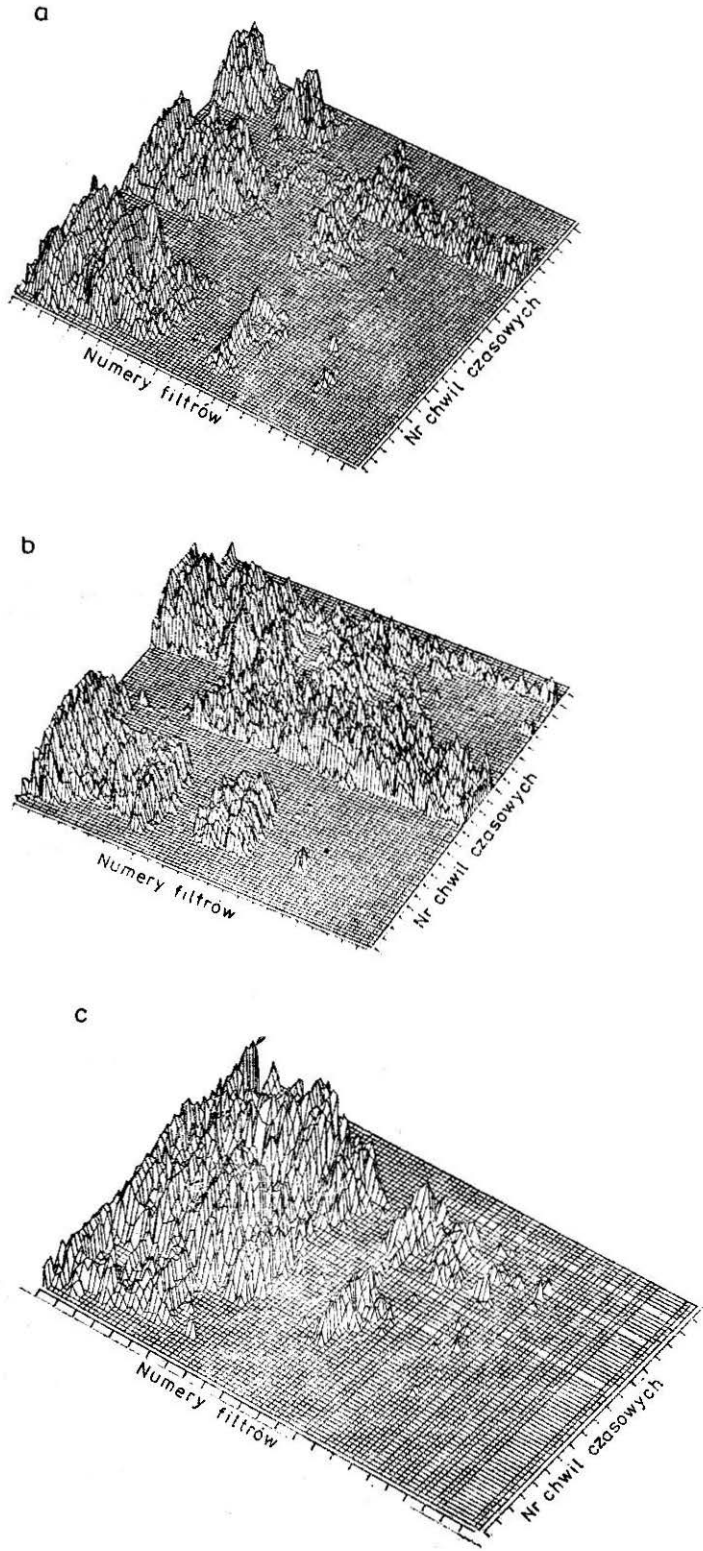


d

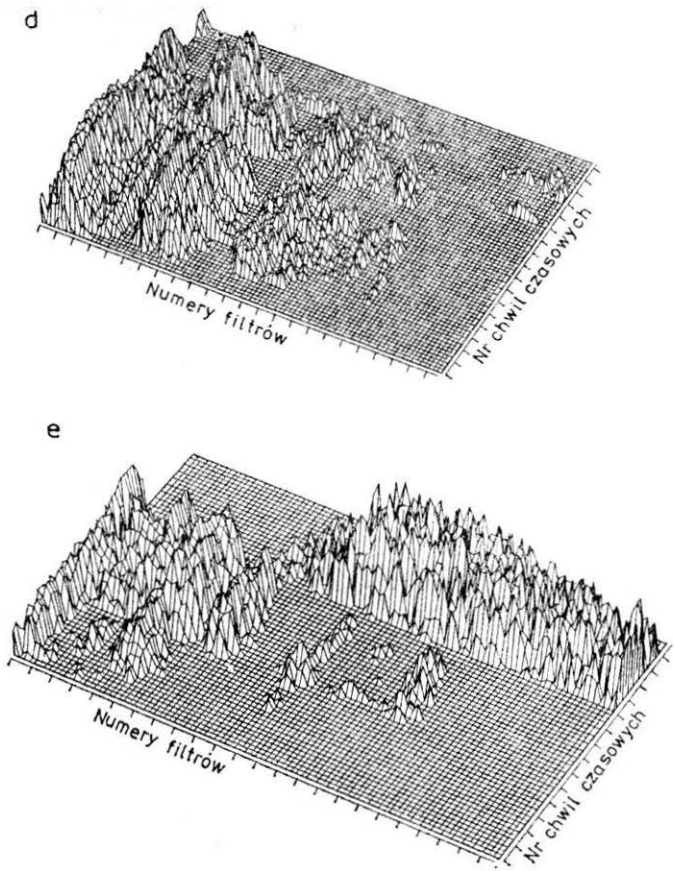


e

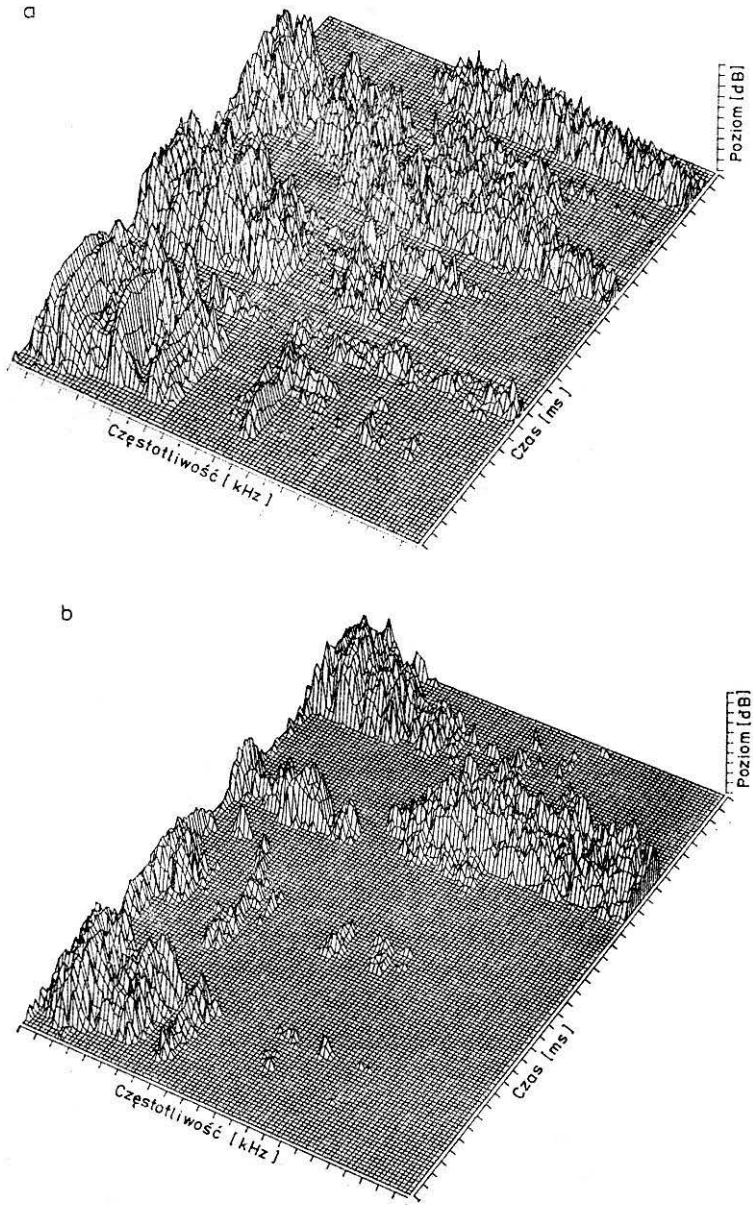





**4-39. Widma trójwymiarowe różnych wypowiedzi:**
**a** — *oferta*; **b** — *wiosło*, **c** — *byłem*, **d** — *wino*, **e** — *sys*

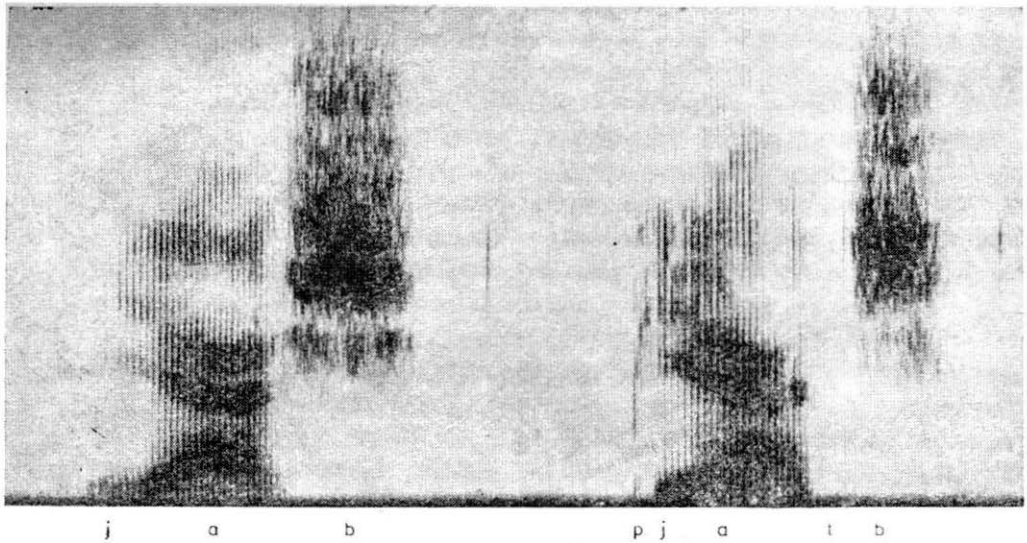



rysunków podobnych do prezentowanych. Z merytorycznego punktu widzenia równoważne są inne przedstawienia, być może skromniejsze wizualnie, ale dostarczające w praktyce tych samych wiadomości na temat analizowanego sygnału, jego widma i zmienności tego widma w czasie. Zazwyczaj wykorzystuje się przy tym prezentację sygnału na dwuwymiarowej płaszczyźnie, której jedna oś oznacza czas, a druga częstotliwość. Trzeci wymiar — amplitudę sygnału — prezentuje się stosując odpowiednią skalę stopnia zaciemnienia papieru (im silniej zaciemniony dany punkt w układzie czas-częstotliwość, tym wyższą amplitudę ma w danym momencie czasu składowa sygnału o wyróżnionej częstotliwości). Przykładowy wykres tego typu pokazano na rys. 4-41. Jak widać, przyjęty sposób prezentacji tworzy rodzaj mapy „górzystego terenu" przedstawionego w pseudoperspektywie na rys. 4-37 ÷ 4-40. Analogia z mapą zwiększa się w niektórych nowocześniejszych aparatach prezentujących czasową zmienność sygnału mowy rozłożonego na poszczególne pasma częstotliwości za pomocą kodu barw. Ten sposób prezentacji, szczególnie przydatny w komputerowych systemach wyposażonych w barwne monitory ekranowe zyskuje na znaczeniu, gdyż umożliwia szybką i precyzyjną lokalizację interesujących struktur w widmowo-czasowym przebiegu sygnału i jest dobrze dostosowany do możliwości percep-




**4-40. Widmo wypowiedzi**
a — *stos drewna*, b — *brat Zygmunta*

cyjnych człowieka. W nowszych systemach tego rodzaju wykorzystuje się dodatkowo możliwości podkreślania wybranych struktur przez technikę rozjaśniania, migotania lub inwersyjnej prezentacji na ekranie, co w połączeniu z możliwością swobodnego wprowadzania na ekran napisów, wartości liczbowych, linii podkreślających obserwowane struktury (na przykład przebieg formantów) daje w sumie dogodne narzędzie, pozwalające w trybie interakcyjnym badać i analizować dowolne szczegóły sygnału.
Urządzenia omawianego typu są jednak bardzo kosztowne i dlatego warto




4-41. Przykładowy sonogram uzyskiwany w aparatach dokonujących czasowo-częstotliwościowej analizy sygnału mowy. Amplituda sygnału zaznaczona jest stopniem zaciemnienia papieru. Spektrogram wykonano w Zakładzie Fonetyki Akustycznej Instytutu Podstawowych Problemów Techniki PAN w Poznaniu i reprodukowany jest za łaskawym zezwoleniem prof. Wiktora Jassema (wypowiedzi *Jaś* oraz *piać*)

wspomnieć o innej możliwości, łatwej do zrealizowania w warunkach krajowych, gdyż w praktyce nie wymagającej żadnych specjalnych urządzeń. Mowa o możliwości sporządzania „map" rozkładu czasowo-częstotliwościowego amplitud analizowanego sygnału w postaci alfanumerycznych wydruków z komputera.

Pewien problem w omawianiu i prezentacji wszystkich form dwuwymiarowego, częstotliwościowo-czasowego odwzorowania sygnału mowy wynika w związku z terminologią. Co to jest przebieg czasowy, to dobrze wiadomo, łatwo też zdefiniować widmo. Ale ta specyficzna hybryda? Przez pewien czas używano określenia sonogram wiążąc to z faktem, że pierwotne formy prezentacji tego typu uzyskiwano z aparatu o nazwie Sona-Graph (patrz rys. 4-41). Potem zaczęto używać nazwy spektrogram dynamiczny — adekwatnej, ale niewygodnej z uwagi na długość. Pojawiają się propozycje nazwania rysunków tego rodzaju wideogramami. Nie czas tu i miejsce na relacjonowanie i próby rozstrzygania tych sporów, warto jednak, aby Czytelnik był świadom tej rozmaitości i nie gubił się przy czytaniu doniesień różnych autorów, nazywających w różny sposób — w istocie jedno i to samo.

Znacznie ważniejszy od nazw jest problem sposobu wyliczania wartości czasowo-częstotliwościowego spektrum sygnału. Niestety, w tym momencie nie uda się dalej ignorować efektu „okna czasowego" i tym problemem należy się teraz zająć.

Obliczając dynamiczne widmo sygnału mowy nie możemy już posługiwać się wzorami (4-23) ÷ (4.43), właściwsze jest natomiast posługiwanie się wzorem

$$G(k, n) = \sum_{l=-\infty}^{\infty} g(l)h(n-1)e^{-j2\pi kl} \tag{4.56}$$

gdzie $G(k, n)$ należy interpretować jako wartość (zespoloną) transformaty dyskretnego sygnału $g(l)$ w chwili $n$ dla dyskretnej częstotliwości $k$. Nie rozważamy tu przypadków sygnału ciągłego lub ciągłej wersji transformaty $G$, gdyż praktycznie krótkookresowa transformata Fouriera (gdyż tak bywa nazywany wzór (4.56)) jest obliczana jedynie z wykorzystaniem sprzętu cyfrowego. Jak widać we wzorze (4.56), centralną pozycję zajmuje funkcja okna $h$. Rozważa się różne funkcje okna. Wspólną ich cechą jest zawsze to, że mają niezerowe wartości jedynie wewnątrz pewnego przedziału swoich wartości. Załóżmy, że w prowadzonych rozważaniach szerokość przedziału, wewnątrz którego spełniony jest warunek $h(n) > 0$, wynosi $N$. Funkcja okna powinna być symetryczna wokół zera, to znaczy powinna być określona zarówno dla $n$ dodatnich, jak i ujemnych. Ponadto powinna spełniać warunek $h(-n) = h(n)$. Na ogół dla wygody zapisu dokonuje się korekty układu współrzędnych zgodnie ze wzorem

$$n' = n + \frac{N-1}{2} \tag{4.57}$$

dzięki czemu niezerowe wartości funkcji $h(n')$ są uzyskiwane wyłącznie dla $n' \geqslant 0$, zaś maksymalna wartość, przypadająca w oryginalnej funkcji $h(n)$ w zerze ($h(0) \geqslant h(n)$ dla wszystkich $n$) w funkcji o skorygowanym argumencie przypada dla wartości $n' = \frac{N-1}{2}$. Przy tak skorygowanym zapisie można z łatwością opisać kilka bardziej popularnych funkcji okna. Najprostsze pojęciowo, ale wprowadzające największe zakłócenia do wynikowego sygnału jest okno prostokątne. Definicja tego okna jest prosta:

$$h(n') = 1 \quad \text{dla} \quad 0 \leqslant n' \leqslant N-1 \tag{4.58}$$

Okno także ma pożądane własności z punktu widzenia selektywności analizy (nie pogarsza rozdzielczości częstotliwościowej w stosunku do szerokości pasma wprowadzanego przez stosowaną metodę analizy częstotliwościowej — na przykład w stosunku do pasma pojedynczego filtru), ale jego widmo, ze względu na dużą liczbę tzw. listków bocznych (rys. 4-42), jest niepożądane. Ponieważ wynikowe widmo sygnału otrzyma się w wyniku splotu widma sygnału i widma okna, listki boczne mogą deformować widmo w istotny i trudny do skorygowania sposób. Dlatego zaproponowano wiele funkcji okna o łagodniej opadających „zboczach", co w rezultacie prowadzi do zmniejszenia wpływu listków bocznych (których amplituda relatywnie maleje) i odtwarzania sygnału bez zniekształceń. Niestety odbywa się to kosztem poszerzenia (a więc na ogół — pogorszenia) pasma analizy.

Najprostszym rozwiązaniem okna o łagodnie opadających zboczach jest okno trójkątne, nazywane w literaturze oknem Bartletta. Wzór opisujący tę funkcję ma postać

$$h(n') = \begin{cases} \dfrac{2n'}{N-1} & \text{gdy } 0 \leqslant n' \leqslant \dfrac{N-1}{2} \\ 2-\dfrac{2n'}{N-1} & \text{gdy } \dfrac{N-1}{2} \leqslant n' \leqslant N-1 \end{cases} \tag{4.59}$$

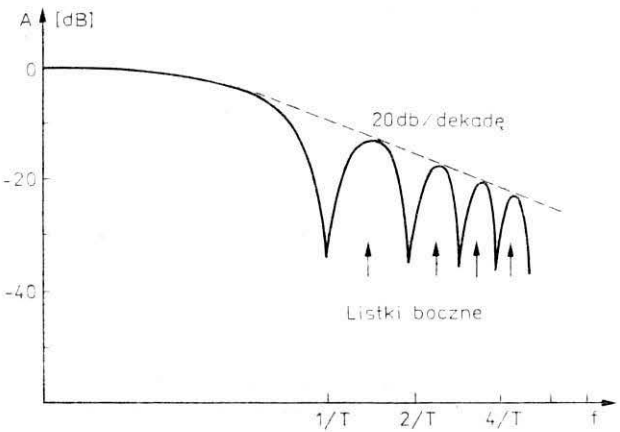
Okno Bartletta charakteryzuje się jednak nadal niewielkim polepszeniem amplitudy sygnału do amplitudy listków bocznych (rys. 4-43). Lepsze pod tym względem jest okno o kształcie funkcji $\cos^2$ nazywane oknem Hanninga. Jest to bardzo często wybierana w praktyce postać funkcji okna, charakteryzująca się dobrymi własnościami (poziom pierwszego listka bocznego stłumiony jest w stosunku do poziomu sygnału o $-32$ dB, a obwiednia kolejnych dalszych listków bocznych szybko opada z nachyleniem 60 dB/ /dekadę — rys. 4-44). W dodatku wbrew pozorom realizacja takiej funkcji okna jest mało kłopotliwa w systemie cyfrowej analizy widmowej sygnału, ponieważ do algorytmu FFT, opisywanego w poprzednim podrozdziale, dysponujemy zwykle w pamięci maszyny gotową tabelą wartości funkcji kosinus. Niestety, dobre własności okna Hanninga okupione są pewnym pogorszeniem rozdzielczości częstotliwościowej prowadzonej analizy, gdyż stosując wskazane okno musimy liczyć się z poszerzeniem wszystkich pasm filtrów analizujących widmo, w wyniku czego rozdzielczość analizy krótkookresowej z wykorzystaniem okna Hanninga wynosi $\Delta f' = 1{,}5\Delta F$, gdzie $\Delta F$ jest rozdzielczością używanej metody analizy (analogowej lub cyfrowej) bez uwzględnienia wpływu okna. Wzór opisujący okno Hanninga wygodnie jest zapisywać w postaci umożliwiającej uniknięcie podnoszenia do kwadratu:

$$h(n') = \frac{1}{2}\left[1-\cos\left(\frac{2\pi n'}{N-1}\right)\right] \quad 0 \leqslant n' \leqslant N-1 \tag{4.60}$$

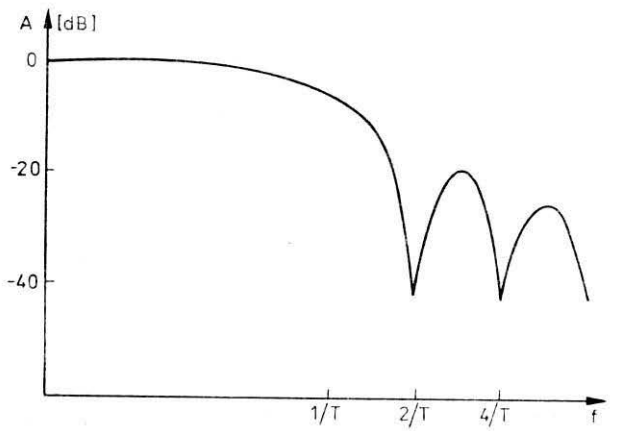
Bardzo często okno Hanninga bywa poddawane modyfikacji, polegającej na podniesieniu okna typu $\cos^2$ na prostokątny piedestał o odpowiednio dobranej wysokości. Taka kombinacja do pewnego stopnia łączy zalety okna prostokątnego i okna Hanninga: pogorszenie pasma częstotliwości jest mniejsze i wynosi tylko 40% ($\Delta f' = 1{,}4\Delta F$), dzięki zaś specyficznej interferencji listków bocznych widm pochodzących od prostokąta i $\cos^2$ następuje dalsze stłumienie poziomu pierwszego listka bocznego do poziomu $-42$ dB. Niestety, malenie dalszych listków bocznych jest przy tym mniejsze niż dla okna Hanninga i wynosi jedynie 20 dB/dekadę (rys. 4-45). Okno o omawianych własnościach nazywane jest w literaturze oknem Hamminga (uwaga na nieznaczną, ale istotną, różnicę w stosunku do poprzedniej nazwy okna Hanninga) i opisywane jest wzorem:

$$h(n') = a-(1-a)\cos\left(\frac{2\pi n'}{N-1}\right) \quad 0 \leqslant n' \leqslant N-1 \tag{4.61}$$

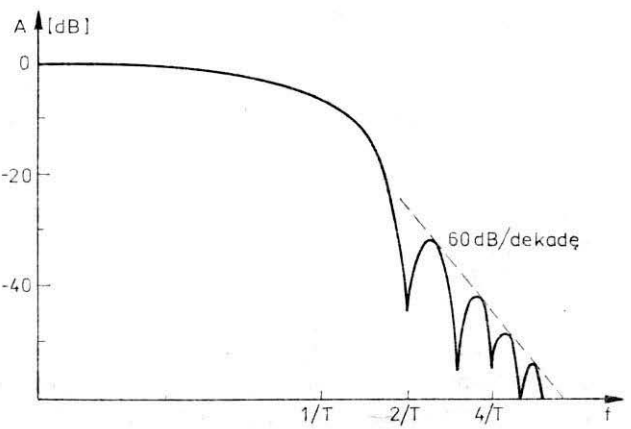
przy czym dobierając parametr $a$ można w pewnym zakresie modulować wpływ odpowiednio składnika prostokątnego i funkcji typu $\cos^2$. Przyjmuje się często, że optymalna wartość $a = 0{,}54$, można jednak prowadzić badania z różnymi wartościami wskazanego parametru.




4-42. Widmo okna prostokątnego (niekorzystny wpływ na wyniki analizy mają widoczne listki boczne)




4-43. Widmo okna Bartletta




4-44. Widmo okna Hanninga (korzystne jest tu szybkie malenie listków bocznych)

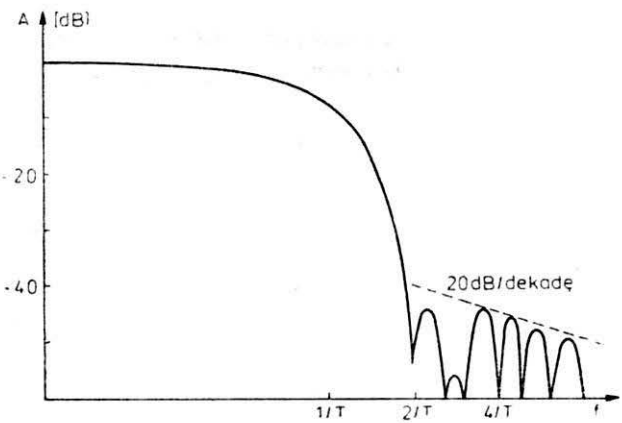
Jeszcze więcej zalet w stosunku do okna Hamminga mają: okna Blackmana;

$$h(n') = 0{,}42 - 0{,}5\cos\left(\frac{2\pi n'}{N-1}\right) + 0{,}08\cos\left(\frac{4\pi n'}{N-1}\right) \tag{4.62}$$

oraz okno Kaisera;

$$h(n') = \frac{I_0\left[a\sqrt{\left(\frac{N-1}{2}\right)^2 - \left(n' - \frac{N-1}{2}\right)^2}\right]}{I_0\left[a\frac{N-1}{2}\right]} \tag{4.63}$$

pozwalające dzięki doborowi parametru $a$ dość swobodnie wymieniać szerokość pasma na amplitudę listków bocznych i na odwrót. Funkcja $I_0$ to zmodyfikowana funkcja Bessela pierwszego rodzaju zerowego rzędu.



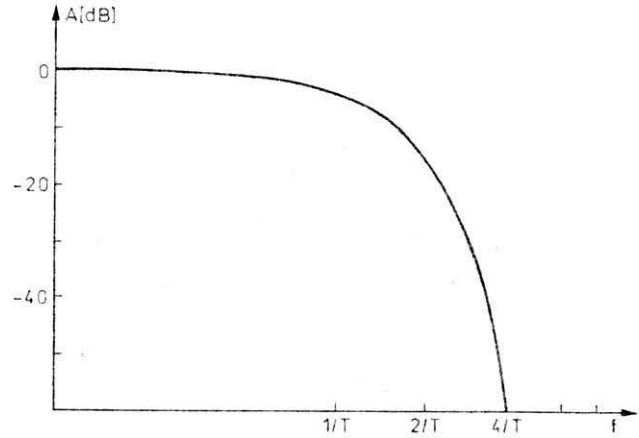

4-45. Widmo okna Hamminga (pierwszy i drugi listek boczny są silniej stłumione, niż w oknie Hanninga, dalsze jednak maleją wolniej)

Najciekawsze własności ma jednak okno Gaussa:

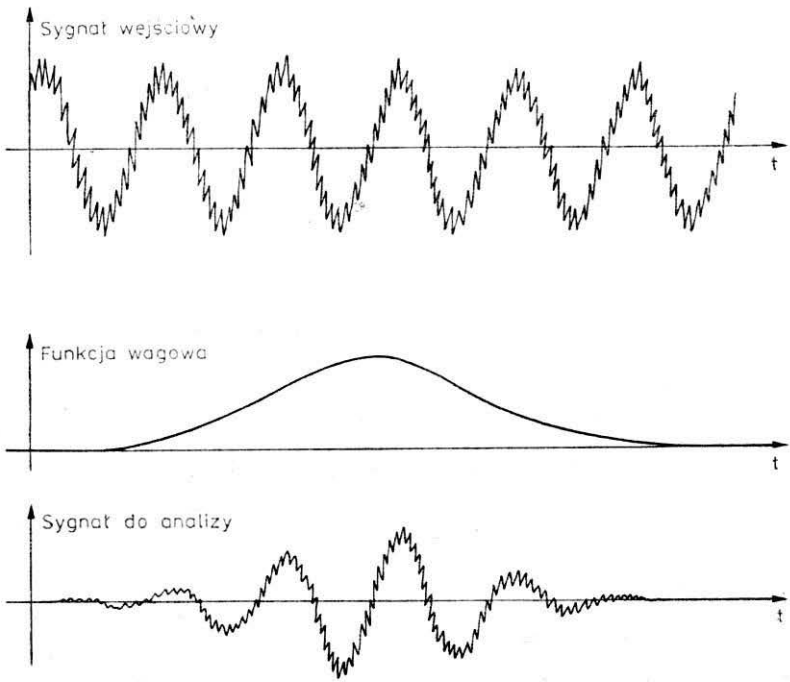
$$h(n') = \frac{1}{\sqrt{2\pi}0{,}14N} e^{-\frac{(n')^2}{0{,}04N^2}} \tag{4.64}$$

Okno tej postaci w ogóle nie wprowadza listków bocznych, gdyż widmo okna Gaussa jest — w skali logarytmicznej — parabolą o ramionach opadających ze wzrastającą stromością (rys. 4-46). Jednak poszerzenie pasma przez to okno jest największe i wynosi 90%. W sumie okno Gaussa jest rzadziej stosowane w analizie mowy niż zasługuje ze względu na swoje zalety. Być może przyczyną jest fakt rutynowego (na przykład w sensie wbudowania w aparaturę przetwarzającą różnych firm) stosowania okien Hanninga lub Hamminga.

W odniesieniu do wszystkich okien czasowych warto odnotować jedną wspólną cechę, stanowiącą istotny czynnik przy metrologii parametrów widmowych różnych sygnałów. Otóż poza oknem prostokątnym, które przenosi do analizy widmowej pełną moc wejściowego sygnału, wszystkie inne okna wycinają jedynie kawałek sygnału, tłumiąc jego fragmenty w po-




4-46. Widmo okna Gaussa całkowicie pozbawione jest listków bocznych



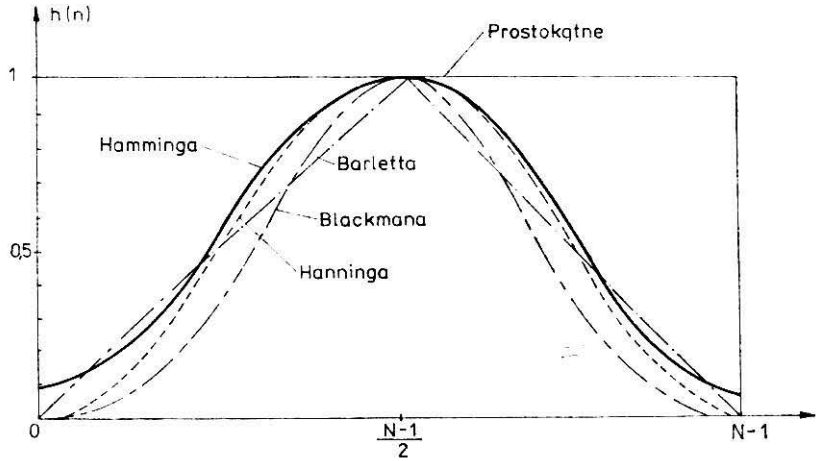

4-47. Ilustracja procesu ważenia czasowego: wejściowy sygnał dźwiękowy (u góry) przemnażany jest przez funkcję wagi (w prezentowanym przypadku jest to funkcja Gaussa pokazana na środku rysunku) w wyniku otrzymuje się sygnał podlegający analizie (w postaci pokazanej na dole)

bliżu końców okna. Zilustrowano to na rys. 4-47 dla okna gaussowskiego, ale porównanie kształtów różnych okien czasowych (rys. 4-48) pozwala stwierdzić, że problem ten występować będzie przy dowolnej funkcji okna. W przypadku analizy mowy i jej rozpoznawania nie ma to istotnego znaczenia, gdyż proporcje pomiędzy składowymi o różnych częstotliwościach nie ulegają przy tym zmianie i kształt widma — będący najczęściej obiektem zainteresowania i podstawą do rozpoznawania — nie ulega zmianie. Jednak

dla porządku należy wskazać, że wprowadzając okno czasowe wprowadzamy także tłumienie sygnału w stopniu możliwym do wyznaczenia z równania

$$P = \frac{1}{N} \sum_{n'=0}^{N-1} [h(n')]^2 \tag{4.65}$$

Przykładowo, dla okna Gaussa $P = 0{,}25$, co oznacza, że analizowany sygnał na skutek „ważenia" go oknem czasowym ma obniżony poziom o około 6 dB.




4-48. Porównanie kształtów omawianych okien czasowych wskazuje na ich duże podobieństwo — z wyjątkiem okna prostokątnego, mającego w rezultacie najmniej korzystne własności

Sposób obliczania widma dynamicznego wynika ze wzoru definicyjnego (4.55), możliwe są tu jednak pewne modyfikacje, o których warto wspomnieć. Wprowadzając oznaczenie $*$ dla operacji splotu:

$$a(n) * b(n) = \sum_{l=-\infty}^{\infty} a(l)b(n-1) \tag{4.66}$$

można równość (4.55) zapisać w postaci

$$G(k, n) = [g(n)e^{-j2\pi kn}] * h(n) \tag{4.67}$$

co odpowiada schematowi pokazanemu na rys. 4-49. Na schemacie tym sygnał wejściowy $g(n)$ jest mnożony przez czynnik zależny od częstotliwości $e^{-j2\pi kn}$, po czym podlega filtracji dolnoprzepustowej (filtr dolnoprzepustowy o odpowiedzi impulsowej $h(n)$). Układ taki jest przydatny do kolejnego wyznaczania przebiegów sygnału w poszczególnych pasmach częstotliwości.

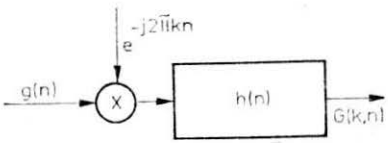
Możliwy jest także odmienny zapis i odmienny schemat. Zapisując zależność (4.55) w postaci:

$$G(k, n) = e^{-j2\pi kn} \sum_{l=-\infty}^{\infty} g(l)h(n-1)e^{j2\pi k(n-1)} \tag{4.68}$$

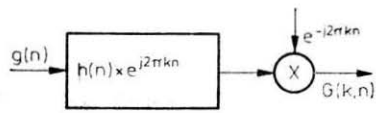
można otrzymać następującą funkcję splotową:

$$G(k, n) = e^{-j2\pi kn}\{g(n) * [h(n)e^{j2\pi kn}]\} \tag{4.69}$$

co odpowiada schematowi pokazanemu na rys. 4-50. Sygnał $g(n)$ jest wówczas poddawany filtracji środkowoprzepustowej (filtr środkowoprzepustowy o odpowiedzi impulsowej $h(n)e^{j2\pi kn}$), co jest bardziej przydatne przy wyznaczaniu wartości widma dla wszystkich częstotliwości $k$ równocześnie.




4-49. Układ wyznaczania dynamicznego widma sygnału mowy, przydatny do kolejnego wyznaczania przebiegów w poszczególnych pasmach częstotliwości




4-50. Układ wyznaczania widma dynamicznego sygnału mowy, przydatny do wyznaczania wartości widm chwilowych dla wszystkich pasm częstotliwości równocześnie

Obliczanie krótkookresowej transformaty Fouriera można usprawnić modyfikując przytoczone wzory na jeden z dwu możliwych sposobów. Pierwszy z nich opiera się na przekształceniu wzoru (4.55) za pomocą podstawienia nowej zmiennej $l = l' + n$.
Wówczas:

$$G(k, n) = \sum_{l'=-\infty}^{\infty} g(l'+n)h(-l')e^{-\frac{j2\pi k(l'+n)}{N}} =$$

$$= e^{-\frac{j2\pi kn}{N}} \sum_{l'=-\infty}^{\infty} g(l'+n)h(-l')e^{-\frac{j\pi l'k}{N}}$$

Zastępując sumowanie w przedziale nieskończonym sumowaniami cząstkowymi w przedziałach niezerowych wartości funkcji okna (o długości $N$) otrzymuje się kolejno:

$$G(k, n) = e^{-\frac{j2\pi kn}{N}} \sum_{m=-\infty}^{\infty} \sum_{l'=mN}^{mN+N-1} g(l'+n)h(-l)e^{-\frac{j2\pi l'k}{N}} =$$

$$= e^{-\frac{j2\pi kn}{N}} \sum_{l'=0}^{N-1} \tilde{g}(l', n)e^{-\frac{j2\pi l'k}{N}} \tag{4.71}$$

gdzie suma w ostatniej wprowadzonej postaci zbioru jest $N$-punktową dyskretną transformatą Fouriera. Może ona być wyliczona za pomocą omówionego wyżej algorytmu FFT dla zmodyfikowanych ciągów $\tilde{g}(l', n)$ opisanych splotem

$$\tilde{g}(l', n) = \sum_{m=-\infty}^{\infty} s(n+l'+mN)h(-l'-mN) \tag{4.72}$$

obliczanym dla $l' = 0, 1, \ldots, N-1$. Warto zauważyć, że wzór (4.72) odwo-

łuje się do sumowania w skończonym przedziale ze względu na własności funkcji okna $h(n)$.

Inne korzystne przekształcenie wzoru (4.15), ułatwiające otrzymanie widma dynamicznego sygnału z wykorzystaniem tzw. świergotowej transformaty $Z$, otrzymuje się po podstawieniu $kl = \frac{k^2}{2} + \frac{l^2}{2} - \frac{(k-1)^2}{2}$:

$$G(k, n) = e^{-\frac{j\pi k^2}{N}} [\hat{g}(l, n) * e^{\frac{j\pi l^2}{N}}] \tag{4.73}$$

gdzie:

$$\hat{g}(l, n) = g(l)h(n-l)e^{-\frac{j\pi l^2}{N}} \tag{4.74}$$

Zasadnicza zaleta wzoru (4.73) w stosunku do wcześniej wprowadzonych polega na tym, że zakłada ona stosowanie filtru o skończonej odpowiedzi impulsowej, której postać jest fragmentem ciągu zespolonego $e^{j\pi l^2/N}$, a filtry tego typu można realizować w szczególnie dogodny sposób.

## 4.4. Parametryczny opis sygnału mowy

W przytoczonych uprzednio rozważaniach uzasadniono tezę, że najbardziej przydatne do analizy mowy jest widmo dynamiczne $G(k, n)$, nazwane także przebiegiem czasowo-częstotliwościowym sygnału, spektrogramem dynamicznym lub wideogramem. Widmo to zawiera jednak bardzo wiele szczegółów, co łatwo zauważyć na rys. 4-35÷4-40. Tak duża liczba szczegółów utrudnia interpretację zapisu przy analizie sygnału, porównywanie z wzorcami przy jego rozpoznawaniu oraz badanie skutków zakłóceń i zniekształceń przy jego przesyłaniu. We wszystkich omówionych przypadkach celowe jest posługiwanie się opisem sygnału mowy i jego zmienności w kategoriach pewnych wybranych parametrów. Przy starannie wybranych parametrach możliwe jest pogodzenie dwu — z pozoru wykluczających się — wymagań: maksymalnej zwartości opisu i zachowania wszystkich, niezbędnych w ustalonym zastosowaniu, szczegółów rozważanego sygnału. Parametrów stosowanych przy analizie i rozpoznawaniu mowy jest wiele, o niektórych z nich będzie jeszcze dodatkowo mowa w kolejnym rozdziale; wszelako szerokie wykorzystanie i powszechne uznanie zdobyły tylko niektóre z nich i dlatego o nich głównie będzie mowa w tym rozdziale. Znaczna część wymienianych parametrów wynika z charakterystyk amplitudowo-częstotliwościowych sygnału i dlatego bywa nazywana p a r a m e t r a m i w i d m o w y m i. Liczne z nich uwzględniają też czasową zmienność widma sygnału i dlatego nazywane bywają p a r a m e t r a m i w i d m o w o - c z a s o w y m i. Są wreszcie i takie, które odwołują się do dziedziny czasu, ale odtwarzanej z widma i dla tych najtrudniej znaleźć wspólną nazwę, będziemy więc mówili o p a r a m e t r a c h k o r e l a c y j n y c h i c e p s t r a l n y c h nie eksponując wspólnej nazwy ich wewnętrznego podobieństwa.

Zaczniemy od parametrów widmowych. Mając określone — dyskretne dla ustalenia uwagi — widmo sygnału mowy $G(k)$ możemy skoncentrować analizę na parametrach opisujących jego kształt. Łatwe do obliczenia i — jak się okazuje — bardzo przydatne w analizie są m o m e n t y w i d m o w e. Moment $m$-tego rzędu określić można ogólnie jako

$$M(m) = \sum_{k=0}^{\infty} |G(k)| \; [f_k]^m \tag{4.75}$$

gdzie $f_k$ jest częstotliwością środkową $k$-tego pasma wyróżnionego w analizie częstotliwościowej. Przyjmuje się przy tym, że $f_0 = 0$ (gdyż $G(0)$ oznacza składową stałą sygnału), dla stałych zaś szerokości pasm analizy, wynoszących $\Delta f$, wartości $f_k$ mogą być wyliczane ze wzoru

$$f_k = (k-1)\Delta f + \frac{\Delta f}{2} \tag{4.76}$$

Z momentów widmowych opisanych wzorem (4.75) najistotniejsze znaczenie ma moment zerowego rzędu $M(0)$*), wykorzystywany do normalizacji momentów wyższych rzędów. Do interpretacji wygodniejsze są bowiem momenty unormowane

$$M_u(m) = \frac{M(m)}{M(0)} \tag{4.77}$$

Przykładowo moment unormowany pierwszego rzędu może być interpretowany jako „środek ciężkości" widma, czyli częstotliwość reprezentująca — przy unimodalnym (jednogarbnym) widmie — wierzchołek widma, a przy widmie zawierającym wiele składowych — ich średnią ważoną. Unormowany moment pierwszego rzędu odgrywa w analizie mowy znaczącą rolę i można wykazać jego przydatność w zadaniach rozpoznawania — szczególnie głosek szumowych. Mniej przydatne są momenty — nawet unormowane — wyższych rzędów, gdyż są one w oczywisty sposób skorelowane ze sobą, a ponadto, w odróżnieniu od momentu pierwszego rzędu, nie mają przekonywającej interpretacji. Aby uzyskać użyteczne i wnoszące istotnie nowe elementy parametry widmowe, trzeba sięgnąć do momentów centralnych — unormowanych lub nie. Przykładowo, użyteczny w analizach jest centralny unormowany moment drugiego rzędu, reprezentujący kwadrat „szerokości" widma. Ogólnie unormowany moment centralny $m$-tego rzędu może być wyliczany ze wzoru:

$$M_{cu}(m) = \sum_{k=0}^{\infty} |G(k)| [f_k - M_u(1)]^m / M(0) \tag{4.78}$$

i użyteczne bywają — obok wspomnianego $M_{cu}(2)$ także momenty rzędów 3, 4 i 5, pozwalające opisywać różne typowe deformacje struktury widma $G(k)$. Najbardziej przydatne są momenty $M_u(1)$ oraz $M_{cu}(2)$. Ich wzory

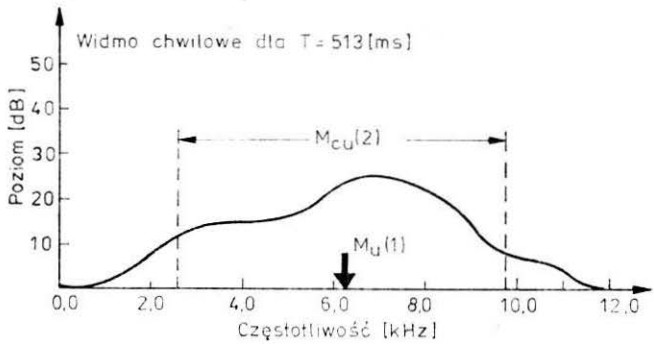
*) Przy obliczaniu momentu zerowego pojawia się problem związany z częstotliwością $f_0$. Będziemy zakładać, że $[f_0]^\circ = 1$, podobnie jak dla wszystkich innych $f_k$.

definicyjne — wynikające z przytoczonych ogólnych formuł — są następujące:

$$M_u(1) = \frac{\sum_{k=0}^{\infty} |G(k)| f_k}{\sum_{k=0}^{\infty} |G(k)|} \tag{4.79}$$

$$M_{cu}(2) = \frac{\sum_{k=0}^{\infty} G(k)\left[f_k - \sum_{\nu=0}^{\infty} |G(\nu)| f_\nu / \sum_{\mu=0}^{\infty} |G(\mu)|\right]}{\sum_{k=0}^{\infty} |G(k)|} \tag{4.80}$$

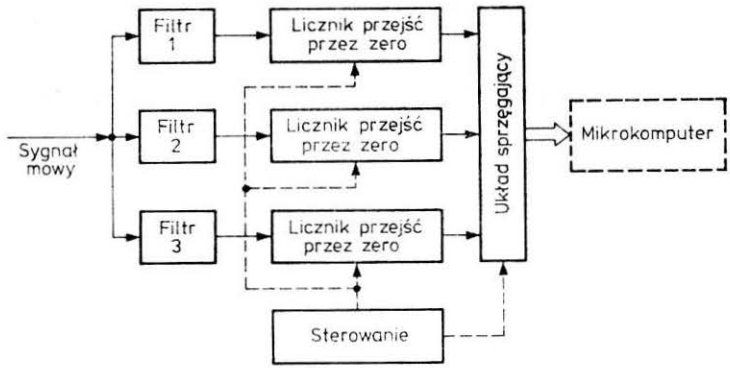
Interpretacja wprowadzonych parametrów jest (rys. 4-51), jak wspomniano, bardzo prosta. $M_u(1)$ można utożsamiać ze średnią (ważoną) częstotliwością widma, a $M_{cu}(2)$ oznacza kwadrat unormowanej szerokości widma.




4-51. Interpretacja dwu wybranych momentów widmowych jako parametrów charakteryzujących generalne własności widma: jego położenie oraz szerokość

Parametry te mogą być wyznaczane dla całego widma, mogą również dotyczyć wydzielonych jego fragmentów. W ten sposób między innymi parametry $M_u(1)$ bywają używane jako zamienniki formantów, jeśli pasma częstotliwości, w których się je wyznacza, odpowiednio ograniczone zostaną do obszarów odpowiadających występowaniu odpowiednich formantów (patrz dalej). Naturalnie korzystając z takiej definicji „przybliżonego formantu" musimy liczyć się z dużymi błędami w ich lokalizacji, w stosunku do położenia rzeczywistych formantów. Zaleta metody, jaką jest prostota pomiaru, przeważa jednak często nad względami teoretycznymi i metodyka zbliżona do omówionej tu bywa stosowana nagminnie w prostszych systemach rozpoznawania mowy dla potrzeb, na przykład, sterowania maszyn rozkazami wydawanymi głosem. Możliwe są zresztą dalsze uproszczenia. W praktycznych realizacjach przyjmuje się niekiedy, że parametr $M_u(1)$ można aproksymować z zadowalającą dokładnością przez obliczanie częstości przejść przez zero rozważanego sygnału (por. p. 4.1). Naturalnie w ten sposób popełnia się kolejne odstępstwo od założeń teoretycznych i jakość rozpoznawania, uzyskiwana takimi uproszczonymi metodami, jest bardzo niska.

Jednak prostota konstrukcji systemu rozpoznawania jest w tym przypadku atutem nie do pogardzenia: układ składa się wówczas z kilku filtrów i liczników przejść przez zero (rys. 4-52). Układ taki można z powodzeniem wykonać w warunkach domowych i użyć — na przykład — do wprowadzania parametrów sygnału mowy do komputera domowego (Sinclair ZX Spectrum C-64, Atari czy nowocześniejszego). Doświadczenia licznych



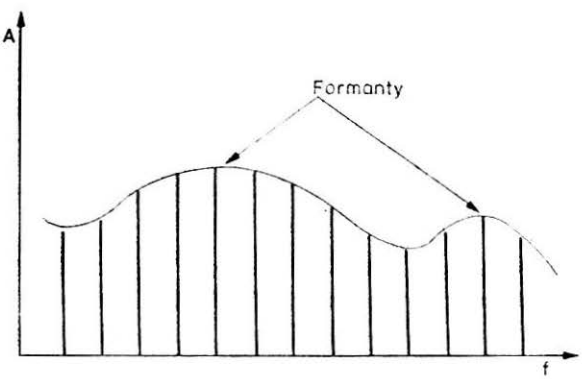

4-52. Skrajnie uproszczony układ wprowadzania parametrów sygnału mowy do maszyny cyfrowej. Filtry nastrojone są na pasma odpowiadające zakresom częstotliwości występowania formantów, a liczniki przejść przez zero określają (od razu w formie cyfrowej) częstości formantów. Układ może być przydatny przy próbach rozpoznawania mowy z wykorzystaniem domowych komputerów

amatorów potwierdzają przydatność takiego układu do rozpoznawania prostych elementów mowy polskiej. Bez trudu można za jego pomocą identyfikować samogłoski, niektóre spółgłoski, proste wyrazy, komendy itp. Nie jest to oczywiście jeszcze system rozpoznawania mowy z prawdziwego zdarzenia, ale jako pomoc dydaktyczna może być nieoceniony, a jako uzupełnienie popularnego komputera osobistego stanowi atrakcję dla tysięcy entuzjastów domowej informatyki.

Wracając do zasadniczego toku wykładu można tytułem uzupełnienia dodać, że również pozostałe parametry wywodzące się z momentów widmowych znajdują zastosowanie w analizie i rozpoznawaniu mowy. Przykładowo parametr $M(0)$ wyznaczany w odpowiednich pasmach częstotliwości może być użyteczny, na przykład, do wykrywania różnicy pomiędzy głoskami dźwięcznymi i bezdźwięcznymi (w głoskach dźwięcznych i tylko w nich występuje obszar koncentracji energii w zakresie niskich częstotliwości (około 100 Hz), co jest związane z występowaniem podstawowej harmonicznej tonu krtaniowego). Wiele innych głosek można rozróżniać, badając stosunki zawartości energii w wybranych pasmach widma — a do tego nadaje się doskonale parametr $M(0)$. Inne momenty widmowe znajdują także ciekawe zastosowania i mają liczne możliwości, jednak niewielu badaczy sięga do tych parametrów i wykorzystuje je w swoich pracach.

Często wykorzystywane są natomiast formanty. Jest to podstawowa grupa parametrów, używana do analizy, przetwarzania i rozpoznawania mowy przez praktycznie wszystkich badaczy zajmujących się problematyką mowy. Wynika to z wielu faktów. Zacząć jednak wypada od definicji formantu,

aby dokładnie wiedzieć, o czym właściwie jest mowa. Jak wynika z rozważań przytoczonych w rozdziale 2, proces artykulacji mowy to świadome kształtowanie obwiedni amplitudowo-częstotliwościowej dźwięku generowanego — (tonu krtaniowego lub/i szumu) za pomocą celowych ruchów języka, żuchwy, warg i podniebienia. Z oczywistych powodów w centrum zainteresowania większości badaczy znajduje się głównie proces formowania dźwięku mowy, a nie — mniej lub bardziej przypadkowy — proces fonacji (generacji tonu). Proces kształtowania artykułowanego sygnału polega — w dużym uproszczeniu to formułując — na tworzeniu struktury dynamicznej traktu głosowego. Transmitancja modelu matematycznego ma wiele biegunów, uwidoczniających się w widmie w postaci maksimów jego obwiedni. Właśnie te maksima nazywa się f o r m a n t a m i, a częstotliwości, przy których występują — c z ę s t o t l i w o ś c i a m i f o r m a n t o w y m i (rys. 4-53).



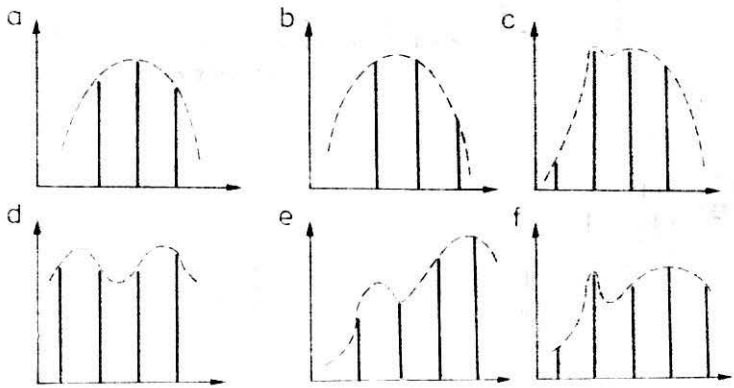

4-53. Ilustracja pojęcia formantu. Widmo ma charakter dyskretny (prążkowy), zatem mówiąc o maksimum lokalnym mamy na myśli maksimum jego obwiedni. Właśnie takie lokalne maksimum obwiedni widma sygnału mowy nazywa się formantem, a częstotliwość, przy której występuje — częstotliwością formantową

Ruchy narządów mowy zmieniają rozmiary i proporcje tworzących się wnęk rezonansowych, wobec tego formanty zmieniają swoje położenie, pojawiają się, znikają, zmienia się ich liczba, wielkość i lokalizacja — a badacz śledząc te zmiany może bardzo dużo powiedzieć o procesie artykulacji, a tym samym o sygnale mowy. W szczególności, na podstawie analizy formantów można określić: co jest mówione (aspekt semantyczny), kto mówi (aspekt osobniczy) i jak mówi (aspekt badawczy, m.in. medyczny). Główna zaleta formantów polega na ich charakterystycznej konfiguracji, możliwej do określenia w charakterze wzorca dla większości głosek (w tym głównie samogłosek) — niezależnie od tego, kto je wypowiada, jak szybki jest proces artykulacji, jakie towarzyszą mu emocje itp. W związku z tą cechą formanty interesują głównie inżynierów łączności, którzy słusznie przypuszczają, że przesyłając w łączu telefonicznym wyłącznie informacje o lokalizacji formantów można uzyskać w urządzeniu odbiorczym zrozumiały sygnał mowy — przy ponad 70% oszczędności objętości przesyłanego sygnału (patrz p. 6.2).

Formanty wydają się również nader interesujące z punktu widzenia procesu automatycznego rozpoznawania mowy. Ich względna stabilność osobnicza przy małej objętości informacyjnej czyni zadość wymaganiom, jakie stawia

się typowo parametrom, na których opiera się proces rozpoznawania — bliższe szczegóły p. 5.2.

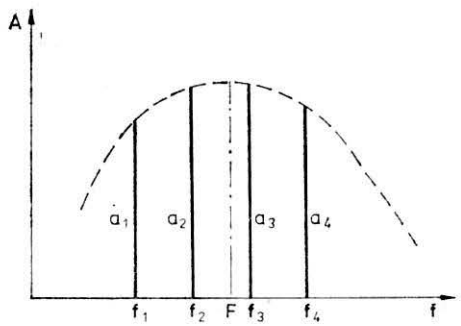
W tym rozdziale, mając świadomość użyteczności formantów, skupimy uwagę na bliższym ich określeniu oraz na problemie skutecznych metod ich wyznaczania. Podana opisowa definicja formantu nie zawsze może być wystarczająca, gdyż pojęcie obwiedni widma, które jest w niej użyte, nie jest wygodne w rozważaniach praktycznych. Rozważmy w szczególności przypadek głosek dźwięcznych. Ich widmo — co wynika z quasi-periodycznego przebiegu sygnału — ma charakter dyskretny. Pojawiają się w nim prążki odległe o wartość, będącą częstotliwością tonu krtaniowego — a więc typowo około 100 Hz. Efekt nałożenia na dyskretne widmo obwiedni zawierającej maksimum jest silnie uzależniony od wzajemnego położenia maksimum i prążków widma, co pokazano na kilku przykładach na rys. 4-54. Na drodze pomiarów możemy (w najlepszym przypadku) określić amplitudy prążków. Natomiast odtworzenie na tej podstawie obwiedni jest




4-54. Ze względu na dyskretny charakter widma położenie maksimum obwiedni może być trudne do zlokalizowania. Pokazano kilka możliwych przypadków wzajemnego ułożenia obwiedni i prążków widma: a — lokalizacja formantu jest łatwa, gdyż wskazuje go wprost prążek o maksymalnej amplitudzie, b — odpowiada „obramowaniu" maksimum przez prążki widma, których wysokość może być jednakowa lub różnić się bardzo mało, c — wysokość dwu prążków jest jednakowa, a odpowiadają one dwu formantom, d — odpowiada połączeniu sytuacji (c) i (b), e — prowadzi do nieuchronnego zgubienia formantu na skutek jego „wąskiej" obwiedni w porównaniu z odstępami prążków, f — pokazuje, że nawet przy bardzo „wąskich" formantach możliwa bywa ich lokalizacja

trudne, nawet bardzo trudne. Można wprawdzie próbować aproksymować przebieg obwiedni założoną postacią funkcji. Przykładowo biorąc pod uwagę trzy prążki można aproksymować położenie wierzchołka obwiedni — a więc lokalizować częstotliwość formantu — zakładając, że przebieg obwiedni jest parabolą. Odpowiedni wzór i jego interpretację podano na rys. 4-55. Nie jest to jednak w istocie rozwiązanie problemu, gdyż wcale niełatwo w praktyce określić, które trzy prążki brać pod uwagę. Dodatkowa trudność wynika przy tym z faktu, że sygnał mowy może zawierać wiele formantów. W badaniach szczegółowych mówi się o przynajmniej pięciu formantach, niektórzy badacze dopatrują się nawet siedmiu. Nawet ograniczając rozważania do trzech podstawowych formantów, najczęściej rozważanych w praktyce, nie unikniemy trudności wynikającej z faktu, że dla niektórych głosek typowe położenie formantów zakłada ich wzajemne od-

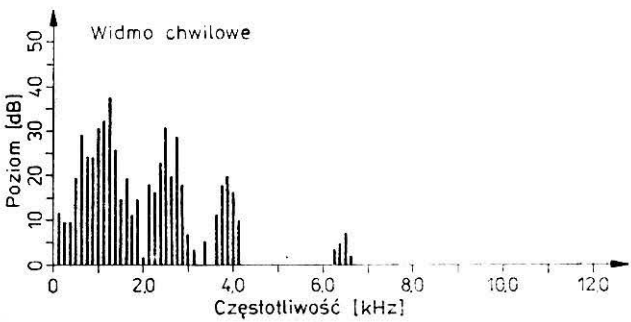
ległości nie przekraczające kilkuset herców, przy odstępie prążków dyskretnego widma wynoszącym ponad 100 Hz. Kolejny problem w wyliczaniu formantów wynika z powodu zakłóceń zniekształcających idealny obrys badanego widma. Jakąkolwiek metodę przyjmie się za podstawę przy określaniu struktury widma chwilowego sygnału mowy — otrzymane widmo będzie zawierało drobne, ale praktycznie nieuniknione deformacje przebiegu.



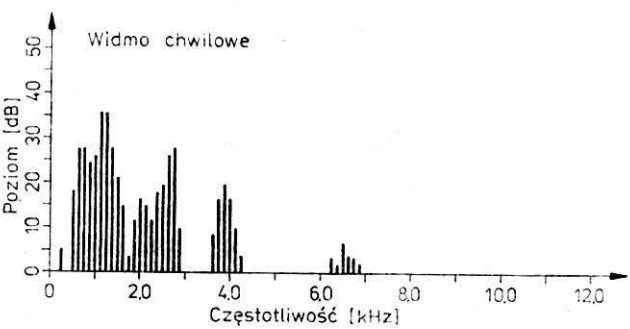

$$\hat{F} = \frac{a_1 f_1 + a_2 f_2 + a_3 f_3}{a_1 + a_2 + a_3}$$

$$\tilde{F} = \frac{a_2 f_2 + a_3 f_3 + a_4 f_4}{a_2 + a_3 + a_4}$$

4-55. Lokalizacja precyzyjnej wartości częstości formantowej $F$ jest na ogół trudna. Dysponując wartościami amplitud dyskretnych wartości widma $a_1, a_2, a_3, a_4$ oraz odpowiadającymi im wartościami częstotliwości $f_1, f_2, f_3, f_4$ możemy wyliczać wartość przybliżoną $\hat{F}$ lub $\tilde{F}$, przy czym wybór jednej z tych wartości bywa trudny




4-56. Widmo chwilowe głoski *a* — na podstawie takiego widma wyznacza się formanty

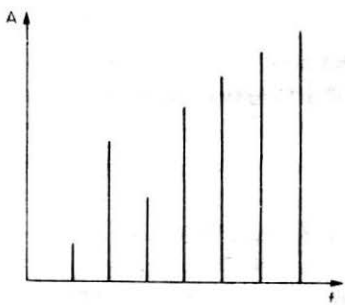



4-57. Widmo chwilowe głoski *a* w innym momencie czasu niż w przypadku opisanym na rys. 4-56. Widać, że lokalizacja formantów będzie w tym przypadku inna niż na poprzednim rysunku, chociaż oba widma pochodzą ze stanu ustalonego głoski *a* w tej samej wypowiedzi i powinny mieć identyczne parametry

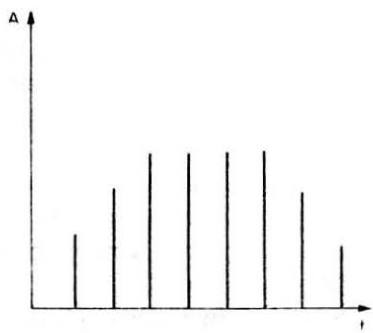
Jeśli dodatkowo uwzględni się fakt, że sygnałowi mowy może towarzyszyć mnóstwo trudnych do usunięcia zakłóceń — problem pojawi się w całej swojej ostrości. Najlepiej ilustrują to konkretne przykłady. Na rysunkach 4-56 i 4-57 pokazano dwie próbki krótkookresowego widma głoski *a*. Obie

próbki pochodzą z tej samej wypowiedzi i odpowiadają tzw. stanowi ustalonemu w wygłosie głoski — zatem powinny być identyczne, a przynajmniej ich formanty powinny być zlokalizowane w przybliżeniu w tym samym miejscu. Tymczasem nawet przy pobieżnym porównaniu obu rysunków powstają wątpliwości, przy czym dla uniknięcia niepożądanych efektów próbka sygnału mowy, którą analizowano, nagrywana była z wykorzystaniem najwyższej jakości sprzętu profesjonalnego w warunkach absolutnego braku zakłóceń — w komorze bezechowej, gdzie dla częstości powyżej 16 Hz poziom ciśnienia akustycznego szumów wynosi praktycznie 0 dB. Zatem nawet w sterylnych warunkach akustycznych i braniu pod uwagę najprostszej formy sygnału — stanu ustalonego samogłosek — istnieją trudności w lokalizacji formantów. Jakich kłopotów należy więc oczekiwać przy lokalizacji maksimów obwiedni w szybkozmiennych partiach widma sygnału, który w dodatku może być zniekształcony przez szumy?

Istnieje jeszcze jeden czynnik, do tej pory nie brany pod uwagę, a bardzo przydatny w analizie. Formanty, jako efekt ruchów artykulacyjnych narządów mowy, nie mogą zmieniać się zbyt szybko. Na tym zresztą w dużym stopniu opiera się ich użyteczność. Można więc wykorzystać do śledzenia przebiegu formantów związki między kolejnymi widmami. Pewna lokalizacja formantu w ustalonym widmie chwilowym, odpowiadającym momentowi czasu zlokalizowanemu wewnątrz rozpatrywanej wypowiedzi, może




4-58. Przypadkowo większa wartość prążka na „zboczu" widma na ogół nie musi wskazywać na obecność formantu




4-59. Nawet rozległe „plateau" widma może zawierać informację o lokalizacji formantu, jeśli wcześniej i później znajdą się zbocza tworzące łącznie wyraźne maksimum lokalne widma

znacznie ułatwić lokalizację tegoż formantu w momentach czasu poprzedzających rozważany lub następujących po nim. W rezultacie urządzenie lub — częściej obecnie — algorytm lokalizujący formanty śledzi ich powolne zmiany i w ten sposób łatwiej „wyławia" je spośród zakłóceń i szumów. Na koniec wreszcie do lokalizacji formantów można użyć algorytmów wykorzystujących nie tylko lokalne, ale także globalne właściwości widma. Nie będzie więc wykryte jako formant pojedyncze maksimum widma położone na „zboczu" (rys. 4-58), gdyż jest wielce prawdopodobne, że jest to typowy artefakt, skutek zakłóceń lub efekt uboczny — na przykład skutek nazalizacji. Będzie natomiast wykryty formant nawet w rejonie pozornie

płaskim (rys. 4-59), jeśli otoczenie wskazuje, że powinien się tam znajdować. Dodatkowym ułatwieniem przy śledzeniu formantów może być uśrednianie rozważanego sygnału — w dziedzinie czasu (uśrednianie sądsiednich widm w spektrogramie dynamicznym) lub w dziedzinie częstotliwości (uśrednianie sąsiednich pasm w każdym momencie czasu). Bardzo przydatne są też empirycznie wyznaczone i znane granice pasm, w których oczekiwać można formantów dla mowy polskiej (tabl. 1). Wykorzystanie tych granic pozwala

Tablica 1. Orientacyjne granice [Hz] pasm częstotliwości, w których mieszczą się trzy pierwsze formanty głosek mowy polskiej

| Numer formantu | 1 | 2 | 3 |
|---|---|---|---|
| Dolna częstotliwość graniczna | 200 | 850 | 2100 |
| Górna częstotliwość graniczna | 880 | 2350 | 3100 |

eliminować niektóre „pseudoformanty", pojawiające się uporczywie we wszystkich analizach rzeczywistego sygnału mowy, a także ułatwia przypisanie znalezionym formantom właściwych numerów porządkowych. Wbrew pozorom nie jest to sprawa drugorzędna, przeciwnie we wszystkich zastosowaniach istotne jest, czy mamy do czynienia z pierwszym, czy na przykład z trzecim formantem o danej częstotliwości, a zakresy pasm częstotliwości formantowych zachodzą na siebie (por. tabl. 1). Ważne jest to również i z tego powodu, że w normalnym sygnale mowy formanty nikną i ponownie się pojawiają. Wobec tego całkowicie możliwa jest sytuacja, kiedy uda się zlokalizować dwa formanty, lecz będą to formanty drugi i czwarty, pierwszego i trzeciego zaś nie będzie. Błędne ponumerowanie formantów uniemożliwi wówczas poprawne rozpoznanie odpowiedniego fragmentu wypowiedzi i bardzo utrudni ich wykorzystanie przy optymalizacji transmisji mowy.

Definicje formantów i opis metod ich lokalizacji w porównaniu z wcześniej omawianymi momentami znacznie trudniej opisać matematycznie i wydobyć z ciągłego sygnału mowy — szczególnie metodami cyfrowymi. Nie wydaje się jednak celowe przytaczanie dość zawiłych wzorów i algorytmów wyliczania formantów w całości. Zainteresowany szczegółami Czytelnik proszony jest o wykorzystanie pozycji źródłowych, wymienionych w spisie literatury. W tej książce i dla potrzeb niniejszego rozdziału zaprezentowane zostaną uproszczone ujęcia formalne i naszkicowane zostaną jedynie główne idee najskuteczniejsze — jak się wydaje — algorytmu lokalizacji formantów.

Punktem wyjścia przy określaniu formantów jest oczywiście widmo sygnału mowy, przy czym w celu wykorzystania w analizie współzależności czasowych konieczne jest przyjęcie za punkt wyjścia spektrogramu dynamicznego $G(k, n)$. Formant w punkcie o współrzędnej czasowej $n$ i współrzędnej częstotliwościowej $k$ wykrywany jest wtedy, gdy

$$G(k, n) \geqslant G(k+1, n) \wedge G(k, n) \geqslant G(k-1, n) \tag{4.81}$$

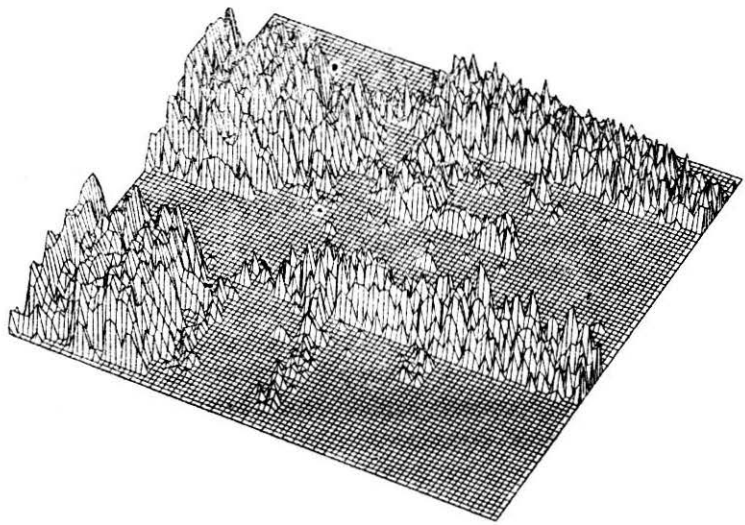
a ponadto

$$k_{\min} \leqslant k \leqslant k_{\max} \tag{4.82}$$

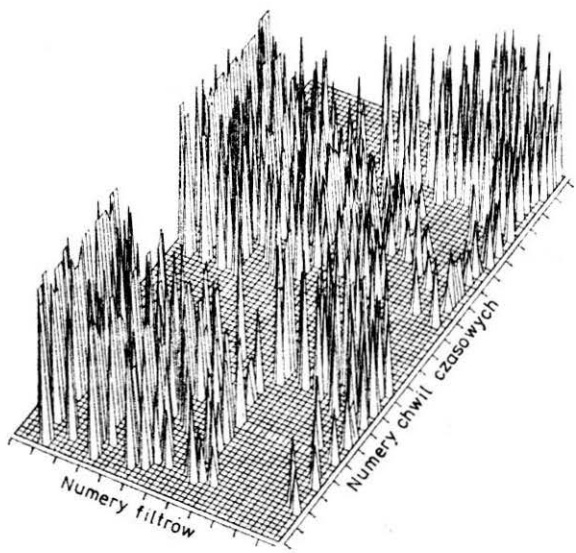
gdzie $k_{\min}$ i $k_{\max}$ są odpowiednio — w dyskretnej skali częstotliwości — wyrażonymi częstotliwościami ograniczającymi pasmo poszukiwanego formantu (por. tabl. 1). Zlokalizowany formant może być akceptowany, jeśli w poprzedniej chwili czasowej $(n-1)$ lub w następnej chwili czasowej $(n+1)$ wykryty był formant w punkcie o współrzędnej częstotliwościowej $k$ lub sąsiednich $(k-1)$ lub $(k+1)$.

Podstawowy problem polega na numeracji formantów. Jeśli wykryto w widmie chwilowym odpowiadającym chwili $n$ pewną liczbę współrzędnych częstotliwościowych spełniających przytoczone warunki, to przypisuje się pierwszy numer współrzędnej o najniższej częstotliwości, drugi kolejnemu w skali rosnących wartości $k$, i tak dalej. Sama koncepcja jest zupełnie elementarna, jednak zapis w postaci wzorów matematycznych analogicznych do (4.81) i (4.82), w tym celu, aby cały opis miał jednolitą formę, nastręcza wiele trudności. Oczywiście można je pokonać komplikując odpowiednio zapis, powstają jednak wzory, których czytelność jest bardzo mała, a użyteczność — jeszcze mniejsza. Poprzestaniemy tu więc na opisie słownym, poszerzając go o stwierdzenie, że graniczne numery pasm częstotliwości we wzorze (4.82) muszą być uzależnione od numeru poszukiwanego formantu, wobec tego kolejność postępowania przy wyznaczaniu formantów jest następująca. Najpierw lokalizuje się wszystkie częstotliwości spełniające warunek (4.81), po czym dokonuje eliminacji wykorzystując kontekst czasowy (sąsiednie widma, dla chwili $(n-1)$ oraz $(n+1)$ oraz ewentualnie szerszy kontekst częstotliwościowy. Dopiero zaakceptowane punkty sprawdza się warunkiem (4.82), poczynając od elementu odpowiadającego najmniejszej częstości, o którym roboczo zakłada się, że jest formantem nr 1, kolejno przechodząc do formantów o większych numerach. W rzeczywistych algorytmach dochodzą dodatkowe czynności, ułatwiające i przyspieszające poszukiwania. Przykładowo w opracowanym do tego celu algorytmie WRMP przebieg widma koduje się najpierw za pomocą funkcji trójwartościowej, określającej relacje między sąsiednimi „prążkami" w widmie (na zasadzie większy-równy-mniejszy, od czego zresztą pochodzi symbol metody), a dopiero potem wyznacza się położenie hipotetycznych formantów i dokonuje kolejnych sprawdzeń. Problem szybkości lokalizacji formantów (lub — stawiając zagadnienie w sposób ogólny — problem szybkości wyznaczania parametrów sygnału mowy, jakiekolwiek by te parametry były) jest ważny w kontekście warunku funkcjonowania w czasie rzeczywistym stawianego typowo systemom analizy i rozpoznawania mowy. Warunek ten oznacza, że czas wyznaczania parametrów widma nie może być dłuższy, niż czas trwania niezerowych wartości okna czasowego służącego do wyznaczania widma. Metoda WRMP daje możliwość pracy w czasie rzeczywistym, gdyż czas wykrycia formantów w widmie chwilowym nie przekracza 4 ms, podczas gdy czas między kolejnymi widmami chwilowymi wynosi w eksploatowanym systemie 9 ms.

W celu zobrazowania metody lokalizacji formantów przytoczymy teraz serię rysunków, pokazujących kolejne ich etapy wykrywania. Na rysunku 4-60 pokazano przebieg wideogramu przykładowo wybranego wyrazu. Na rysunku 4-61 pokazano skutki zastosowania reguły wykrywania każdego maksimum lokalnego w widmie i traktowania go jako formantu. Widać wyraźnie, że

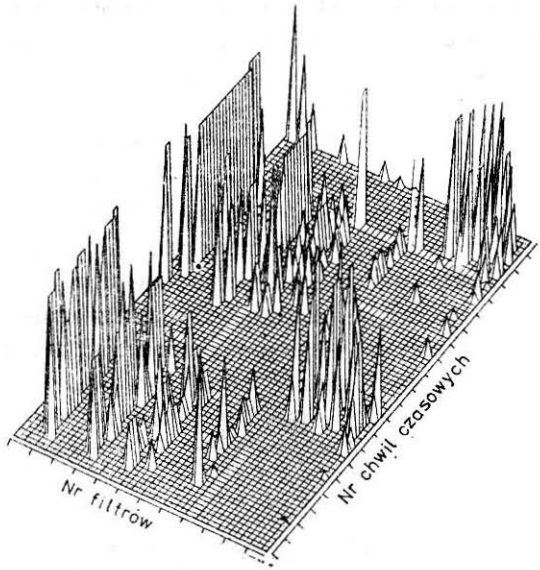


4-60. Wideogram sygnału, w którym poszukiwane są formanty (wypowiedź *serce*)

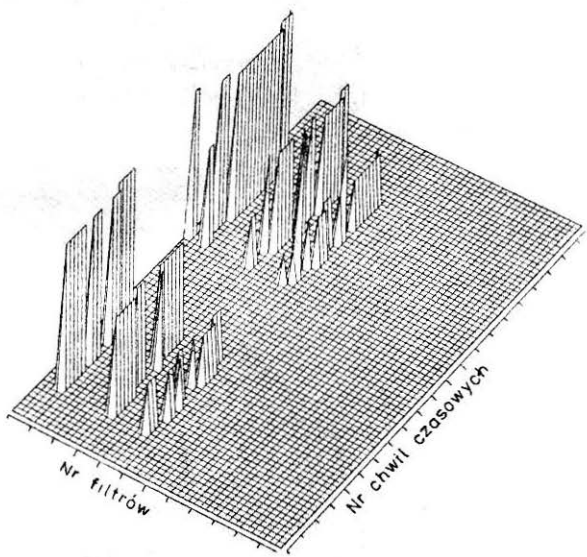



4-61. Pierwszy krok wydzielania punktów, odpowiadających prostej definicji formantu, prowadzi do wykrycia takiej liczby punktów, że tworzą one absolutnie chaotyczny, nieprzydatny do analizy obraz

nawet w oryginalnym sygnale rzeczywiste formanty rysowały się wyraźniej niż w chaosie pików sygnalizujących wszystkie lokalne maksima. Tak więc reguła dana wzorem (4.81), przytaczana niekiedy jako definicja formantu, jest w najwyższym stopniu niewystarczająca do poprawnej lokalizacji. Pewne polepszenie i uporządkowanie obrazu dają reguły pozwalające uwzględniać w lokalizacji formantu kontekst czasowy i szersze widmo, z jego



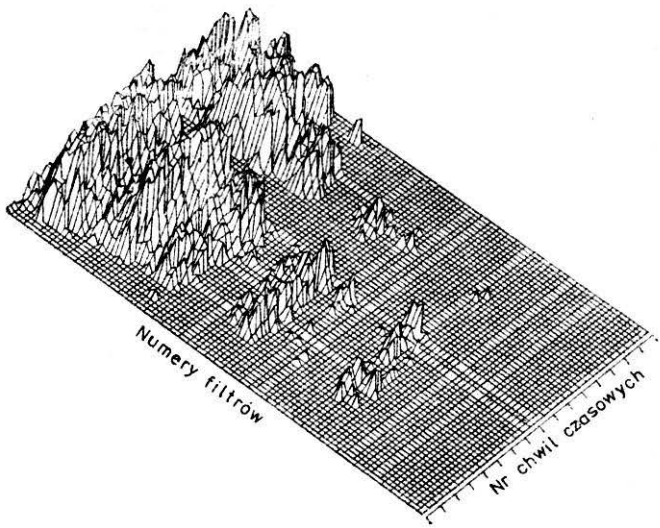

4-62. Obraz formantów staje się bardziej czytelny, gdy dokona się ich selekcji z uwzględnieniem kontekstu



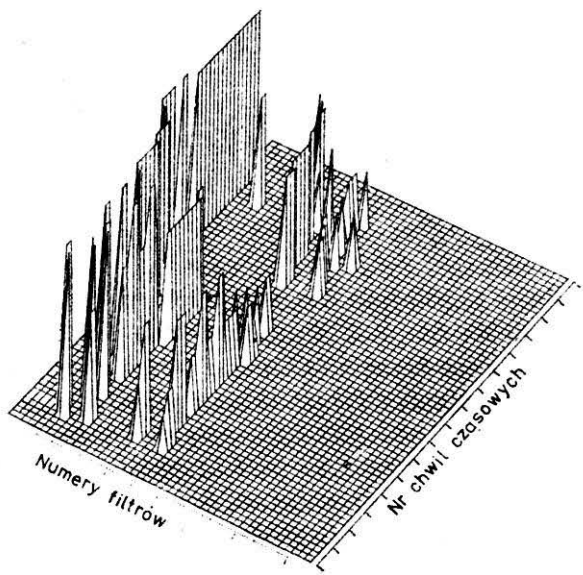

4-63. Ostateczny obraz przebiegu formantów zlokalizowanych rozważaną metodą jest klarowny i pozwala uprościć proces rozpoznawania. Obraz pokazany na tym rysunku zawiera istotnie mniej szczegółów w stosunku do obrazu źródłowego — rys. 4-60 i jest bardziej uporządkowany w stosunku do obrazu uzyskanego z bezpośredniej definicji formantu — rys. 4-61

cechami globalnymi. Wynik zastosowania tych kryteriów selekcyjnych pokazano na rys. 4-62. Ostateczne wyselekcjonowanie fragmentów widma spełniających kryteria (4.82) oraz wprowadzenie warunków czasowej kontynuacji porządkuje obraz w sposób ostateczny, co pokazano na rys. 4.63. Przejście od rys. 4-60, będącego zapisem spektrogramu dynamicznego rozważanego sygnału, do rys. 4-63, będącego obrazem czasowej zmienności wybranych parametrów sygnału (w rozważanym przypadku — formantów), obrazuje drogę radykalnej redukcji ilości informacji zawartej w rozważanym odcinku sygnału mowy.

Do zapisania w pamięci komputera w celu późniejszego rozpoznania lub do przesłania w kanale telekomunikacyjnym wygodniejsze jest podanie sygnału w formie parametrycznej, danej na rys. 4-63, niż w postaci źródłowego spektrogramu z rys. 4-60 lub — tym bardziej — źródłowego przebiegu czasowego sygnału. Można się o tym przekonać, porównując rys.4-64 i 4-65 — prezentujące zestawienie: spektrogram wyrazu i jego reprezentacja za pomocą przebiegu selekcjonowanych formantów. Podkreślmy raz jeszcze, że głównym celem i podstawowym atutem opisu sygnału mowy w formie parametrycznej jest redukcja informacji. Opis parametryczny — na przykład z wykorzystaniem formantów — zawiera znacznie mniej informacji, łatwiej więc go umieścić w systemie rozpoznającym lub przesłać na odległość.



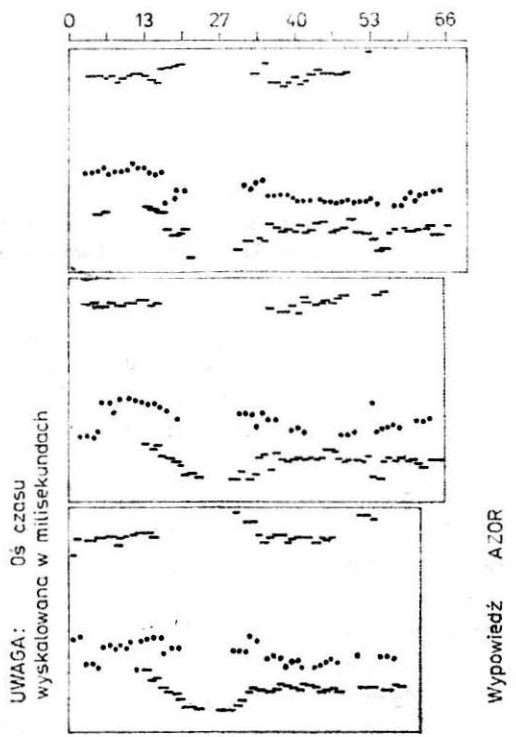

4-64. Inna wypowiedź, w której można lokalizować formanty, *ryba*




4-65. Przebieg formantów wykrytych w wyrazie *ryba*

Temu samemu celowi służą zresztą także i inne opisy sygnału mowy, wykorzystujące różne rodzaje parametrów. Podstawowy problem, jaki przy tym występuje, polega na zachowaniu w zredukowanym zestawie informacji mieszczących się w wytypowanych parametrach, wszystkich informacji istotnych z określonego, wybranego punktu widzenia. Formanty są parametrami istotnymi z punktu widzenia semantycznej treści wypowiedzi. Ich rejestrowanie pomaga w procesie rozpoznawania mowy, a ich przesyłanie pozwala na odbiorczym końcu łącza rozumieć treść nadawanego komunikatu. Z punktu widzenia innych celów analizy lub przy odmiennych celach przetwarzania mowy bardziej przydatne okazują się inne zestawy parametrów, przy czym o niektórych spośród nich będzie jeszcze mowa.

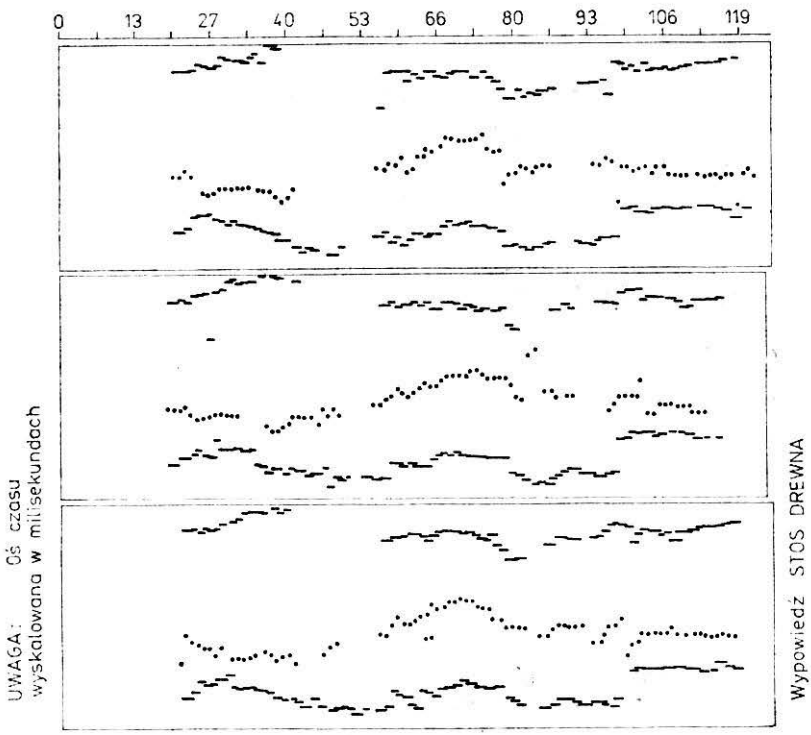
Przytaczane „trójwymiarowe" rysunki formantów służyły do oceny działania procedur wydzielających je z ciągłego sygnału mowy oraz pozwalały zorientować się w roli poszczególnych etapów procesu wydzielania. Natomiast do oceny przydatności formantów do rozpoznawania wypowiedzi oraz do oszczędnego przesyłania ich łączem telekomunikacyjnym bardziej użyteczne są „mapki", powstające przy oglądaniu płaszczyzny „czas — częstotliwość" z góry i oznaczaniu trasy zmienności formantów na tej płaszczyźnie. Na rysunku 4-66 pokazano „trajektorie" formantów dla trzech różnych wypowiedzi tego samego wyrazu. Nawet bez wnikania w szczegóły rysunku łatwo zauważyć podobieństwo kształtu zarysu zmienności formantów w poszczególnych wypowiedziach, które poza tym różniły się od siebie znacznie — nawet czasem trwania. „Mapki" te są charakterystyczne dla




4-66. „Mapa" przebiegu formantów w trzech próbkach wypowiedzi *azor*

określonej konkretnej wypowiedzi i mogą być podstawą rozpoznawania. Łatwo się o tym przekonać porównując identyczną prezentację innej wypowiedzi, pokazaną przykładowo na rys. 4-67.

Znaczenie formantów jest tak duże, że poszukiwano możliwości wyznaczania ich wartości i czasowych zmian w ciągłym sygnale mowy na drodze procesów przetwarzania informacji odległych od tradycyjnych metod przetwarzania sygnałów. Jedną z takich prób było konstruowanie tzw. sieci neuropodobnych, wykrywających i śledzących formanty na bieżąco, w cza-

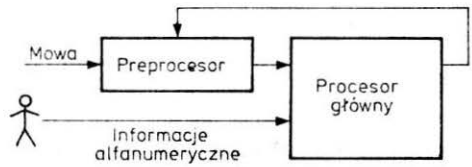



4-67. „Mapa" przebiegu formantów w trzech próbkach wypowiedzi *stos drewna*

sie rzeczywistym — bez angażowania komputera. Nie wdając się w szczegóły, które w razie potrzeby znaleźć można w podanej na końcu książki literaturze, można stwierdzić, że angażowanie dużych mocy obliczeniowych w proces wykrywania formantów jest nieracjonalne, gdyż jest to — z punktu widzenia całego systemu komputerowego — proces pomocniczy do procesu pomocniczego, czyli w sumie uboczny. Równocześnie przeprowadzona dyskusja wskazała, że operacje wydobywania formantów z ciągłego sygnału mowy są złożone i pracochłonne. Jedyny sposób na usunięcie rysującej się sprzeczności polega na „wysunięciu" procesu wydobywania formantów do procesora specjalistycznego — preprocesora obsługującego głosowe wejście do komputera w sposób nie angażujący procesora głównego maszyny.

Przyjmując takie założenie (rys. 4-68) możemy również zastanowić się nad optymalną architekturą preprocesora. Wiele względów przemawia tu za

użyciem techniki przetwarzania równoległego, odmiennej od technik stosowanych w tradycyjnej informatyce, a bliższych zasadom działania struktur nerwowych w mózgu. Jest to uzasadnione ze względu na narzucającą się, wynikającą z logiki wykonywanych operacji, równoległość procesów przetwarzania informacji w poszczególnych kanałach — odpowiadających poszczególnym pasmom częstotliwości, wydzielonym przez filtry. Jest to

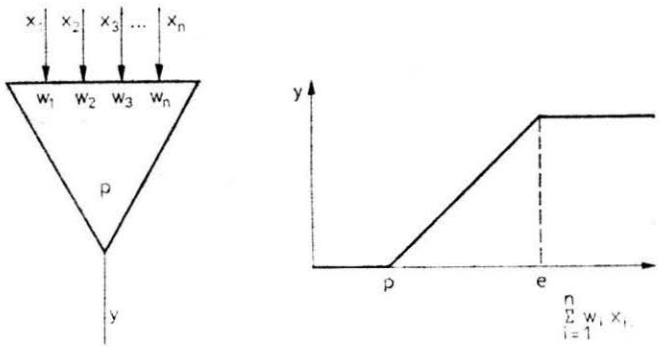



4-68. Struktura systemu komputerowego z preprocesorem do rozpoznawania i przetwarzania mowy

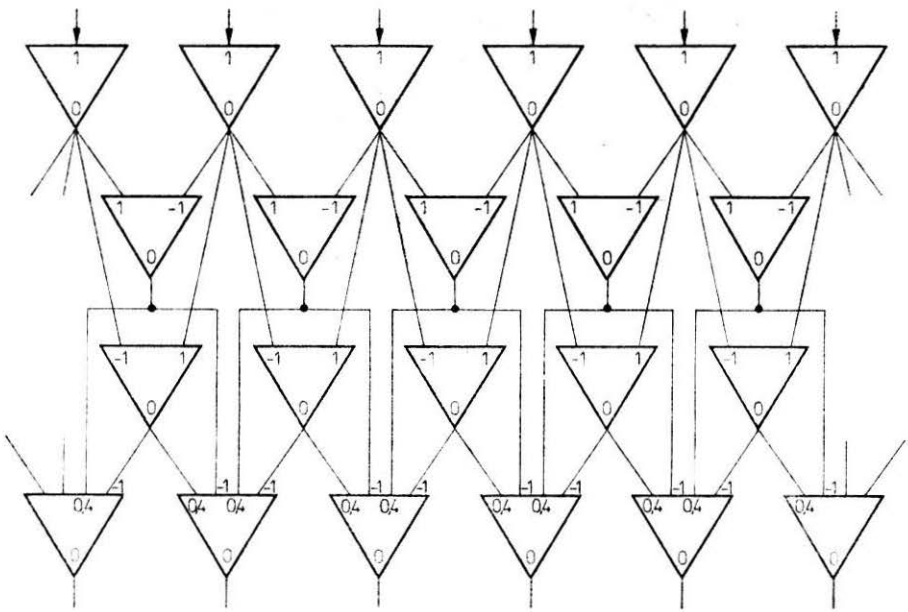
także hipotetycznie uzasadnione faktem, że w mózgu — prawdopodobnie — zachodzą procesy analizy struktur widma sygnału mowy i przypuszczalnie lokalizacja maksimów obwiedni amplitudowo-częstotliwościowej sygnału odgrywa w tym procesie poczesne miejsce. Zachodzi jedynie pytanie, w jaki sposób i z jaką dokładnością procesy te dla potrzeb technicznych modelować. Można przyjąć, że jako element przetwarzający informację może być akceptowany element progowy (rys. 4-69). W elemencie takim sygnały wejściowe $x_1, x_2, \ldots, x_n$ są mnożone przez odpowiednio dobrane współczynniki („wagi") $w_1, w_2, \ldots, w_n$, sumowane i porównywane z progiem $p$. Sygnał wyjściowy $y$ otrzymywany z takiego elementu (nazywanego elementem neuropodobnym, ze względu na swoje ograniczone analogie z rzeczywistym neuronem, budującym struktury nerwowe mózgu) można przedstawić za pomocą wzoru:

$$y = \begin{cases} 0 & \text{gdy } \sum_{i=1}^{n} w_i x_i < p \\ k\left(\sum_{i=1}^{n} w_i x_i - p\right) & \text{gdy } p \leqslant \sum w_i x_i \leqslant e \\ k(e-p) & \text{gdy } \sum_{i=1}^{n} (w_i x_i) > e \end{cases} \tag{4.83}$$

Okazuje się, że za pomocą takiego elementu można budować struktury o bardzo bogatych możliwościach w zakresie przetwarzania sygnałów, przy zachowaniu prostej realizacji technicznej i łatwej organizacji procesu przetwarzania w strukturach równoległych. Naturalnie do każdego konkretnego zastosowania należy dobrać odpowiednią strukturę połączeń rozważanych elementów, tak aby powstała sieć neuropodobna wydobywała z podanego sygnału odpowiednie parametry sygnału. W licznych pracach kilkoma różnymi metodami znaleziono wiele sieci neuropodobnych wydobywających formanty z wejściowego sygnału mowy. Na przykład, na rysunku 4-70 pokazano strukturę, która poza możliwością skutecznej lokalizacji formantów, jest niewrażliwa na zakłócenia i nieregularności „mikrostruktury" analizowanego widma.



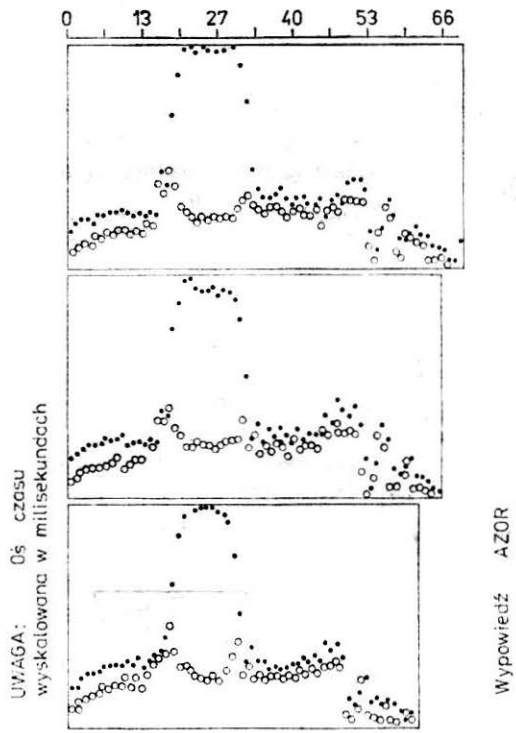

4-69. Uproszczony schemat komórki nerwowej, użyteczny przy projektowaniu systemów przetwarzania sygnałów, działających na zasadzie tzw. sieci neuropodobnych. Sygnały wejściowe oznaczono $x_1, \ldots, x_n$, sygnał wyjściowy $y$, a parametry $w_1, \ldots, w_n$ oznaczają wagi synaptyczne. Literą $p$ oznaczono próg zadziałania komórki zgodnie z fizjologiczną zasadą „wszystko albo nic". Obok podano przebieg charakterystyki statycznej neuronu, stanowiącej uproszczoną wersję charakterystyk znanych z doświadczeń biologicznych



4-70. Struktura sieci neuropodobnej, która może być użyta do wykrywania formantów. Możliwe jest użycie sieci o lepszych własnościach, niewrażliwych na większość możliwych deformacji sygnału, ale ich struktura jest bardzo rozbudowana. Szczegóły budowy tych sieci podane są w literaturze

W uzupełnieniu prezentowanych rozważań warto powrócić raz jeszcze do omówionych na wstępie podrozdziału momentów widmowych. Nie stanowią one — jak się wydaje — konkurencji dla formantów, lecz są cennym uzupełnieniem informacji zawartej w formantach. Formanty z zasady określane są w dźwięcznych fragmentach sygnału mowy, a swoją szczególną przydatność wykazują w analizie stanów ustalonych samogłosek i w śledzeniu stanów przejściowych większości spółgłosek dźwięcznych. Momenty wykazują szczególną przydatność w analizie głosek szumowych, dla których

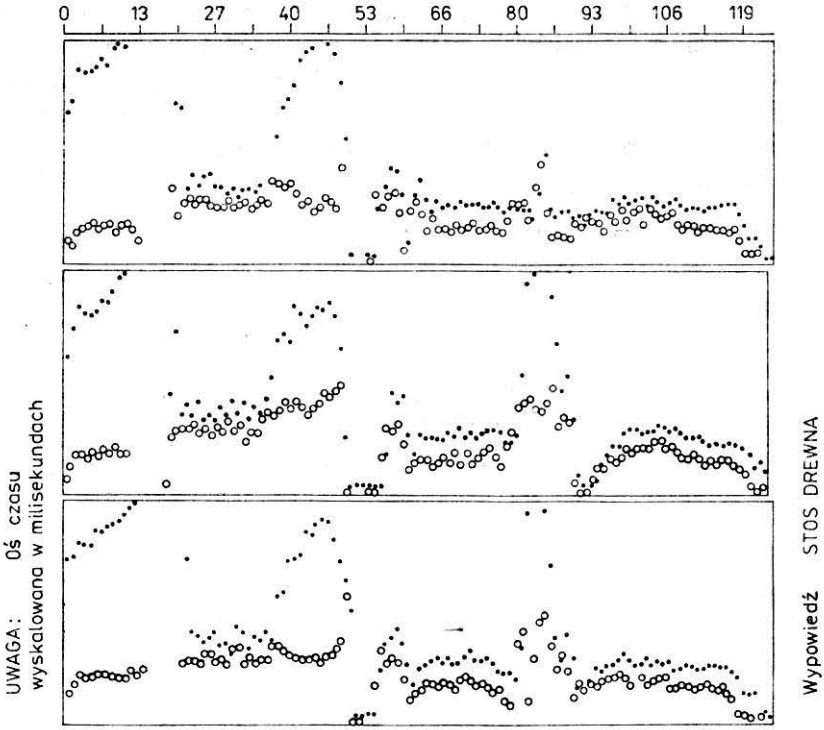
badanie lokalnych własności widma, takich jak formanty, pozbawione jest na ogół sensu, natomiast wiele informacji zawiera głobalny opis widma, dostarczany między innymi przez momenty. Kreśląc „mapy" zmienności momentów, w analogiczny sposób jak uprzednio dla formantów, możemy również zauważyć regularności w ich przebiegu (rys. 4-71 i 4-72). Porównując te „mapki" z przebiegami podanymi na rys. 4-66 i 4-67 łatwo znajdujemy



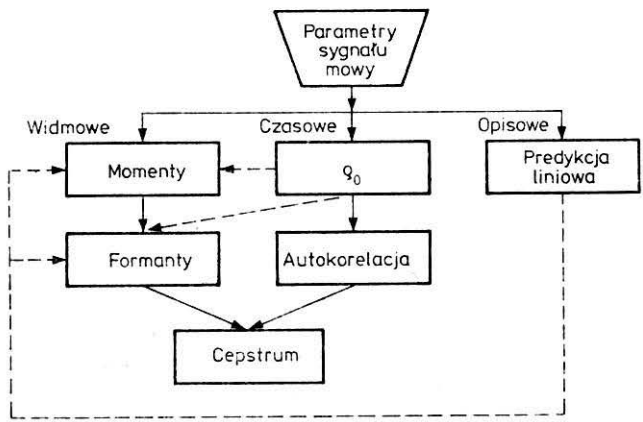

4-71. Mapka przebiegu momentów widmowych w wypowiedzi *azor*. Momenty uzupełniają informację daną w postaci formantów (por. rys. 4-66) i stanowią cenne uzupełnienie informacji przy rozpoznawaniu

potwierdzenie tezy o uzupełniającym charakterze informacji zawartej w momentach widmowych w stosunku do tej, która jest wznoszona przez formanty. Na marginesie można także odnotować fakt, że opierając się na samych momentach widmowych możliwe jest także rozpoznawanie niektórych prostszych klas głosek — na przykład bez trudu można rozpoznawać opierając się wyłącznie na momentach wszystkie samogłoski oraz głoski szumowe.

Momenty widmowe i formanty nie stanowią jedynych parametrów, których można używać przy opisie sygnału mowy. Do określonych celów można sygnał mowy opisywać stosując różne parametry, tak dobierane, aby w sumie ich objętość informacyjna była wydatnie mniejsza od objętości wejściowego sygnału mowy, ale by zachowane były w nich te cechy źródłowego sygnału, które są przydatne z punktu widzenia ustalonego celu analizy. Parametrami takimi mogą być między innymi: omówiony wyżej parametr $\varrho_0$ (gęstość przejść przez zero sygnału i ewentualnie także jego pochodnych i całek), przebieg funkcji autokorelacji sygnału (przydatny przy określaniu




4-72. „Mapka" przebiegu momentów widmowych w wypowiedzi *stos drewna*




4-73. Podział parametrów opisujących sygnał mowy. Liniami przerywanymi pokazano drogi uzyskiwania parametrów metodami pośrednimi

periodyczności sygnału i badaniu funkcjonowania krtaniowego źródła tonu), parametry cepstrum sygnału (przydatne do rozplatania wpływu traktu głosowego i źródła dźwięku na ostateczną postać sygnału) są wreszcie — warte osobnego omówienia — parametry metod liniowej predykcji sygnału mowy (rys. 4-73).

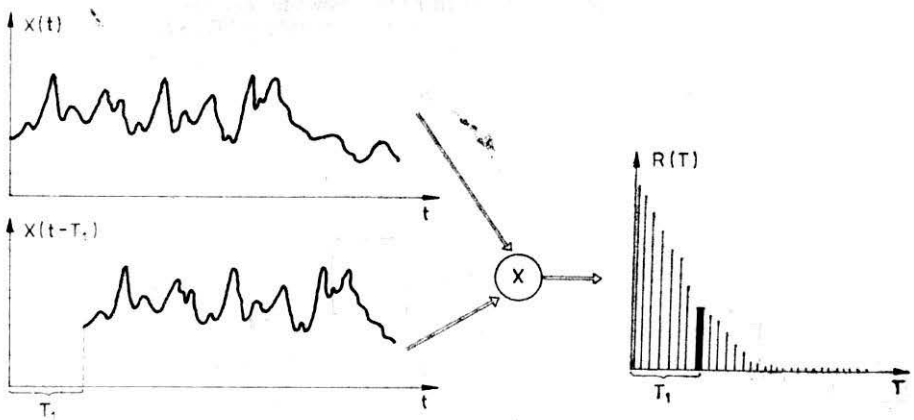
Nie o wszystkich parametrach można tu napisać tak obszernie i dokładnie jak by należało, gdyż konieczne jest zachowanie właściwych proporcji pomiędzy opisem metod, które już zyskały powszechne uznanie (takich jak wyszukiwanie formantów czy liniowa predykcja) a technikami zapożyczo-

nymi w istocie z innych dziedzin przetwarzania sygnałów i tam mających swoje obszerne uzasadnienie, bogatą literaturę i szczegółowo opracowaną metodologię. Wzmiankując więc raczej, niż dokładnie dyskutując, przedstawimy teraz kolejno metody autokorelacji i technikę cepstralną.

Funkcja autokorelacji (rys. 4-74) sygnału opisanego przebiegiem czasowym $x(t)$ dana jest wzorem:

$$R(T) = \int_{-\infty}^{\infty} x(t)x(t-T)\mathrm{d}t \tag{4.84}$$

Wykorzystywaną w analizie sygnału mowy własnością funkcji autokorelacji jest możność wykrywania na jej podstawie okresowości sygnału. Istotnie, obok wydatnego maksimum przy wartości $T = 0$ (wartość $R(0)$ jest miarą

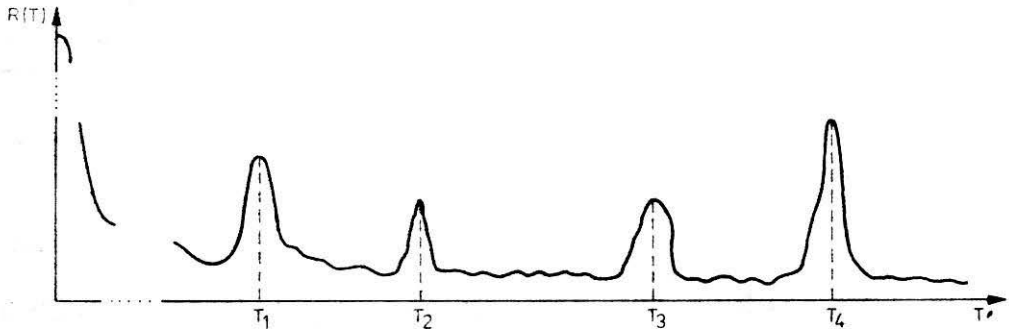



4-74. Tworzenie funkcji autokorelacji. Wartość tej funkcji dla wybranej wartości argumentu $T_1$ (pogrubiony prążek na wykresie) wyznaczana jest na drodze uśredniania iloczynu danego przebiegu czasowego $x(t)$ i przebiegu przesuniętego o wartość $T_1$ (patrz lewa strona)

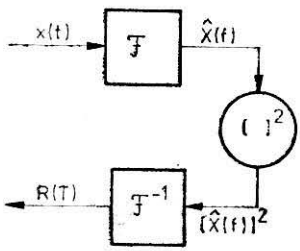
wariancji przebiegu $x(t)$ i zdecydowanie dominuje nad pozostałymi wszystkimi wartościami $R(T)$), w funkcji autokorelacji wykryć można wyraźnie maksima w punktach $T_1$, $T_2$, ..., $T_k$, przy czym każde maksimum sygnalizuje obecność w sygnale $x(t)$ składowej periodycznej o okresie $T_1$, $T_2$, ... ..., $T_k$ (rys. 4-75). W odniesieniu do sygnału mowy technika ta bywa wykorzystywana do wyznaczania częstotliwości tonu krtaniowego w dźwięcznych segmentach sygnału. Funkcję autokorelacji wygodnie jest wyznaczać za pomocą widma sygnału. Istotnie, oznaczając przez $\mathscr{F}$ transformatę Fouriera sygnału możemy zapisać:

$$R(T) = \mathscr{F}^{-1}\{[\mathscr{F}(x(t))]^2\} \tag{4.85}$$

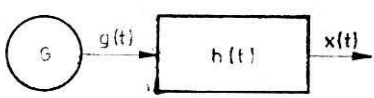
przy czym prawdziwość zależności (4.85) wynika ze wzoru (4.84) i znanych własności przekształcenia Fouriera w stosunku do całki splotowej. Nawiasem mówiąc w przeszłości zależność (4.85) usiłowano wykorzystywać w odwrotną stronę, upatrując w niej wygodną metodę obliczeniowego wyznaczania transformaty Fouriera. Obecnie opracowanie algorytmu FFT tak uprościło obliczeniowe wyznaczanie widma sygnału, że chętnie sięga się do możli-



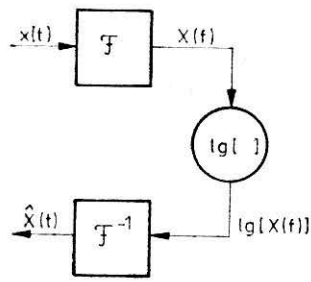

4-75. Obecność składowych okresowych w sygnale $x(t)$ powoduje powstawanie maksimów funkcji autokorelacji przy wartościach $T$ równych odpowiednim wartościom okresu składowych periodycznych. Rysunek ma charakter ilustracji sygnalizowanej tezy, a nie dokładnego wykresu; w szczególności oś $T$ musi być traktowana jako nieciągła (wykropkowany fragment) z tego powodu, że oznaczając obecność maksimum w punkcie $T_1$ musimy liczyć się z pojawieniem kolejnych w punktach $2T_1$, $3T_1$ itd., analogicznie z $T_2$, $T_3$ i $T_4$. Założono więc, że $T_4 - T_1 \ll T_1$ i narysowano jedynie interesujący fragment osi $T$. Podobnie oś wartości funkcji korelacji musiała być przerwana ze względu na zachowanie czytelnego obrazu przy jednoczesnym występowaniu zależności $R(0) \gg R(T)$ dla wszystkich $T > 0$




4-76. Technika obliczenia funkcji autokorelacji za pomocą prostego i odwrotnego przekształcenia Fouriera jest obecnie — ze względu na dostępność algorytmu FFT — najwygodniejszą drogą postępowania




4-77. Model generacji sygnału mowy, używany przy analizie cepstralnej. Funkcja cepstrum umożliwia rozdzielenie w sygnale $x(t)$ składowych pochodzących od własności generatora $G$ i przebiegu wymuszającego $g(t)$, z oddzieleniem ich od składowych pochodzących od własności dynamicznych toru głosowego, wyrażających się jego odpowiedzią impulsową $h(t)$




4-78. Sposób obliczenia cepstrum z z wykorzystaniem transformaty Fouriera. W użyciu znajduje się również wariant metody, oparty na transformacie $\mathscr{Z}$. Zasadniczym problemem przy analizie cepstralnej jest sposób traktowania operacji logarytmowania, oznaczonej w kółku. Jeśli przyjąć, że logarytmowaniu podlegają zespolone wartości $\hat{X}(f)$, wówczas mamy do czynienia z cepstrum zespolonym — dokładnym, ale kłopotliwym w analizie i obliczeniach. Jeśli natomiast brać pod uwagę logarytm modułu, wówczas obliczenia stają się prostsze, ale gubione są zależności fazowe sygnału

wości wykorzystania podanego wzoru do wyliczania funkcji autokorelacji (rys. 4-76).

Inną techniką opierającą się również na przekształceniu Fouriera jest analiza homomorficzna, czyli głównie wyznaczanie cepstrum sygnału mowy i badanie jego przebiegu. Analiza homomorficzna opiera się na dość ogólnych założeniach i dysponuje obszerną teorią, której nie ma potrzeby tu w ca-

łości przytaczać, zaczniemy zatem rozważania od punktu dogodnego dla praktyki analizy sygnału mowy. Niech

$$X(f) = \mathcal{F}[x(t)] \tag{4.86}$$

będzie transformatą Fouriera sygnału mowy $x(t)$. Wówczas cepstrum zespolonym sygnału $x(t)$ nazwiemy przebieg czasowy obliczony ze wzoru:

$$\hat{X}(T) = \mathcal{F}^{-1}\ [\ln X(f)] \tag{4.87}$$

Ważną własnością cepstrum (którego nazwa pochodzi od czytanego wspak słowa spectrum) jest możliwość nader łatwego rozdzielenia w nim wpływów generatora sygnału i własności układu go modulującego. Istotnie, niech sygnał $x(t)$ powstaje (tak, jak to ma miejsce przy artykulacji mowy) przez kształtowanie sygnału generatora $g(t)$ (na przykład tonu krtaniowego) przez układ o zmiennej odpowiedzi impulsowej $h(t)$ (tor głosowy) (rys. 4-77). Zakładając, że mamy do czynienia z układami liniowymi, możemy sygnał $x(t)$ przedstawić w postaci całki splotowej sygnałów $g(t)$ oraz $h(t)$:

$$x(t) = \int_{-\infty}^{\infty} g(T)h(T-t)\mathrm{d}T \tag{4.88}$$

Jak wiadomo, splotowi w dziedzinie czasu odpowiada iloczyn w dziedzinie transformat Fouriera

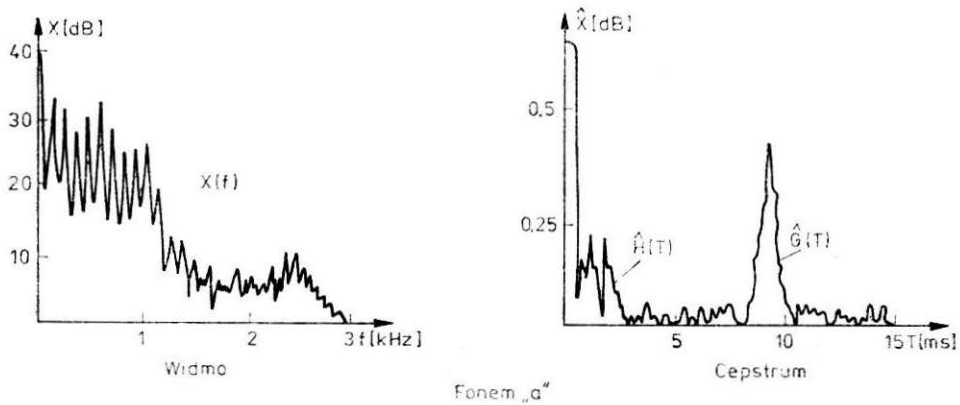
$$X(f) = G(f)H(f) \tag{4.89}$$

gdzie oczywiście $G(f) = \mathcal{F}[g(t)]$ oraz $H(t) = \mathcal{F}[h(t)]$. Kolejna operacja wyznaczania cepstrum, logarytmowanie, zamienia iloczyn ze wzoru (4.89) na jeszcze prostszą i wygodniejszą w dalszych operacjach sumę, która nie zmienia się po dokonaniu operacji odwrotnej transformacji Fouriera. W związku z czym:

$$\hat{X}(T) = \hat{G}(T) + \hat{H}(T) \tag{4.90}$$

W tej postaci rozdzielenie składników pochodzących od pobudzenia krtaniowego i składowych zależnych od procesu artykulacji sygnału w trakcie głosowym jest już łatwe (rys. 4-78). Następnie, zależnie od potrzeb, można koncentrować uwagę wyłącznie na parametrach pobudzenia (na przykład do diagnostyki foniatrycznej) lub wyłącznie na efektach procesu artykulacji (na przykład do automatycznego rozpoznawania treści wypowiedzi). Warto zwrócić uwagę, że nader użyteczne parametry cepstralne nie zyskały jeszcze wystarczającego upowszechnienia w technice analizy i przetwarzania sygnału mowy, przy czym główna przyczyna tkwi w fakcie, że operacje wymagane przy obliczaniu cepstrum możliwe są w praktyce do wykonania jedynie na drodze cyfrowej, natomiast znacząca część badań nad sygnałem mowy była i jest prowadzona metodami analogowymi. Jedna z głównych trudności, jakie pojawiają się przy stosowaniu analizy cepstralnej, wynika z konieczności operowania liczbami zespolonymi, gdyż już pierwsza zastosowana transformacja Fouriera powoduje, że przebieg $X(f)$ staje się zespolony. Na szczęście w zastosowaniach praktycznych zamiast cepstrum zespolonego można stosować cepstrum wyznaczane w dziedzinie liczb rzeczywistych.

Możliwość takiego uproszczenia wynika z pewnej własności przekształcenia cepstralnego, którą podamy niżej bez dowodu, odsyłając bardziej dociekliwych Czytelników do pozycji literatury zamieszczonych na końcu książki. Otóż wśród przebiegów czasowych poddawanych przekształceniu cepstralnemu wyróżnić można klasę przebiegów minimalnofazowych. Dla tych przebiegów dowodzi się, że możliwe jest zastąpienie zespolonej wartości $X(f)$ przez moduł widma $|X(f)|$, a co za tym idzie — możliwość zastąpienia logarytmowania liczb zespolonych logarytmowaniem liczb rzeczywistych. Dla przebiegów minimalnofazowych jest to postępowanie zapewniające taką samą dokładność, jak obliczenia za pomocą dokładnych, pełnych wzorów. Niestety, sygnał mowy w większej części swego przebiegu nie jest minimalnofazowy. Ma to jednak mało istotny wpływ na przebieg analizy, gdyż dla sygnałów nieminimalnofazowych wartości cepstrum zachowują pełną informację o module widma, a nie o jego fazie. Natomiast w większości praktycznie prowadzonych analiz faza sygnału mowy, jak wielokrotnie podkreślano, nie jest brana pod uwagę. Wobec tego, utrata informacji o fazie, związana z korzystaniem z uproszczonej reguły wyliczania cepstrum, nie jest stratą ważną.




4-79. Widmo głoski *a* oraz obliczone na jego podstawie cepstrum (po prawej stronie rysunku), w którego przebiegu wyraźnie zaznacza się składowa pochodząca od pobudzenia krtaniowego $\hat{G}(T)$ oraz, w okolicy $T = 0$, składowa pochodząca od procesu artykulacji $\hat{H}(T)$. Rozdzielenie tych dwu składowych jest teraz łatwe, a wynik — w postaci wyodrębnionego przebiegu $\hat{H}(T)$ — jest bardzo przydatny przy rozpoznawaniu mowy

Można wykazać, że składniki odpowiadające własnościom kanału głosowego mieszczą się w cepstrum w pobliżu wartości $T = 0$, co powoduje, że rozdzielenie składników $\hat{H}(T)$ od $\hat{G}(T)$ we wzorze (4.89) może odbywać się przez przemnożenie przebiegu $\hat{X}(T)$ przez funkcję „okna" o wartościach różnych od zera w pobliżu $T = 0$ (rys. 4-79). Wydzielony w ten sposób składnik $\hat{H}(T)$ może zostać użyty do wielu celów. Po dokonaniu operacji odwrotnych do używanych przy tworzeniu cepstrum otrzymuje się sygnał o gładkiej obwiedni widma, odpowiadającej ruchom artykulacyjnym narządów mowy. Taka czynność, nazywana wygładzaniem cepstralnym, jest

nieocenioną pomocą przy wszelkich dalszych analizach i badaniach sygnału mowy. Możliwe jest także wykorzystanie analizy cepstralnej do wydzielenia tonu krtaniowego (składowej $\hat{G}(T)$) oraz do tworzenia mowy syntetycznej. Często stosuje się ją do usuwania z sygnału różnych form zakłóceń, echa, pogłosu, zniekształceń wnoszonych przez proces rejestracji sygnału (tą metodą „czyści" się archiwalne nagrania o dużej wartości historycznej lub artystycznej). Pojawiają się wciąż nowe i coraz bardziej interesujące doniesienia na temat wykorzystania analizy cepstralnej. Należy oczekiwać, że w okresie, jaki upłynie od napisania tej książki, do chwili, kiedy dotrze ona do Czytelników, pojawią się nowe, ważne publikacje, nie ujęte w podanym spisie literatury.

Zwróćmy jeszcze uwagę na cztery ważne własności cepstrum zespolonego, wymieniane w literaturze, a mające zastosowanie w analizie mowy. Własności te sformułujemy dla dyskretnych postaci zarówno sygnału $x(t)$, jak i jego cepstrum $\hat{X}(T)$, gdyż — jak wspomniano — cepstrum jest wyznaczane wyłącznie metodami cyfrowymi. Zapiszmy zatem oryginalny sygnał mowy w postaci dyskretnego ciągu wartości cyfrowych $x(n)$. Zamiast transformacji Fouriera użyjemy jej dyskretnej analogii, to znaczy transformacji $\mathcal{Z}$, zapisując ją $X(z)$, dyskretną zaś postać obliczonego cepstrum zapiszemy w postaci ciągu $\hat{X}(N)$.

Przy takich oznaczeniach wspomniane własności można sformułować w następującej postaci.

1. Własność dotycząca składowej sygnału pochodzącej od odpowiedzi impulsowej traktu głosowego. Sformułować ją można w postaci twierdzenia: jeśli ciąg wejściowy $x(n)$ ma transformatę wymierną daną wzorem

$$X(z) = \frac{\prod_{k=1}^{m_l} (1 - a_k z^{-1}) \prod_{k=1}^{m_0} (1 - b_k z)}{\prod_{k=1}^{p_l} (1 - c_k z^{-1}) \prod_{k=1}^{p_0} (1 - d_k z)} \tag{4.91}$$

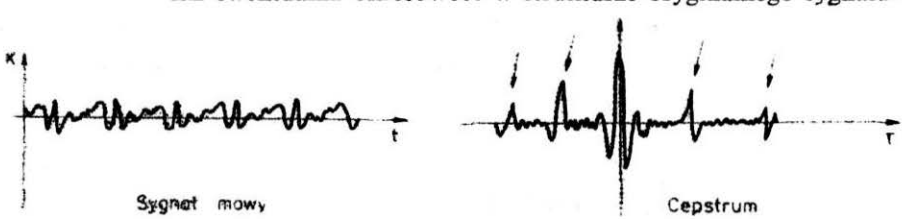
to jego cepstrum zespolone można zapisać:

$$\hat{X}(N) = \begin{cases} \sum_{k=1}^{m_0} \frac{b_k^{-N}}{N} - \sum_{k=1}^{p_0} \frac{d_k^{-N}}{N} & \text{gdy } N < 0 \\ \ln |A| & \text{gdy } N = 0 \\ -\sum_{k=1}^{m_l} \frac{a_k^N}{N} + \sum_{k=1}^{p_l} \frac{c_k^N}{N} & \text{gdy } N > 0 \end{cases} \tag{4.92}$$

Zauważmy, że jeśli wszystkie współczynniki $a_k$, $b_k$, $c_k$, $d_k$ są co do modułu mniejsze od jedności (co się zakłada przy wprowadzaniu wzoru (4.91), to składniki zawierające $a_k$ oraz $c_k$ odpowiadają minimalnofazowej części sygnału, natomiast składniki zawierające $b_k$ oraz $d_k$ są odpowiedzialne za jego nieminimalnofazowość. W szczególności, rozważając dalej przypadek minimalnofazowego sygnału będziemy zakładali wszystkie $b_k$ oraz $d_k$ równe zeru. Niezależnie jednak od minimalnofazowości sygnału — lub jej braku —

widać, że ciąg dany wzorem (4.92) maleje szybko ze wzrostem bezwzględnej wartości $N$. Szybkość tego malenia jest nie mniejsza od szybkości malenia ciągu $1/|N|$. Upoważnia to do twierdzenia (wykorzystywanego wyżej), że składowych cepstrum pochodzących od sygnału odpowiedzi impulsowej traktu głosowego szukać trzeba w okolicy $T = 0$, co dla dyskretnych ciągów odpowiada wartości $N = 0$.

2. Własność cepstrum zespolonego dotycząca sygnałów będących ciągami dyskretnych impulsów (w dziedzinie czasu) o różnych amplitudach. Cepstrum takiego przebiegu ma także postać ciągu impulsów o takich samych odstępach, jak w sygnale oryginalnym. Zmianie ulegają jedynie amplitudy. Może to być wykorzystane do wykrywania w cepstrum składników odpowiadających pracy generatora krtaniowego, którego sygnał można uważać za periodyczny lub quasi-periodyczny ciąg impulsów ciśnienia akustycznego. Oczywiście wniosek ten ma charakter przybliżony. Cytowana własność cepstrum dotyczyła impulsów idealnych o postaci zbliżonej do impulsów Diraca, podczas gdy ton krtaniowy ma formę ciągu impulsów piłozębowych. Niemniej impulsowy charakter tego sygnału dość wyraźnie uwidacznia się w cepstrum, co między innymi bywa wykorzystywane do wykrywania obecności pobudzenia krtaniowego (rozróżnianie dźwięcznych i bezdźwięcznych segmentów mowy) oraz do określania częstości tonu krtaniowego, gdyż odstępy między maksimami cepstrum odpowiadają wiernie okresowości tonu, a są znacznie wyraźniej widoczne (dla wartości $T$ dalekich od zera) niż ewentualna okresowość w strukturze oryginalnego sygnału (rys. 4-80).




4-80. Okresowość przebiegu sygnału mowy (po lewej stronie rysunku) manifestuje się bardzo wyraźnymi pikami cepstrum (po prawej stronie, wskazane strzałkami). Jest to jedna z wygodniejszych metod wydzielania tonu krtaniowego i określenia jego parametrów (na przykład wysokości głosu w intonografie)

3. Własność pozwalająca wiązać cepstrum wyliczone z modułu transformaty Fouriera (uproszczone) z cepstrum pełnym — dla sygnałów spełniających warunek minimalnofazowości. Rozważając raz jeszcze wzór (4.82) dostrzegamy, że dla ciągów (sygnałów) minimalnofazowych cepstrum ma wartość 0 dla wartości $N < 0$ ($T < 0$). Rozpatrując zatem zależność między cepstrum dokładnym $\hat{X}(N)$, a cepstrum przybliżonym (wyliczanym przez logarytmowanie modułu transformaty $X(z)$ lub $X(f)$ odpowiednio dla procesów ciągłych) możemy stwierdzić, że dla wszystkich $N > 0$ zależność ta ma postać

$$\hat{X}(N) = 2\hat{X}_p(N) \tag{4.93}$$

gdzie $\hat{X}_p(N)$ oznacza cepstrum wyznaczone z modułu $X(z)$ lub $X(f)$. Osobnego

rozważenia wymaga jedynie przypadek $N = 0$. Okazuje się bowiem, że $\hat{X}(0) = \hat{X}_p(0)$.

Przydatność przedstawionej własności jest bezdyskusyjna; korzystaliśmy z niej wcześniej często. Jednak sygnał mowy nie może być traktowany jako minimalnofazowy i dlatego w praktycznym stosowaniu wzoru (4.93) przydatna jest kolejna, czwarta odnotowywana własność cepstrum zespolonego. Otóż jeśli rozpatrywany sygnał $x(n)$ nie jest minimalnofazowy, to wówczas cepstrum $\hat{X}(N)$ wyznaczone zgodnie ze wzorem (4.93) (przy uwzględnieniu wyjątkowości przypadku $N = 0$ oraz przy wyzerowaniu wartości $\hat{X}(N)$ dla $N$ ujemnych) będzie cepstrum innego minimalnofazowego sygnału $x_m(n)$, mającego jednak identyczny jak sygnał $x(n)$ moduł transformaty Fouriera. Innymi słowy — co było wyżej również wykorzystane — w przypadku kiedy sygnał mowy nie spełnia rygorów minimalnofazowości cepstrum przestaje reprezentować stosunki fazowe w rzeczywistym sygnale, ale nadal wiernie oddaje moduł jego widma.

Reasumując należy raz jeszcze podkreślić charakterystyczne elementy opisu sygnału mowy. Zależnie od potrzeb można dobrać różne parametry w celu opisania tych własności sygnału, które z tego punktu widzenia są najbardziej przydatne. Parametry te — niezależnie od tego, jakie są, zmieniają się w czasie, gdyż sygnał mowy jest kształtowany w procesie artykulacji, jest zmienny w czasie i niesie w różnych momentach czasu różne informacje. Ponieważ opis parametryczny nie jest na ogół celem sam w sobie, lecz służy do optymalizacji przesyłania sygnału mowy przez łącze telekomunikacyjne lub do oszczędnej budowy algorytmów automatycznego rozpoznawania mowy, to wydaje się celowe na koniec zajęcie stanowiska wobec mnogości różnych parametrów i dokonanie próby wyboru parametrów, które — zdaniem Autora — najlepiej nadają się do wymienionych celów. Otóż nie biorąc pod uwagę dyskutowanych w dalszym ciągu zagadnień predykcji liniowej, celowe wydaje się rekomendowanie zestawu złożonego z dwu dyskutowanych obszerniej momentów widmowych oraz trzech pierwszych formantów (a dokładniej — częstotliwości formantowych) — jako zestawu przenoszącego bardzo dużo informacji o treści analizowanej, czy przesyłanej wypowiedzi — bez nadmiernej rozbudowy ilości informacji zawartej w wyselekcjonowanych parametrach.

## 4.5. Technika predykcji liniowej w opisie sygnału mowy

Rozwój metod komputerowych w analizie sygnałów — w tym także sygnału mowy — prowadzi do sięgania przez badaczy do takich metod i technik przetwarzania, które metodami analogowymi w ogóle nie mogły być realizowane, natomiast z użyciem szybkich, efektywnych metod komputerowych mogą być prowadzone z konkurencyjnymi, w stosunku do tradycyjnych metod, rezultatami. Jedną z najefektywniejszych i również często stosowanych metod tego rodzaju jest omawiana tu technika predykcji liniowej. Metoda ta ma wiele wariantów i odmian, bywa zresztą stosowana do różnych celów i jest przydatna do analizy wielu różnych rodzajów sygnału.

W kontekście analizy mowy uzyskuje się za jej pomocą bardzo zróżnicowane wyniki — od opisu sygnału w formie ułatwiającej jego rozpoznawanie, przez oszczędny opis, wykorzystywany do skompandowanego przesyłania sygnału przez łącza, aż do celów badawczych, gdzie za pomocą predykcji liniowej wyznacza się geometryczne parametry tonu głosowego w trakcie procesu artykulacji określonych głosek.

Istota metody polega na następującym stwierdzeniu. Ponieważ sygnał mowy $x(n)$ powstaje w wyniku przekształcania sygnału źródła krtaniowego $g(n)$ w trakcie głosowym o transmitancji (funkcji przejścia, będącej transformatą $\mathscr{Z}$ odpowiedzi impulsowej $h(n)$) wyrażającej się wzorem:

$$H(z) = G\frac{1+\sum_{k=1}^{q} b_k z^{-k}}{1+\sum_{k=1}^{p} a_k z^{-k}} \tag{4.94}$$

to może on być wyliczany ze wzoru — wynikającego natychmiast ze struktury transmitancji (4.94) — postaci:

$$x(n) = -\sum_{k=1}^{p} a_k x(n-k) + G\sum_{k=1}^{q} b_k g(n-k) + Gg(n) \tag{4.95}$$

Warto zwrócić uwagę na predykcyjny charakter wzoru (4.95). Wartość sygnału $x$ w chwili $n$ jest przewidywana na podstawie poprzednich wartości sygnału $x$ i sygnału $g$ oraz na podstawie bieżącej wartości sygnału $g$.

Niestety, przydatność wzoru (4.95) jest ograniczona ze względu na to, że na ogół nie znamy wartości $g(n)$ dla żadnej wartości $n$. Dlatego opis traktu głosowego, który powinien mieć postać daną wzorem (4.94), zamieniamy do postaci zawierającej tylko bieguny:

$$H(z) = \frac{G}{1+\sum_{k=1}^{p} a_k z^{-k}} \tag{4.96}$$

czemu odpowiada autoregresyjna zależność:

$$x(n) = -\sum_{k=1}^{p} a_k x(n-k) + Gg(n) \tag{4.97}$$

O sygnale $g(n)$ nie mamy oczywiście żadnej informacji, zakładamy więc, że jego wartości są przypadkowe, i stawiamy zadanie znalezienia takich wartości $a_k$ ($k = 1, 2, \ldots, p$), aby minimalizować sumę kwadratów błędów, to znaczy rozbieżności między wartościami rzeczywistego sygnału $x(n)$ a ich przybliżonymi wartościami, wyznaczonymi z zależności (4.97), przy pominięciu składnika $Gg(n)$, który jest nie znany. Minimalizowaną funkcję można zapisać w postaci:

$$E = \sum_{n}\left[x(n)+\sum_{k=1}^{p} a_k x(n-k)\right]^2 \tag{4.98}$$

a jej minimum można osiągnąć, wyznaczając współczynniki $a_k$ z $p$ równań postaci:

$$\frac{\partial E}{\partial a_k} = 0 \quad k = 1, 2, \ldots, p \tag{4.99}$$

lub po rozpisaniu:

$$\sum_{i=1}^{p} a_i \sum_{n} x(n-i)x(n-k) = -\sum_{n} x(n)x(n-k) \quad k = 1, 2, \ldots, p \tag{4.100}$$

Gdyby zakres zmienności $n$ w powyższych wzorach rozciągał się na przedział nieskończony, wówczas odpowiednie sumy byłyby współczynnikami ciągu autokorelacyjnego sygnału $x(n)$ i rozwiązywanie równań (4.100) byłoby znacznie uproszczone. Macierz układu równań (4.100) jest w takim przypadku macierzą Toeplitza, to znaczy wartości elementów wzdłuż każdej przekątnej byłyby identyczne. Niestety, wartości $n$, dla których znane są wartości $x(n)$, są ograniczone i obliczenia nieco się komplikują, niemniej rozwiązanie jest zawsze osiągalne. W literaturze podanej na końcu książki można znaleźć zarówno propozycje różnych metod rozwiązywania układu równań (4.100), jak i teksty programów komputerowych (głównie w języku FORTRAN) do ich rozwiązywania.

Po wyznaczeniu współczynników $a_k$ można przyjąć ich wartości jako parametry, na podstawie których będzie prowadzony proces rozpoznawania odpowiednich fragmentów sygnału mowy. Można je także przesyłać łączem telekomunikacyjnym, aby — wykorzystane na odbiorczym końcu łącza — służyły do syntezowania mowy, przesyłanej tym sposobem ze znaczną oszczędnością objętości informacyjnej łącza. Można wreszcie, co często bywa głównym celem wyznaczania współczynników predykcji liniowej, obliczać ze znanych współczynników predykcji widmo sygnału. Dysponując wzorem (4.96) i znając wartości współczynników liniowej predykcji $a_k$ można na drodze prostych przekształceń znaleźć widmo sygnału a dokładniej, obwiednię widma, zależną tylko od własności artykulacyjnych narządów mowy i wolną od elementów przypadkowych i zbędnych szczegółów (na przykład prążków pochodzących od harmonicznych tonu krtaniowego). Doświadczenia wykazują, że widmo wyznaczone z wykorzystaniem techniki liniowej predykcji jest przynajmniej równie gładkie, jak widmo poddane wygładzaniu cepstralnemu, a znacznie „spokojniejsze" niż widmo wyznaczane techniką FFT — nawet przy stosowaniu wyszukanych postaci okna czasowego.

Technika predykcji liniowej, przedstawiona tu skrótowo w swojej podstawowej postaci, zawiera wiele problemów, wymagających dodatkowego uściślenia. Przykładowo otwarty jest problem liczby składników $p$ we wzorze (4.97). Jest ona na ogół wybierana arbitralnie. Podobnie arbitralnie wybierany jest zakres zmienności parametru $n$ we wzorach (4.98) i (4.100) — co ma szczególnie duże znaczenie przy stosowaniu predykcji liniowej zmiennej w czasie („kroczącej" za zmianami struktury sygnału). Jak wspom-

niano wyżej, możliwe jest wykorzystywanie predykcji liniowej do wyznaczania parametrów kanału głosowego. W niektórych pracach wyznaczano nawet profile narządów mowy w trakcie artykulacji poszczególnych głosek (rozkłady średnic wzdłuż osi traktu głosowego). O jakości metod predykcji liniowej świadczyć może fakt dobrej zgodności takich teoretycznie wyznaczonych profili z rzeczywistym przebiegiem rozmiarów traktu głosowego, ustalonym na podstawie danych anatomicznych i fotografii rentgenowskich narządów mowy w trakcie artykulacji ustalonych głosek.

W sumie technika predykcji liniowej może w dużym stopniu zastępować wszystkie wcześniej omówione techniki analizy sygnału mowy, gdyż może służyć do analizy widmowej sygnału, pozwala wykrywać szczegóły jego obwiedni (na przykład formanty), dostarcza parametrów umożliwiających efektywne rozpoznawanie sygnału i jego oszczędne przesyłanie, wreszcie stanowi mało poznane, a zapewne efektywne narzędzie w diagnostyce medycznej narządów mowy.

## 4.6. Opis sygnału mowy z punktu widzenia teorii informacji

Z poprzednich rozdziałów wynikało jednoznacznie, że środki informatyki, cyfrowe metody analizy i przetwarzania sygnałów, a także algorytmy i programy komputerowe, odgrywają współcześnie coraz istotniejszą rolę także i w dziedzinie analizy mowy. Logicznym następstwem tego faktu jest patrzenie na sygnał mowy z punktu widzenia teorii informacji i rozpatrywanie go jako strumienia bitów, koniecznego do wprowadzenia do systemu, przetworzenia, zapamiętania i ewentualnie także wyprowadzenia na zewnątrz.

Jak wiadomo, teoria informacji zajmuje się, wbrew swojej obiecującej nazwie, jedynie pewnym aspektem informacji, mianowicie jej ilością. W dodatku Shannonowska definicja ilości informacji rozmija się w wielu przypadkach z potoczną intuicją, gdyż jako jedyne kryterium ilości informacji zawartej w określonym sygnale brane są jego odpowiednie miary probabilistyczne. Mierzona jest tu niepewność, wyrażana prawdopodobieństwami, i definiowana jest ilość informacji, jako stopień zmniejszenia tej niepewności. Zaletą takiego podejścia jest jego asemantyczność, gdyż miara ilości informacji zawartej w sygnale nie jest związana z sensownością i przydatnością tej informacji (te pojęcia nie dają się wyrażać w sposób sformalizowany), a jedynie z parametrami fizycznymi sygnału. Istotnie, w kanale łączności lub w pamięci komputera informacja zajmuje tyle samo miejsca niezależnie od tego, co oznacza i czy ma sens.

Przedstawiając niżej elementarny zapis opisu sygnału mowy w kategoriach teorii informacji należy uprzedzić Czytelnika, że zarówno z punktu widzenia tej pięknej, wysoce zmatematyzowanej teorii, jak i z punktu widzenia wiedzy o sygnale mowy — jest to opis niepełny. Istnieją jednak i są łatwo dostępne podręczniki, z których problematykę tę można sobie dodatkowo przestudiować, a ta książka ma być raczej przewodnikiem problemowym, a nie

encyklopedią. Zadaniem tego rozdziału jest więc jedynie zasygnalizowanie problemów i możliwości.

Sygnał mowy z punktu widzenia teorii informacji można rozpatrywać jako łańcuch zdarzeń. Zdarzeniami są kolejne wypowiadane głoski, ich liczba jest ograniczona i dlatego nasza niepewność co do tego, która z nich będzie artykułowana, może być wyrażona ilościowo. Proces mówienia tę niepewność usuwa, możemy więc wiązać z sygnałem mowy taką ilość informacji, jaka była pierwotna niepewność co do tego, jaka głoska będzie wypowiedziana. Jednym z pierwszych i podstawowych osiągnięć teorii informacji było określenie związku między pojęciem niepewności elementarnego zdarzenia a wartością jego prawdopodobieństwa. Oznaczając przez $A$ i $B$ rozważane zdarzenia, przez $p(A)$ oraz $p(B)$ ich prawdopodobieństwa oraz przez $H(A)$ i $H(B)$ ich niepewności można zapisać dość oczywiste wymagania:

$$\text{jeśli } p(A) > p(B), \text{ to } H(A) < H(B) \tag{4.101}$$

$$\text{jeśli } p(A) = 1, \text{ to } H(A) = 0 \tag{4.102}$$

Wprowadzając dodatkowo dla niezależnych zdarzeń $A$ i $B$ oznaczenie $AB$ dla ich jednoczesnego zajścia można postawić żądanie:

$$\text{jeśli } p(AB) = p(A)p(B), \text{ to } H(AB) = H(A)+H(B) \tag{4.103}$$

Łatwo wykazać, że istnieje tylko jedna formuła matematyczna, spełniająca wszystkie postawione postulaty:

$$H(A) = -\log_a p(A) \tag{4.104}$$

Miara nieoznaczoności, dana wzorem (4.104), nosi w literaturze miano e n t r o p i i i jest bardzo użyteczna we wszystkich pracach związanych z teorią informacji. Pozostaje jedynie problem wyboru podstawy logarytmu $a$ w podanym wzorze. Decyzja co do jej wyboru jest równocześnie decyzją odnośnie jednostek, w jakich niepewność, a w dalszej kolejności także ilość informacji będziemy wyrażali. Najczęściej wybierana jest podstawa $a = 2$, w związku z czym jednostka niepewności jest dwójkowa. Za jednostkową niepewność uważa się nieoznaczoność sytuacji wyboru dychotomicznego, tzn. istnienie alternatywy dwóch jednakowo prawdopodobnych zdarzeń. Od angielskiej nazwy tej jednostki: „binary unit" pochodzi popularny i często używany skrót: bit. Jeden bit informacji pozwala więc odpowiedzieć na proste, elementarne pytanie: tak lub nie.

Do analizy mowy miara nieoznaczoności dana wzorem (4.104) jest niewystarczająca, gdyż w przypadku śledzenia łańcucha głosek swobodnie wypowiadanych mamy do czynienia w każdym momencie z problemem wyboru jednej spośród $N$ możliwości. Następną głoską może bowiem być dowolna dopuszczalna w danym języku (łącznie z pauzą międzywyrazową), przy czym prawdopodobieństwo wystąpienia poszczególnych głosek może być wyznaczone empirycznie na podstawie badań językoznawczych i fonetycznych. Przykładowo w tablicy 2 zestawiono prawdopodobieństwa określone dla poszczególnych fonemów języka polskiego. Widać, że pojawienie się kolejnego fonemu w ciągu rozpatrywać należy jako wybór konkretnej wartości zmiennej losowej (której wartościami są na przykład numery fo-

nemów w tablicy 2, z odpowiednimi prawdopodobieństwami ich wystąpienia), nie zaś elementarne zdarzenia, o których można mówić w kategoriach, że zaszły lub nie zaszły. Można się tu zresztą dopatrzyć także drugiej zmiennej losowej, którą jest entropia. Z każdym numerem $i$ w tabl. 2 związane jest prawdopodobieństwo $p_i$ oraz obliczona na jego podstawie entropia $-\log_2 p_i$. Można więc podjąć próbę określenia wartości oczekiwanej entropii

Tablica 2. Prawdopodobieństwa występowania poszczególnych fonemów języka polskiego. Zapis fonemów podano w konwencji wynikającej z międzynarodowego systemu transkrypcji fonematycznej. W ostatniej kolumnie podano przykłady wyrazów, w których te fonemy występują

| Fonem | Częstość | Przykład użycia |
|---|---|---|
| (–) | 0,140 | (pauza) |
| (e) | 0,088 | chleb |
| (a) | 0,080 | brat |
| (o) | 0,078 | skok |
| (j) | 0,039 | jodła |
| (t) | 0,038 | tor |
| (ɨ) | 0,035 | ryba |
| (n) | 0,034 | nora |
| (i) | 0,034 | igła |
| (r) | 0,031 | ryba |
| (m) | 0,030 | matka |
| (v) | 0,030 | woda |
| (u) | 0,029 | kruk |
| (p) | 0,027 | pole |
| (s) | 0,026 | sarna |
| (k) | 0,023 | kot |
| (ɳ) | 0,022 | koń |
| (d) | 0,019 | dom |
| (w) | 0,019 | dłoń |
| (l) | 0,018 | lody |
| (ʃ) | 0,017 | koszyk |
| (z) | 0,015 | ząb |
| (ɕ) | 0,013 | świat |
| (t͡s) | 0,013 | cena |
| (f) | 0,013 | fajka |
| (g) | 0,013 | góra |
| (b) | 0,013 | burza |
| (t͡ɕ) | 0,011 | cichy |
| (ʒ) | 0,010 | żaba |
| (t͡ʃ) | 0,010 | czytać |
| (x) | 0,009 | herbata |
| (d͡z) | 0,007 | dzwon |
| (ŋ) | 0,007 | bank |
| (c) | 0,006 | kino |
| (ʑ) | 0,002 | źrebię |
| (d͡ʑ) | 0,002 | dziura |
| (ɟ) | 0,001 | gil |
| (d͡ʒ) | poniżej 0,001 | drożdże |

pojedynczego symbolu (głoski) w ciągłym sygnale mowy. Odpowiedni wzór podano niżej:

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{4.105}$$

Pozwala on, wraz z wartościami podanymi w tabl. 2, wyliczyć entropię pojedynczego fonemu w mowie polskiej. Odpowiednia wartość wynosi 4,06 bit/głoskę. Przy czym warto zauważyć, że jest ona mniejsza od wartości entropii maksymalnej, osiągalnej przy takiej samej liczbie głosek. Istotnie, łatwo wykazać, poszukując maksimum wyrażenia (4.105) ze względu na prawdopodobieństwa $p_i$ ($i = 1, 2, \ldots, n$), że największą wartość entropii osiąga się przy równomiernym rozkładzie prawdopodobieństwa. Ponieważ suma prawdopodobieństw $p_i$ musi wynosić 1 (jakiś fonem zawsze jest wypowiadany, skoro pauzę zaliczyliśmy do nich), wobec tego wszystkie $p_i$ są dla maksymalnej entropii równe $1/n$ i wartość maksymalnej entropii można wyliczyć z prostszej formuły

$$H_{\max} = \log_2 n \tag{4.106}$$

Przy przyjętej liczbie fonemów wartość $H_{\max}$ wynosi ponad 5 bitów/fonem, natomiast nierównomierny rozkład prawdopodobieństwa fonemów powoduje zmniejszenie tej wartości. Skoro tak, to uwzględnienie dotychczas pomijanego faktu istnienia kontekstu i jego wpływu na prawdopodobieństwa poszczególnych fonemów zapewne jeszcze bardziej obniży zawartość informacyjną pojedynczego fonemu. Przypuszczenie to jest w pełni uzasadnione. Rozbudowując wzór (4.105) w sposób umożliwiający wykorzystanie prawdopodobieństw warunkowych, a także wykorzystując prezentowane w literaturze prawdopodobieństwa warunkowe par, trójek i większych zestawów fonemów stwierdzamy, że w miarę rozszerzania kontekstu i uwzględniania powiązań coraz większej liczby głosek entropia pojedynczego fonemu systematycznie maleje. Spadek ten jest dość wyraźny do kontekstu około 5 fonemów, potem malenie entropii jest wolniejsze. Przy kontekstach rzędu 10 i więcej fonemów entropia praktycznie się ustala i przyjmuje najniższą obserwowaną — a jednocześnie prawidłową z punktu widzenia rozważania zawartości informacyjnej sygnału mowy jako całości — wartość, wynoszącą około 1 bit/fonem.

Warto skupić uwagę na tym wyniku: okazuje się, że mowa jako system komunikacyjny charakteryzuje się dużą redundancją (nadmiarowością). Ze względu na nierównomierne częstości występowania poszczególnych fonemów w sygnale mowy, a także z powodu istnienia związków kontekstowych między elementami mowy jej rzeczywista nośność informacyjna wynosi niespełna 20% teoretycznych możliwości. Innymi słowy, spojrzenie na sygnał mowy z punktu widzenia teorii informacji ujawnia redundancyjność tego sygnału. Nadmiarowość, o której mowa, odgrywa ważną rolę przy przekazywaniu mowy, gdyż zabezpiecza zwiększoną niezawodność przekazywania informacji. Dzięki temu, że nie wszystkie fonemy są jednakowo prawdopodobne, możemy odgadnąć fonem, który rozmówca zniekształcił

podczas szybkiej i niezbyt starannej wypowiedzi. Na skutek istnienia związków kontekstowych pomiędzy elementami mowy można rozumieć wypowiedź częściowo zagłuszoną szumami. System porozumiewania przy braku redundancji jest narażony na bezpowrotne straty części informacji, co może prowadzić do całkowitej niemożności komunikowania się.

Nadmiarowość ta, widziana oczami inżyniera analizującego mowę dla potrzeb jej automatycznego rozpoznawania, jest korzystna z podobnych przyczyn. Niedoskonałości systemu identyfikacji fonemów czy analizy na płaszczyźnie akustycznej mogą być — zapewne — kompensowane za pomocą analizy kontekstu i zastępowania błędnych, bezsensownych identyfikacji elementów mowy — jej kontekstowo uzależnionymi, sensownymi i prawdopodobnie poprawnymi zamiennikami.

Natomiast z punktu widzenia inżyniera telekomunikacji redundancja to balast. Owszem, wprowadza się przy transmisji danych — na przykład cyfrowych — dodatkowe bity zabezpieczające przed przekłamaniami. Bywa ich kilka w kilkunastobitowym słowie, stanowią więc poniżej 10% całości transmitowanej informacji. Ale żeby wprowadzać nadmiarowość ponad 80%? Dlatego w systemach telekomunikacyjnych sygnał mowy usiłuje się pozbawić informacyjnego balastu, poszukując metod oszczędnych transmisji, dokonuje się rozmaitymi technikami kompresji sygnału. Na razie — mało skutecznie. Nadal przesyłane są dziesiątki niepotrzebnych bitów, gdyż mowa — obok nadmiarowości strukturalnej, którą oceniano uprzednio — ma przynajmniej trzy dodatkowe źródła nadmiaru, możliwe do wykrycia przy dyskutowaniu jej własności z punktu widzenia teorii informacji.

Pierwsze ze wzmiankowanych źródeł tkwi w „rozwlekłości" sygnału mowy rozpatrywanego jako przebieg czasowy. Teoria informacji pozwala bowiem określić ilość informacji zawartą w sygnale o czasie trwania $T$, szerokości pasma częstotliwości $F$ i zakresie dynamiki $D$. Można tego zresztą dokonać na kilka sposobów, na przykład wprowadzając we wzorze (4.105) zmienną losową ciągłą w miejsce dyskretnej i zastępując sumowanie całkowaniem, albo odwołując się do procesu dyskretyzacji sygnału (por. rozdz. 4.1), który w istocie zamienia sygnał ciągły na zbiór dyskretnych wartości. Nie wdając się tu w rozważania teoretyczne możemy posłużyć się wzorem

$$H = cFDT \tag{4.107}$$

który pozwala wyznaczyć nieoznaczoność (a więc i pojemność informacyjną sygnału). Współczynnik skalujący $c$ może być przyjęty jako równy 1/3 (konieczność wprowadzenia tego mnożnika wynika z faktu, że przy wyliczaniu decybeli stosuje się logarytmy dziesiętne, a nie dwójkowe, jak we wzorach (4.105) i (4.106) i wówczas dla $F$ [Hz], $D$ [dB] i $T$ [s] wartość $H$ wyznaczana jest w bitach. Próby przeliczeń wykonane z użyciem wzoru (4.107) prowadzą do interesujących wyników, zwłaszcza jeśli porównać je z wcześniej określonymi wartościami asymptotycznymi pojemności informacyjnej sygnału przeliczonej na pojedynczy fonem. Przy założeniu pełnego pasma akustycznego, wynoszącego 20 000 Hz, i pełnego zakresu dynamiki, sięgającego 80 dB, sygnał mowy reprezentuje strumień informacji o objętości

ponad 500 000 bitów/s. Krótka wypowiedź, np. *Ala ma Asa*, trwająca przy powolnej artykulacji blisko 2 s, odpowiada objętości informacyjnej $10^6$ bitów — to jest pojemności pamięci średniej wielkości komputera! Tymczasem rzeczywista objętość informacji tej krótkiej, dziesięciofonemowej wypowiedzi nie przekracza — ustaliliśmy to wszak uprzednio — 10 bitów. Stosunek objętości informacyjnej sygnału do jego nośności w sensie treści semantycznych wyraża się więc liczbą rzędu $10^5$. Jest to bez wątpienia wynik szokujący.

Oczywiście przytoczonym rozważaniom można zarzucić, zupełnie słusznie, demagogiczność. Mowę można rozumieć przy pasmie znacznie węższym od 20 kHz, a rozpiętość dynamiki rzędu 80 dB osiągalna jest jedynie w warunkach laboratoryjnych. Ograniczmy się więc do realnych, a nawet minimalnych wartości pasma i dynamiki. Niech pasmo ograniczone zostanie do szerokości 3 kHz (a więc mniej niż standard telefoniczny), a zakres dynamiki niech wynosi $< 40$ dB. Nawet w tych warunkach objętość informacyjna sygnału będzie znaczna: 40 000 bitów/s. Dużo, dla wielu zastosowań o wiele za dużo. Z tego właśnie powodu poszukiwaliśmy w poprzednich podrozdziałach takiej reprezentacji sygnału, która oszczędniej koduje przydatne z rozważanego punktu widzenia aspekty sygnału, a pozwala usuwać zbyteczny nadmiar. Warto jednak mieć świadomość ograniczoności wyników, uzyskiwanych tymi metodami. Na przykład, orientacyjna objętość sygnału mowy zredukowanego do postaci spektrogramu dynamicznego wynosi przynajmniej 10 000 bitów/s, a reprezentacja za pomocą formantów i momentów widmowych wymaga (zależnie od dokładności) kilkuset bitów na sekundę.

To nadal dużo, bardzo dużo. Najdoskonalsze systemy dokonujące kompresji sygnału mowy do jej oszczędnego przesyłania na duże odległości (na przykład działające z wykorzystaniem metody predykcji liniowej) dają sygnał o objętości około 2000 bitów/s. Tymczasem, powtórzmy to raz jeszcze, nawet przy bardzo szybkiej artykulacji rzeczywista objętość informacyjna sygnału mowy nie przekracza 10 bitów/s (przyjmując „oszczędną" reprezentację, zajmującą zaledwie jeden bit informacji na jeden fonem). Dlatego tak wiele nadziei budzą udane próby automatycznego rozpoznawania mowy. Gdyby mowę przed wysłaniem rozpoznawać, kodować cyfrowo i na odbiorczym końcu łącza resyntezować — oszczędności byłyby ogromne.

Zagadnienie nadmiarowości informacyjnej sygnału mowy na tym się nie kończy. Rozważaliśmy nadmiarowość wynikającą na poziomie akustycznym (wynikającym ze stosowania wzoru (4.107) oraz na poziomie fonematycznym korzystając ze wzoru (4.105)). Przechodząc na kolejny poziom, to znaczy interesując się całymi wyrazami, spotykamy się z kolejnym przejawem nadmiarowości sygnału mowy. Liczbę wyrazów w konkretnym języku trudno dokładnie ocenić, są ich jednak z pewnością dziesiątki tysięcy. Tymczasem w częstym użyciu jest ich znacznie mniej. Językoznawcy znają to zjawisko i badają je, układając tak zwane słowniki częstościowe. Słownik częstościowy jest spisem wyrazów badanego podzbioru języka z podaniem częstości ich występowania. Dysponując takim słownikiem można bez trudu

obliczyć, ile wyrazów (a także które) pozwala na zrozumienie określonej części wypowiedzi — lub formułując to samo w inny sposób — znajomość jakiej liczby wyrazów zapewnia z określonym prawdopodobieństwem możliwość zrozumienia dowolnej wypowiedzi. Badania te są ważne i interesujące z tego powodu, że wykazują, jak niewiele w istocie wyrazów jest w częstym użyciu i jak duża nadmiarowość mieści się w dużych słownikach.

W celu uproszczenia dalszych rozważań przyjmiemy za podstawę ujęcie analityczne — przybliżone w swojej istocie, ale dające wystarczający dla naszych potrzeb, klarowny obraz. Otóż zależność między pozycją (numerem kolejnym „i") określonego wyrazu w słowniku częstościowym, a prawdopodobieństwem użycia tego wyrazu wyraża tak zwane prawo Zipfa:

$$p_i = \frac{A}{B+Ci} \tag{4.108}$$

Parametry $A$, $B$ oraz $C$ zależne są od wziętego pod uwagę podzbioru języka, inaczej bowiem kształtują się proporcje określone prawem Zipfa dla języka potocznego, inaczej dla języka literackiego, jeszcze inaczej dla rozważań naukowych. Co więcej, każdy autor ma sobie tylko właściwą skłonność do używania jednych i unikania innych wyrazów, zatem określając parametry $A$, $B$ i $C$ dla dostatecznie długiej próbki tekstu można wnioskować — chociaż nigdy nie jest to wnioskowanie całkowicie pewne — spod czyjego pióra tekst ten wyszedł. Próbowano tej metody na przykład w celu ustalenia autorstwa wielu dzieł, które tradycja przypisuje Szekspirowi. Pozostawiając jednak to językoznawcom skupmy się na użytkowej własności wzoru (4.108). Otóż, stosując go można dość łatwo określić wymaganą liczbę wyrazów $N$, aby w dowolnej odbieranej wypowiedzi nie znalazł się ani jeden wyraz nieznany — z założonym prawdopodobieństwem $P < 1$. Oznacza to, że poszukujemy $N$ spełniającego warunek:

$$\sum_{i=1}^{N} p_i = \sum_{i=1}^{N} \frac{A}{B+Ci} > P \tag{4.109}$$

Dla konkretnego podzbioru języka i dla założonej wartości $p$ można ze wzoru (4.109) każdorazowo wyznaczyć $N$, przy czym okazuje się, że nawet przy wartościach $P$ przekraczających 90% uzyskiwane liczebności $N$ są zaskakująco małe, poniżej tysiąca*). Innymi słowy, znając zaledwie niespełna tysiąc wyrazów można z prawdopodobieństwem przewyższającym 90% rozumieć dowolne wypowiedzi. Zestawienie tego wyniku ze słownikami liczącymi dziesiątki lub setki tysięcy haseł stanowi wymowny dowód kolejnego źródła nadmiarowości sygnału mowy.

Na koniec warto dołączyć kilka uwag na temat gramatyki, jako źródła ko-

*) Rozważania matematyczne oparte na wykorzystaniu prawa Zipfa potwierdzono wielokrotnymi badaniami eksperymentalnymi na naturalnym sygnale mowy. Między innymi badania firmy Bell w latach trzydziestych oparte na analizie zarejestrowanych rozmów telefonicznych wykazały, że wśród 80 tys. wyrazów, 30 z nich stanowiło 50% ogółu wypowiedzianych słów, 155 — 80%, a 737 — 96%. Inne badania, prowadzone przez Glenna i Hitchcoocka w symulowanym systemie kontroli ruchu powietrznego, wykazały, że możliwe jest rozpoznanie 13 tys. zdań opierając się na słowniku składającym się tylko z 54 wyrazów. (Przypis sporządzono na podstawie uwag Recenzenta książki, dr. Ryszarda Gubrynowicza).

lejnej nadmiarowości sygnału mowy. Reguły składniowe, narzucające określone uporządkowanie szyku wyrazów i ich form powodują, że nawet brak wyrazu lub całego fragmentu zdania nie zawsze musi wiązać się z brakiem możliwości zrozumienia całej wypowiedzi. Przeciwnie, okazuje się, że kluczowe znaczenie dla zrozumienia sensu zdania ma jedynie kilka spośród tworzących go wyrazów, pozostałe zaś pełnią funkcje uzupełniające. Rozkład wyrazów w zdaniu, ich porządek, zestawienie, kontekst — niosą niekiedy tak wiele informacji, że odtworzenie brakujących elementów może być dokonane ze stuprocentowym prawdopodobieństwem. W naturalnym języku jest to środek zabezpieczający zrozumiałość mowy niestarannej, niedokładnie artykułowanej, niekiedy wadliwej gramatycznie lub zakłóconej w inny sposób. W systemach technicznych może to być źródłem dodatkowych możliwości lub dodatkowych problemów, zależnie od tego, czy będziemy starali się to zjawisko wykorzystać, czy przeciwnie, podejmiemy próbę eliminacji jego skutków.

Podsumowując ten rozdział można stwierdzić co następuje. Teoria informacji dostarcza narzędzia, które w zastosowaniu do sygnału mowy pozwala oszacowywać jej objętość informacyjną, a tym samym pozwala określić wymagania odnośnie pamięci komputerów przetwarzających mowę i pojemności kanałów telekomunikacyjnych, wykorzystywanych do jej transmisji. Przy okazji tych rozważań udało się wskazać na bardzo istotny problem redundancji (nadmiarowości) sygnału mowy. Nadmiarowość ta występuje zarówno na płaszczyźnie akustycznej (w czasowym, częstotliwościowym i amplitudowym wymiarze sygnału mowy), fonematycznej, leksykalnej i syntaktycznej. Jednoznaczna ocena tej nadmiarowości jest niemożliwa. Jej obecność w podobnych proporcjach w każdym bez wyjątku*) języku świata dowodzi, że w warunkach naturalnej komunikacji głosowej nadmiarowość ta jest niezbędna. Istnienie tej nadmiarowości warunkuje niezawodną komunikację w obecności zakłóceń. W systemach technicznych redundancja jest szkodliwa, gdyż powoduje niepotrzebne zajęcie pamięci komputera lub ogranicza przepustowość łącza telekomunikacyjnego. Usuwa się ją zatem wszelkimi dostępnymi środkami, mając przy tym na względzie możliwość ewentualnego uprzedniego wykorzystania podlegającego eliminacji nadmiaru dla podniesienia wiarogodności analizy lub rozpoznania sygnału mowy. Konkretne przykłady takiego postępowania podane zostaną w kolejnych rozdziałach.

*) Badano metodami teorii informacji odmienne od mowy systemy komunikacji międzyludzkiej. Analizowano „mowę" afrykańskich tam-tamów, różne systemy pisma, egipskie hieroglify i węzełkowe pismo prekolumbijskich kultur Ameryki. Okazało się, że nadmiarowość występuje wszędzie, żaden system komunikacji nie jest od niej wolny, a procentowe wielkości stopnia nadmiarowości okazywały się dla wielu zupełnie odmiennych w swej naturze systemów porozumiewania — bardzo zbliżone. Widocznie człowiekowi dla komfortu odbioru informacji taki nadmiar jest niezbędny, widocznie oszczędniejsze kodowanie informacji, tak korzystne i chętnie stosowane w technice jest obce naszej psychice, a nadmiarowość, będąca wygodnym „spadochronem" dla naszej niestarannej mowy czy pisma jest czymś koniecznym. Fakt ten trzeba brać pod uwagę przy opracowywaniu systemów komunikacji pomiędzy człowiekiem i maszyną, które w obecnej postaci są zbyt technocentryczne i dlatego męczące, niewygodne i nie akceptowane przez człowieka.

# 5

# Sygnał mowy w automatyce

## 5.1. Rola sygnału mowy w systemach sterowania

W poprzednich rozdziałach przedstawiono informacje o własnościach sygnału mowy, sposobach jego analizy i prezentacji. Niniejszy rozdział wraz z następnym są poświęcone dwu najbardziej typowym obszarom praktycznego wykorzystania tej wiedzy.

Na gruncie automatyki sygnał mowy jest wykorzystywany oczywiście do komunikacji pomiędzy ludźmi a systemem sterującym określony obiekt. Obiekt może mieć rozmaity charakter. Najczęściej rozważane są procesy produkcyjne: chemiczne, metalurgiczne, wydobywcze, energetyczne, maszynowe. Nie wyczerpuje to jednak listy rozważanych obiektów, gdyż możliwe jest także rozważenie systemów automatycznego sterowania różnych pojazdów i systemów komunikacyjnych (od pojedynczego samolotu czy okrętu do całych zautomatyzowanych lotnisk, portów, sieci metra lub systemów transportu wewnątrzzakładowego). Do klasy rozważanych systemów należeć też mogą wszelkie systemy komputerowe o rozmaitym przeznaczeniu: do obliczeń naukowo-technicznych, do przetwarzania danych, edukacyjne, doradcze, banki informacji, systemy ekspertowe. Wszystkie te złożone zastosowania komputerów także wymagają sterowania, a systemy sterujące proces świadczenia usług informatycznych muszą — w natu-

ralny sposób — komunikować się z ludźmi. Zresztą rozgraniczenie między wymienionymi wyżej typami obiektów ma w całości charakter umowny. W skład większości systemów produkcyjnych wchodzą także procesy transportowe, a komputery o rozmaitym przeznaczeniu „wrosły" już w większość bardziej złożonych systemów produkcyjnych.

W istocie zatem nie rodzaj automatyzowanego obiektu, lecz zakres i charakter kontaktów pomiędzy człowiekiem a systemem jest głównym wyznacznikiem potrzeb w zakresie wykorzystania sygnału mowy i dlatego raczej ten punkt widzenia będziemy prezentować w dalszych rozważaniach. Na wstępie należy zwrócić uwagę na fakt, że przy rosnącej złożoności podlegających sterowaniu procesów, a także przy postępującej ich automatyzacji charakter kontaktu pomiędzy człowiekiem a sterowanym procesem technicznym nabiera cech dialogu pomiędzy dwiema inteligentnymi indywidualnościami. Człowiek stawia systemowi zadania, a nie tylko — jak w rozwiązaniach prymitywniejszych — steruje jego pracę. Z kolei system komunikuje człowiekowi wysoce przetworzoną i opracowaną informację o swoim stanie, a nie tylko udostępnia wyniki pomiarów ustalonych zmiennych, parametrów i wskaźników. W rezultacie komunikacja pomiędzy człowiekiem i maszyną upodabnia się — w sensie zakresu, charakteru i tematyki — do komunikacji pomiędzy ludźmi, traci natomiast podobieństwo do sterowania maszyn poprzedniej generacji. Tym samym także środki techniczne, używane poprzednio do sterowania: dźwignie, pokrętła, wyłączniki, a także klawiatury przestają być przydatne, natomiast poszukuje się metod i form kontaktu pomiędzy człowiekiem i maszyną dostosowanych do nowych zadań. W tym kontekście rozważać trzeba możliwość i celowość wprowadzenia sygnału mowy jako nośnika informacji wymienianej pomiędzy człowiekiem i maszyną. Rozważania te muszą osobno dotyczyć obydwu kierunków: od maszyny do człowieka i od człowieka do maszyny.

Sygnał mowy jest niewątpliwie najbardziej naturalnym i najszybszym*) sposobem porozumiewania pomiędzy ludźmi, dlatego jego użycie przy przekazywaniu informacji od czy do operatora lub dyspozytora zautomatyzowanego systemu zapewnia w wielu przypadkach nieosiągalny na innej drodze komfort psychiczny. Osiągnięcie takiego komfortu nie jest wyłącznie kwestią wygody człowieka i miarą nowoczesności całej konstrukcji. Przeciwnie, jest to sprawa konkretnych i wymiernych korzyści, gdyż wiadomo już od długiego czasu, że w złożonych systemach, obejmujących zarówno ludzi, jak i maszyny, słabym punktem nieodmiennie okazuje się rejon styku. Zapewnienie operatorowi komfortu w obcowaniu ze sterowaną maszyną wpływa na podniesienie szybkości i trafności jego działania, a równocześnie ogranicza do rozsądnego, akceptowalnego minimum prawdopodobieństwo pojawiania się błędów w pracy operatora. Należy także zwrócić uwagę, że w warunkach szczególnie trudnych (brak oświetlenia, przeciążenia, wibracje,

*) Przekazując informacje za pomocą mowy osiąga się prędkość 50 bit/s przy tempie mówienia 10 głosek/s, natomiast za pomocą dalekopisu można przesłać 30 bit/s przy tempie 60 słów/min. Kod Morse'a pozwala pracować w tempie 6 bit/sek (dla wprawnego operatora).

zagrożenie) sterowanie głosem może okazać się jedyne zapewniające rozsądną sprawność działania. Nie przypadkiem zagadnienia rozpoznawania mowy rozważane są w ośrodkach badań kosmicznych, w instytutach lotniczych, w wojsku.

Ustaliliśmy więc, że rola głosowego, wykorzystującego sygnał mowy, wejścia do systemu sterującego będzie nieodmiennie rosła w miarę postępu w automatyzacji i robotyzacji. Warto dodać, że przy komunikacji w przeciwną stronę, przydatność i użyteczność sygnału mowy nie wydaje się tak bezsporna. Człowiek jest „wzrokowcem", najwięcej informacji w najkrótszym czasie może odebrać i zanalizować za pomocą oczu, nie wspominając o tym, że wzrokowo bardzo łatwo wykrywa się wszelkie prawidłowości, regularności, symetrie itp. własności prezentowanej informacji, lub — przeciwnie — wydobywa się i ustala nieregularności i zakłócenia. Doprawdy wiele głębokiej prawdy jest w chińskiej maksymie, liczącej już blisko trzy tysiące lat: *Jeden obraz daje więcej niż sto słów.* W komunikacji pomiędzy ludźmi informacja obrazowa nie odgrywa tak doniosłej roli, jak by można było oczekiwać po tych wszystkich uwagach, lecz przyczyna tego stanu rzeczy jest trywialna: człowiek nie dysponuje sprawnym efektorem obrazowym, każdy rysunek wymaga pracy — zbyt dużej, jak na potrzeby doraźnego kontaktu. Chociaż — ileż to razy uciekamy się do szkicu, diagramu wykresu w fachowej dyskusji lub chociażby określając drogę do określonego punktu w nieznanym mieście. Natomiast dysponując swobodą wyboru środków przekazywania informacji od systemu sterowania do obsługujących go ludzi możemy wybrać rozwiązanie oparte na graficznej prezentacji informacji — szczególnie, że istnieje obecnie bardzo wiele metod i środków technicznych służących do generacji rysunków i wielobarwnych obrazów przez komputery. Nie oznacza to bynajmniej rezygnacji z omawianego tu sygnału mowy, przeciwnie, syntezatory mowy są urządzeniami bardzo wygodnymi w użyciu i do wielu zastosowań wprost niezastąpionymi (na przykład w zadaniach wymagających przesłania odpowiedzi systemu z wykorzystaniem typowej sieci telefonicznej). Jednak ich względne znaczenie — w stosunku do systemu rozpoznawania mowy, pozwalającego wprowadzać sygnał wprost do komputera — jest wyraźnie mniejsze. Zresztą — o czym była obszernie mowa w p. 2.3 — zadanie generacji mowy z wykorzystaniem sztucznych syntezatorów jest właściwie, od strony koncepcyjnej, całkowicie rozpracowane, pozostają jedynie prace nad doskonaleniem szczegółów technicznych i poprawianiem jakości mowy syntetycznej. W dalszym ciągu tego rozdziału skoncentrujemy więc uwagę głównie na problemach automatycznego rozpoznawania mowy, jako ważniejszych dla systemów automatyki i trudniejszych do praktycznej realizacji. W dziedzinie głosowego „wyjścia" z systemu sterowania poprzestaniemy na dotychczas przytoczonych uwagach, poszerzonych o stwierdzenie, że pewne konkrety na ten temat są zawarte w cytowanym już p. 2.3, a także w zamieszczonej na końcu książki literaturze nawiązującej do tego podrozdziału.

W dość ogólnym zarysie przedstawiono już argumenty przemawiające za stosowaniem układów rozpoznawania mowy do wprowadzania informacji

przez sterującego pracą systemu człowieka do komputera, który pełni bezpośrednie funkcje wykonawcze i regulacyjne. Wspomniano już, że postęp automatyzacji i robotyzacji nie wyeliminował konieczności udziału człowieka w procesach podlegających sterowaniu, zmienił natomiast zakres i charakter jego działań. Uwolniony od czynności bezpośredniego nadzoru i sterowania rozważanego procesu, operator musi wymieniać z nim informacje i polecenia, stawiając mu w trybie dialogowym zadania, odbierając od niego zbiorcze raporty i oceniając realizację wymaganych czynności. Dialog pomiędzy człowiekiem i maszyną stał się faktem, a jakość środków, jakie pozostawi się do dyspozycji człowiekowi, może decydować o niezawodności i jakości wykonania zadań przez cały system, obejmujący zarówno maszynę, jak i współdziałających z nią ludzi. Warto dodać, że dotychczas włożony wysiłek w doskonalenie maszynowego składnika całego „hybrydowego" systemu przyniósł w wielu zastosowaniach ten rezultat, że osiągnięte zostały parametry jakościowe i niezawodnościowe daleko wykraczające poza wartości analogicznych parametrów określanych dla ludzi. Jedyną drogą dalszego postępu jest stworzenie człowiekowi optymalnych warunków pracy, tak aby mógł jak najlepiej wykorzystywać swoje możliwości. W tej sytuacji sterowanie głosowe, ze swoimi zaletami, takimi jak:

— szybkość działania (wypowiedź daje się sformułować sprawniej niż jakąkolwiek manipulację),

— brak związania operatora z jakimkolwiek pulpitem, zestawem manipulatorów, klawiaturą itp.,

— możliwość sprawnego działania w ciemności, w warunkach przeciążenia, stresu fizycznego czy psychicznego,

— naturalność i wygoda sterowania, uwalniające od konieczności długotrwałego treningu i przyuczania personelu,

może stanowić sensowną propozycję w zakresie metod komunikacji człowieka z maszyną. Wyniki stosowania sterowania za pomocą mowy są trudne do wyrażenia w kategoriach ekonomicznych i koncentrują się w sferze zwiększenia efektywności działania. Ostateczny efekt może jednak mieć wymiar ekonomiczny, gdyż sterowanie za pomocą mowy może oznaczać lepsze działanie zautomatyzowanego obiektu i mniejszą uciążliwość pracy dla personelu.

## 5.2. Możliwości automatycznego rozpoznawania mowy

Z uwag zawartych w poprzednim podrozdziale wynikała celowość prac zmierzających do skonstruowania systemu automatycznego rozpoznawania mowy. Ten podrozdział ma z kolei za zadanie pokazanie, że zadanie to jest wykonalne, chociaż w ogólnym sformułowaniu na obecnym etapie jeszcze bardzo trudne i — na razie — nie rozwiązane dla żadnego z rzeczywistych języków. Dokonamy tego dwuetapowo: na początku wymienimy operacje i procesy, które składają się na proces rozpoznawania mowy, a następnie dokładniej je omówimy.

Podstawową operacją, poprzedzającą jakiekolwiek próby rozpoznania, jest

wprowadzenie sygnału mowy do pamięci komputera. Operacja ta bynajmniej nie należy do prostych z uwagi na dużą objętość informacyjną sygnału dyskutowaną w p. 4.6. Trzeba wybrać wewnętrzną postać reprezentacji sygnału mowy w komputerze, a także określić metody jej wprowadzania, rozmieszczania w pamięci, operowania tą nietypową z komputerowego punktu widzenia informacją i dziesiątki innych szczegółów.

Po wprowadzeniu informacji do komputera następuje etap określenia jej parametrów przydatnych do rozpoznawania, to znaczy takich, które redukując w zasadniczym stopniu objętość informacyjną sygnału wydobywają te jego cechy, które są przydatne z punktu widzenia procesu rozpoznawania. Proces ten bywa w większym albo mniejszym stopniu spleciony z uprzednio omówionym, gdyż w celu zaoszczędzenia pamięci urządzenia rozpoznającego pewne cechy wydobywa się przed wprowadzeniem sygnału do maszyny, w innych zaś przypadkach, dysponując maszyną o odpowiednio pojemnej pamięci i dużej szybkości działania, można proces wydobywania parametrów pozostawić do realizacji na drodze czysto cyfrowej, co zawsze jest łatwiejsze, a niekiedy bywa także znacznie bardziej efektywne.

Dysponując wybranym parametrycznym opisem rozpoznawanego sygnału trzeba dokonać jego segmentacji, to znaczy podzielić go na odcinki, podlegające rozpoznawaniu. Problem segmentacji może być w ogólnym przypadku bardzo złożony, gdyż sygnał mowy ma charakter ciągły i jedyne wyraźne granice występują (a i to nie zawsze) pomiędzy wyrazami. Z punktu widzenia segmentacji najwygodniej jest rozpoznawać całe wyrazy, co jednak nie jest optymalnym rozwiązaniem ze względu na inne kryteria. W szczególności liczba podlegających rozpoznawaniu wyrazów musi być bardzo duża — chyba że zdecydujemy się na budowę systemu funkcjonującego z ograniczonym słownikiem dopuszczalnych wyrazów. Ponadto, co jest oczywiste, rozpoznanie wyrazu jest na ogół trudniejsze niż rozpoznanie fonemu — niezależnie od tego, jaką metodą będziemy dokonywać samego rozpoznawania. Z tej argumentacji wynika, że celowe jest dokonanie segmentacji sygnału mowy na fonemy i rozpoznawanie fonemów. Jest ich niewiele, a ponieważ są stosunkowo proste i fonetycznie jednorodne — rozpoznawanie większości z nich (z wyjątkiem spółgłosek płynnych lub nosowych) jest względnie łatwe. Niestety segmentacja ciągłej mowy na fonemy jest bardzo trudna.

Można oczywiście podejść do tego zagadnienia jeszcze w inny sposób. Istnieją metody tzw. analizy skupień, pozwalające ustalić sposób podziału złożonego zestawu danych zgodnie z ich naturalną tendencją do grupowania się. Próba zastosowania tych skupień do grupowania elementów mowy dostarcza segmentów przydatnych do rozpoznawania i dodatkowo łatwych do wydzielenia, bo wynikających z naturalnych tendencji opisujących samą strukturę danych. Być może, że właśnie takie segmenty okażą się najbardziej przydatne przy rozpoznawaniu. Jest to jednak kwestia nadal otwarta, warta dalszych badań. Zagadnienie to będzie dalej dokładniej omówione.

Kolejnym etapem po wydzieleniu segmentów jest ich rozpoznawanie. W literaturze opisano bardzo wiele metod rozpoznawania, przy czym sto-

sunkowo słabo zbadane są ich wzajemne relacje i mało znana jest ich względna przydatność w zadaniu rozpoznawania mowy. Jedyne informacje, na jakie można liczyć, dotyczą wykorzystywania tychże metod w innych, niż rozpoznawanie mowy, zagadnieniach (rozpoznawanie obrazów, identyfikacja złóż surowców mineralnych, prognozowanie pogody, automatyzacja diagnostyki technicznej i medycznej itp.). Opierając się na tych danych możliwe jest wskazanie kilku szczególnie obiecujących metod rozpoznawania i skupienie badań nad rozpoznawaniem mowy na — początkowo przynajmniej — badaniu przydatności tych znanych, wypróbowanych metod. Kolejny etap polega na „złożeniu" elementarnych (być może błędnych w pewnej części) rozpoznań w jedno rozpoznanie globalne. Mówiąc prościej, zachodzi potrzeba przetworzenia sekwencji rozpoznanych segmentów (powiedzmy — fonemów) na rozpoznanie całej wypowiedzi. Jednym z głównych problemów, jaki się przy tym pojawia, jest problem normalizacji czasu. Nawet różne wypowiedzi tej samej kwestii przez tego samego człowieka różnią się znacznie pomiędzy sobą czasem trwania poszczególnych segmentów. Jeszcze większe i bardziej znaczące różnice powstają przy porównywaniu wypowiedzi różnych ludzi. Podkreślić należy przy tym, że zmiana dotyczy całej skali czasu, to znaczy w dwu różnych wypowiedziach tej samej kwestii mogą pojawiać się w sposób trudny do przewidzenia zarówno segmenty krótsze, jak i segmenty trwające dłużej, przy czym łączny czas trwania wypowiedzi tylko w niewielkim stopniu może tu stanowić wskazówkę. Zdarza się bowiem, że w wypowiedzi krótszej niektóre segmenty mogą mimo to trwać dłużej niż w wypowiedzi dłuższej, zatem deformacja skali czasowej ma wybitnie nieliniowy charakter. Ilustrując to trywialnym przykładem można przedstawić następujące ciągi rozpoznanych elementów, wykryte przez automatyczną procedurę rozpoznającą w kilkunastu wypowiedziach wyrazu *sowa*

| | |
|---|---|
| *ssssoooowwaaa* | (wzorcowe nagranie) |
| *csooooffaa* | (obecność szumów) |
| *sssooaaoowwaeaa* | (niewyraźna wymowa) |
| *zzzsuoooaaa* | (głos kobiecy użyty w badaniach) |

Warto zauważyć, że zniekształcenia wypowiedzi na tym etapie dotyczą zarówno czasu trwania poszczególnych segmentów, jak i mogą przejawiać się przekłamanymi rozpoznaniami oraz „gubieniem" niektórych segmentów. Oczywiście problemu tego nie należy wyolbrzymiać. Przytoczone wyżej przykłady zostały celowo, tendencyjnie dobrane w ten sposób, aby ilustrowały możliwe zniekształcenia. W istocie najczęstszym problemem są zmiany skali czasu, a więc pojawianie się niezliczonej liczby wariantów w rodzaju (dla przyjętego przykładu):

*sssssooooowwwwaaaaaaaa*
*ssooooooowwaaaaaaaa*

Usuwanie skutków przytoczonych zjawisk możliwe jest na dwu drogach. Z jednej strony można wykorzystywać słownik wzorców, na którego podstawie wyszukuje się najbliższy, zgodnie z określonym kryterium, wzorzec

wyrazu, odpowiadający (a przynajmniej niesprzeczny z nim) przyjętemu łańcuchowi elementarnych rozpoznań. Z drugiej strony można wykorzystywać reguły kontekstowe do eliminacji błędów i odtworzenia przypuszczalnej prawidłowej postaci rozpoznawanego wyrazu. Pierwsze podejście ma swoje zalety w sytuacji, kiedy rozpoznawaniu podlega ograniczony zbiór wzorców. Drugie stosowane jest wtedy, kiedy dąży się do rozpoznawania możliwie nieograniczonego zbioru wyrazów (na przykład przy próbach konstrukcji „automatycznej sekretarki" — maszyny piszącej pod dyktando). Zaletą pierwszego podejścia bazującego na wzorcach jest istnienie sprawnych i szybko działających algorytmów rozpoznawania, opartych najczęściej na koncepcji programowania dynamicznego, które mogą — przynajmniej w teorii — pracować w czasie rzeczywistym. Zaletą drugiego podejścia jest jego ogólność, okupiona niestety na ogół bardzo dużymi wymaganiami odnośnie do mocy obliczeniowej (pojemności pamięci i szybkości przetwarzania) systemu realizującego wspomniane algorytmy.

Po rozpoznaniu całych wypowiedzi następuje etap ich dalszej analizy, najpierw pod względem strukturalnym (analiza syntaktyczna, rozbiór gramatyczny), a następnie pod względem semantycznym. Zagadnienia związane z tymi etapami procesu analizy będą potraktowane skrótowo, należą bowiem raczej do obszernej i rozwijającej się dziedziny przetwarzania języka naturalnego za pomocą komputerów niż do zagadnień analizy i rozpoznawania mowy jako takiej. Z chwilą bowiem kiedy w wyniku etapu identyfikacji wypowiedzi otrzymamy rozpoznanie badanej wypowiedzi w postaci — przyjmijmy dla przykładu — ciągu znaków odpowiadających poszczególnym fonemom, zagadnienie staje się identyczne z problemami analizy języka naturalnego, badanymi intensywnie w ramach tak zwanej „sztucznej inteligencji" oraz programem szumnie zapowiadanej piątej generacji komputerów. Problemy te polegają — w uproszczeniu to przedstawiając — na tworzeniu reguł wydobywania sensu ze swobodnie formułowanych przez człowieka w języku naturalnym poleceń i informacji, przy czym ze względu na omówione już trudności związane z automatycznym rozpoznawaniem mowy — badania te są z reguły prowadzone na podstawie wypowiedzi wprowadzanych do komputera z użyciem technik znakowych (klawiatur alfanumerycznych, kart i taśm perforowanych, nośników magnetycznych). Zadaniem tej książki nie jest streszczanie czy omawianie bardzo licznych i istotnych osiągnięć, jakie w dziedzinie analizy języka naturalnego odnotowano już w informatyce, chodzi raczej o to, aby zasygnalizować rysującą się tutaj więź prac prowadzonych nad sygnałem mowy z badaniami sztucznej inteligencji, wskazać na wzajemne uwarunkowania postępu w obydwu dziedzinach, a także o to, by z naciskiem podkreślić, że na rozpoznaniu fonemów, sylab czy nawet całych wyrazów problem rozpoznawania mowy bynajmniej się nie kończy. Urządzenie sygnalizujące na ekranie lub wypisujące na drukarce tekst wypowiedzianej i rozpoznanej kwestii ma swoją samodzielną użyteczność — na przykład w systemach telekomunikacyjnych, gdzie takie rozpoznane segmenty mogą być oszczędnie przysyłane przez łącze i użyte do syntezy sygnału mowy w urządzeniu odbiorczym, jednak główny cel, do którego

dążymy, jest znacznie dalszy i wyraźnie ambitniejszy. Nie chodzi o rozpoznanie wypowiedzi, lecz o jej automatyczne zrozumienie. Mówiąc krótko idzie o poprawne wykonanie nakazanej przez wypowiedzianą kwestię czynności. W systemie sterowania, jako wspomożenie obsługującej obiekt sterowania automatyki, znajdzie zastosowanie jedynie taki system, który pozwoli człowiekowi formułować wszystkie zapytania, polecenia i uwagi dla systemu nie tylko głosem, ale również przy zachowaniu pełnej swobody formułowania wypowiedzi w sensie doboru słów, użycia (byle poprawnego) dowolnych form gramatycznych i dokonania dowolnych (byle sensownych) przekształceń wypowiadanej komendy — z pełną gwarancją, że będzie ona właściwie zrozumiana, poprawnie zinterpretowana i bezbłędnie wykonana. Można się spierać, czy takie poprawne z operacyjnego punktu widzenia wykorzystanie treści wypowiedzi można uznać za równoważne jej zrozumieniu, toczone są zażarte polemiki, których celem jest wykazanie „nonsensowności" twierdzeń o maszynowym rozumieniu czegokolwiek, atakowany jest termin „sztuczna inteligencja" — przy czym wszystko to są właściwie spory o słowa. Nikt bowiem nie ma wątpliwości, że konstrukcja systemu zapewniającego możliwość poprawnego interpretowania wypowiedzi jest z praktycznego punktu widzenia bardzo potrzebna, a z technicznego punktu widzenia — całkowicie możliwa, chociaż może jeszcze nie dziś. Bez tego syntaktycznego i semantycznego uzupełnienia prace nad systemami automatycznego rozpoznawania mowy są oderwane od podstawowego praktycznego celu, a osiągane wyniki mogą być traktowane jak zwykłe kuglarstwo.

Podsumowując to ogólne wprowadzenie trzeba stwierdzić, że jako wniosek z przeprowadzonych rozważań można wskazać na wielopoziomową strukturę problemu rozpoznawania mowy. Wyróżnić bowiem możemy:

— poziom akustyczny, związany z wprowadzaniem sygnału do systemu rozpoznającego i jego (ewentualnym) wstępnym przetwarzaniem,

— poziom parametryczny, związany z problemami wydzielenia parametrów sygnału i redukcją jego zapisu do operacyjnie wygodniejszej, a merytorycznie równoważnej formy opisu parametrycznego,

— poziom strukturalny, związany z podziałem sygnału na podlegające rozpoznawaniu segmenty (wraz z problematyką wyboru tych fragmentów i ich optymalizacji),

— poziom identyfikacyjny, związany z metodami automatycznego rozpoznawania wydzielonych fragmentów wypowiedzi oraz z zagadnieniami uczenia, które dla większości metod rozpoznawania są nieodłącznym elementem poprzedzającym proces identyfikacji,

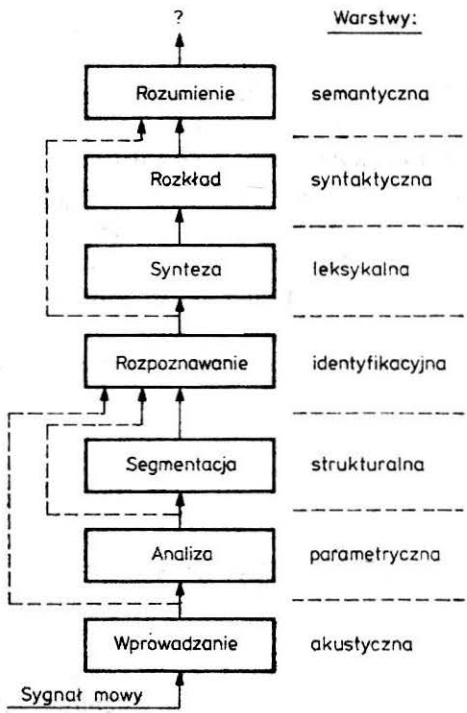
— poziom leksykalny, odpowiedzialny za syntezę rozpoznanych elementów fonetycznych w całościowe elementy rozpoznania — najczęściej wyrazy,

— poziom syntaktyczny, odpowiedzialny za analizę gramatyczną wypowiedzi i zapisanie jej struktury w postaci przydatnej do dalszej analizy,

— poziom semantyczny, związany z problemami identyfikacji treści wypowiedzi i z wydobywaniem jej „sensu".

Ta wielopoziomowa struktura problemu powoduje podobnie wielopo-

ziomowo zorganizowaną, hierarchiczną strukturę systemów rozpoznawania mowy. Wyróżnić w nich można na ogół (rys. 5-1) poszczególne podukłady, odpowiadające wyżej wymienionym poziomom, obrazującym strukturę problemu rozpoznawania mowy i jego składniki. Granice wymienionych podukładów bywają płynne, w konkretnych realizacjach niektóre wydzielone tu piętra hierarchiczne są pominięte, inne zaś mogą się zlewać.




5-1. Struktura problemu rozpoznawania mowy (poziomy wyróżnione po prawej stronie rysunku) i czynności konieczne do zrealizowania dla pełnego rozpoznania i „zrozumienia" sygnału mowy. Niektóre czynności mogą być pomijane (linie przerywane na diagramie). Realizowane obecnie systemy obejmują najwyżej cztery pierwsze poziomy struktury

Nie ma to jednak zasadniczego znaczenia. Wydzielenie odpowiednich poziomów w problemie rozpoznawania i postulat hierarchicznej organizacji struktury układu rozpoznającego posłuży nam do uporządkowania dalszej dyskusji i do systematycznego przedstawienia metod i problemów wchodzących w skład całego zadania.

Tak więc po ogólnym wprowadzeniu, którego celem było wskazanie na możliwości automatycznego rozpoznawania mowy, przystąpimy teraz do zapowiedzianej analizy szczegółowej, prezentując konkretnie, w jaki sposób można skonstruować system rozpoznawania mowy, jakie problemy są już znane i rozpracowane, a jakie zagadnienia stanowią teren dociekań naukowych i prób prototypowych. Kolejność prezentacji zagadnień zgodna będzie z kolejnością ich wprowadzania w dokonanym wyżej przeglądzie problematyki, chociaż z konieczności w pewnych przypadkach wyłamujących się z przyjętego schematu, także schemat opisu będzie modyfikowany.

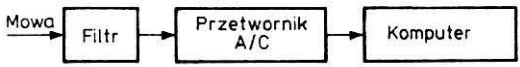
## 5.3. Wprowadzanie sygnału mowy do systemu jej rozpoznawania

Systemy rozpoznawania mowy można podzielić na cyfrowe, analogowe oraz hybrydowe, przy czym kryterium podziału tkwi w oczywisty sposób w naturze sygnału wewnątrz systemu. Sygnał wejściowy jest bowiem zawsze sygnałem analogowym, sygnał zaś wyjściowy (wynik rozpoznania) jest z natury swojej cyfrowy. W najprostszym przypadku sygnałem wyjściowym systemu rozpoznawania mowy jest łańcuch rozpoznanych elementów (a raczej ich kodów), w bardziej złożonych sytuacjach sygnał wyjściowy może mieć formę zbioru konkretnych sygnałów sterujących, w których wyniku dochodzi do wykonania zawartego w wypowiedzi człowieka polecenia. W obu wymienionych skrajnych sytuacjach, a także we wszystkich możliwych do wyobrażenia stanach pośrednich, sygnał wyjściowy z systemu jest wyborem jednej z wielu dyskretnych możliwości — a więc może i powinien być rozważany jako cyfrowy. Podział na systemy analogowe i cyfrowe jest więc dość umowny, w istocie bowiem każdy system rozpoznawania mowy jest hybrydowy, jednak dla konkretyzacji dalszych rozważań przyjmiemy, że interesować nas będą systemy, w których proces wprowadzania sygnału można jeszcze zaliczyć do procesów analogowych, wszystkie dalsze natomiast procesy są czysto cyfrowe. Odpowiada to aktualnym tendencjom obserwowanym w laboratoriach zajmujących się analizą i rozpoznawaniem mowy oraz jest to racjonalne z punktu widzenia konstrukcji urządzeń wykorzystujących rozpoznawanie mowy w praktyce.

Przyjmiemy zatem, że w dalszych podrozdziałach zajmować się będziemy operacjami realizowanymi na drodze tylko cyfrowej, a także zakładamy, że analizę będziemy prowadzić na poziomie struktur algorytmów odpowiednich procesów — rozumiejąc, że jeśli nawet wykonawcą tych operacji nie będzie uniwersalna maszyna cyfrowa, to najtańsza realizacja sprzętowa i tak musi opierać się na zastosowaniu mikroprocesora. Wobec tego budowa specjalizowanego systemu w praktyce oprze się na oprogramowaniu, tyle że zrealizowanym na poziomie języka wewnętrznego mikroprocesora i zapisanym do pamięci stałej systemu. W tym podrozdziale jednak musimy od tego wygodnego punktu widzenia odstąpić i rozważać sygnał mowy w postaci analogowej — takiej jaka jest dostępna na wyjściu przetwornika elektroakustycznego.

Metodami analogowymi musi więc być dokonywane wstępne przetwarzanie sygnału mowy, przynajmniej do etapu filtracji dolnoprzepustowej, odcinającej wszystkie składowe sygnału powyżej częstotliwości Nyquista w celu uniknięcia nakładania się widm (rys. 5-2). Po tym filtrze może znajdować się już układ przetwarzania analogowo-cyfrowego. Pełny sygnał, bez żadnych zmian i korekt, może być przesłany do maszyny cyfrowej, dokonującej wszystkich dalszych niezbędnych transformacji (por. p. 4.1). Taka droga postępowania charakterystyczna jest dla systemów, w których mamy do dyspozycji dużą moc obliczeniową i możemy ją bez ograniczeń angażować dla potrzeb systemu rozpoznawania mowy, a także w tych przypadkach,

kiedy opracowywany system ma charakter eksperymentalny, badawczy. Łatwiej bowiem poszukiwać właściwej drogi przetwarzania sygnału i dobierać różne rozwiązania strukturalne procesu analizy i rozpoznawanie, gdy wszystkie te etapy mają charakter odpowiednich modułów programowych i mogą być zmieniane przez dopisanie lub usunięcie kilku instrukcji.




5-2. Najprostszy system wprowadzania mowy do maszyny cyfrowej. Prostota systemu okupiona jest niestety bardzo dużym zajęciem pamięci komputera przez wprowadzony, nie przetworzony sygnał. Na rysunku tym, podobnie jak na kilku dalszych, użyto skrótowego określenia *mowa* do oznaczenia wejścia, przez które wprowadzany jest sygnał mowy. Zakłada się przy tym, że wejście takie musi być wyposażone w przetwornik elektroakustyczny, wzmacniacz, ewentualnie także w układy formujące sygnał

Znacznie mniej swobody ma eksperymentator w przypadku kiedy proces przetwarzania zdeterminowany jest sprzętowo lub kiedy poprawka w algorytmie wymaga przebudowy licznych układów elektronicznych czy konstrukcji nowych elementów. W takich okolicznościach obok bariery technicznej, utrudniającej prowadzenie badań, pojawia się bariera psychologiczna. Badacz poszukuje rozwiązań stojących przed nim problemów nie w obszarze wszystkich możliwych form i metod przekształcania sygnału, lecz w obszarze wytyczonym przez możliwości wykorzystywanej techniki i dopuszczalne modyfikacje używanej aparatury. Ograniczenie, o którym mowa, jest tym groźniejsze, że funkcjonuje najczęściej w sposób dla samego badacza nieuświadomiony. Pozostawiając jednak na uboczu te metodologiczne dygresje warto uświadomić sobie, że schemat przetwarzania, przedstawiony na rys. 5-2, dlatego jest mało przydatny, że stawia przed częścią cyfrową systemu bardzo wysokie, trudne do zaspokojenia w warunkach polskich, wymagania. Istota problemu tkwi w dyskutowanej w p. 4.6 ogromnej objętości informacyjnej sygnału mowy. Istotnie, jeśli z transmisją sygnału mowy wiąże się strumień informacji o objętości setek tysięcy bitów na sekundę, to analiza odcinków mowy obejmujących całe wypowiedzi — nawet proste polecenia lub komendy dla systemu automatyki — wymaga komputerów o megabajtowych pamięciach, a szybkość wykonywania operacji, wymagana przy obliczaniu, przekracza wszelkie rozsądne granice, jeśli wymaga się pracy systemu w czasie rzeczywistym.

Zatem nie na zasadzie wyboru optymalnego wariantu, lecz przyciśnięci do muru dysproporcją potrzeb i możliwości, badacze sygnału mowy i konstruktorzy urządzeń automatycznego rozpoznawania tego sygnału decydują się na rozbudowę analogowej części aparatury i na dokonywanie procesów wstępnego przetwarzania sygnału jeszcze przed jego wprowadzeniem do komputera. Przetwarzanie, o którym mowa, może kierować się w stronę różnych parametrycznych i bardzo oszczędnych reprezentacji sygnału mowy, względnie może ograniczać się do przekształcenia sygnału do postaci widma dynamicznego (za pomocą zestawu filtrów, demodulatorów, układów uśredniających itd., zgodnie z zasadami podanymi w p. 4.2). To drugie rozwiązanie bywa zwykle preferowane, ponieważ objętość informacyjna

widma dynamicznego sygnału mowy jest na tyle mniejsza od objętości oryginalnego sygnału, że dokonane ograniczenie wystarcza do rozsądnego pomieszczenia wymaganych odcinków sygnału mowy w pamięciach komputerowych o łatwo dostępnych pojemnościach. Równocześnie przekształcenie sygnału do postaci jego widma dynamicznego jest — nawet metodami analogowymi — łatwe do przeprowadzenia. Co więcej stosunkowo rozpowszechniona i dostępna jest profesjonalna, wysokiej jakości aparatura, umożliwiająca dokonywanie takiej transformacji.

Tak więc z jednej strony prostota operacji zmierzających do znalezienia wymaganej formy sygnału, z drugiej natomiast zadowalający wynik w postaci wystarczającego ograniczenia objętości sygnału preferują łącznie użycie krótkookresowej analizy widmowej do badania charakterystyk sygnału i do wprowadzania ich do maszyny cyfrowej. Oczywiście podejmując decyzję o zastosowaniu zestawu filtrów pasmowych do wydzielenia charakterystyk sygnału mowy, przydatnych do jego wprowadzania do systemu rozpoznającego, musimy dodatkowo określić dużą liczbę szczegółowych parametrów tego procesu wstępnego rozpoznawania i przetwarzania, którego własności mogą w decydujący sposób wpływać na jakość procesu rozpoznawania w całości. Trzeba bowiem mieć świadomość, że dokonując wstępnego przetwarzania sygnału mowy przed jego wprowadzeniem do systemu rozpoznającego, bezpowrotnie tracimy pewną część informacji. Zresztą to właśnie jest celem wstępnego przetwarzania. Problem jedynie w tym, żeby tracona informacja była — z punktu widzenia celu rozpoznawania — bezwartościowa, natomiast aby tracić jak najmniej informacji użytecznej. Postulat taki łatwiej sformułować, niż zapewnić jego realizację.

Przyjmując, co wydaje się wysoce prawdopodobne, że niezbędna informacja mieści się w charakterystyce amplitudowo-częstotliwościowej sygnału, oraz zakładając, że do określenia charakterystyki posłużymy się zestawem filtrów analogowych, pozostaje nadal wiele pytań szczegółowych, na które należy udzielić odpowiedzi, zanim dokończy się projektu systemu wprowadzania mowy do maszyny cyfrowej. Są to między innymi następujące zagadnienia:

— ile pasm częstotliwości zamierzamy wyróżnić,

— czy mają być one rozłożone liniowo, czy w sposób logarytmiczny (stała szerokość pasma, czy stały stosunek szerokości do częstotliwości środkowej pasma),

— jak szerokie zastosować okno czasowe i jakim rodzajem okna się posłużyć,

— jaką przyjąć metodę demodulacji sygnału (prostowanie dwupołówkowe, podnoszenie do kwadratu, detekcja impulsowa itp.),

— jak dokonywać uśredniania sygnału (liniowo, wykładniczo czy według innej funkcji wagowej) oraz jaki ma być czas uśredniania sygnału,

— jak często próbkować sygnały wyjściowe używanych filtrów,

— jaką dokładność amplitudową zapewnić przy przetwarzaniu sygnałów wyjściowych z filtrów (ile przyjąć poziomów dyskryminacji amplitudy i jak je rozmieścić — równomiernie, czy według zasady gęściejszego obsadzenia poziomów o niższych amplitudach),

— jak rozmieszczać informację o sygnale mowy w pamięci komputera (czy przeznaczać jedną komórkę lub bajt na pojedynczy odczyt amplitudy sygnału w pojedynczym pasmie częstotliwości i w jednym ustalonym momencie czasu, czy też „upakowywać" informację na pojedynczych bitach słowa maszynowego),
— jakich kodów użyć do rejestracji odczytów z poszczególnych filtrów (naturalnych binarnych, Graya, z zabezpieczeniem przed przekłamaniami czy bez nich),
— ile bitów przeznaczyć na zapamiętanie pojedynczego kwantu informacji.
Przytoczona lista jest bez wątpienia niekompletna. Pominięto w niej bardzo wiele zagadnień szczegółowych, ściśle technicznych, na przykład: wybór techniki realizacji filtrów (bierne czy aktywne, rezonansowe czy drabinkowe, LC czy RC), metod uśredniania (analogowe czy cyfrowe) czy konkretnego typu używanego konwertera (bezpośredni, wagowy, całkujący, kombinowany) oraz sposobu jego zastosowania (czy ma być jeden konwerter dla wszystkich filtrów i układ musi wzbogacić się o komutator kanałów analogowych czy też użyty będzie w każdym torze oddzielny konwerter). Przytoczona lista zagadnień ma jedynie sygnalizować, jak wiele problemów wiąże się z prostą pozornie i nie budzącą wątpliwości koncepcją dokonywania wstępnego przetwarzania sygnału mowy przed jego wprowadzeniem do komputera — także wówczas, kiedy już się podejmie kluczową dla dalszych działań decyzję, że typ przyjętego przetwarzania będzie wynikał z zastosowania krótkookresowej transformaty Fouriera. Na wszystkie przytoczone tu i pominięte pytania trzeba konkretnie i szczegółowo odpowiedzieć przy budowie układu wstępnego przetwarzania sygnału mowy. Przyjęte odpowiedzi — determinujące strukturę i działanie zbudowanego systemu — wynikają z głębokiej analizy własności sygnału, podlegającego przetwarzaniu, z rozważenia pozostających do dyspozycji możliwości sprzętowych (zarówno w zakresie używanego komputera, jak i w zakresie aparatury analogowej, którą po odpowiedniej adaptacji zamierzamy wykorzystać w torze wstępnego przetwarzania mowy), a także z arbitralnych rozstrzygnięć, wynikających z osobistych preferencji badacza prowadzącego próby rozpoznawania mowy. Na ten ostatni składnik każdej podejmowanej decyzji zwraca się na ogół zbyt mało uwagi, tymczasem marginesy pozostawione w tym miejscu przez ścisłą wiedzę są dość szerokie i wpływ arbitralnych rozstrzygnięć może być w sumie znaczący. Rezultatem takiego stanu rzeczy są trudności wynikające przy próbach porównywania wyników uzyskiwanych przez różne zespoły.
Liczba wyróżnianych pasm częstotliwości jest najważniejszym i w największym stopniu arbitralnie wybieranym parametrem. Wydaje się, że proponowana niekiedy liczba 5 wyróżnionych pasm częstotliwości (biorąca się z rozumowania, że jest to najmniejsza liczba pozwalająca wykrywać obecność lub brak trzech pierwszych formantów) jest niewystarczająca. Również zbyt mała wydaje się liczba kilkunastu wyróżnionych pasm częstotliwości proponowana przez niektórych autorów na podstawie doświadczeń z wokoderami pasmowymi (por. p. 6.2). Doświadczenie wykazuje, że dla systemu

rozpoznawania mowy niezbędna jest rozdzielczość odpowiadająca kilkudziesięciu — blisko stu — wydzielonym pasmom częstotliwości. Przy większej liczbie pasm zasadniczy cel stosowania analizy, mianowicie ograniczenie objętości informacyjnej sygnału, staje się problematyczny. Przy mniejszej (wyraźnie mniejszej) liczbie pasm mowa zachowuje informacje wystarczające do jej zidentyfikowania przez człowieka, natomiast ilość informacji okazuje się zbyt mała dla algorytmów dokonujących rozpoznawania na drodze automatycznej. Powtarza się tu omawiana wcześniej sytuacja (por. p. 4.1), w której pewien kwant informacji (poprzednio była to częstość przejść przez zero) jest dostateczny dla rozpoznania sygnału mowy przez najdoskonalszy system rozpoznający, a mianowicie przez mózg człowieka, natomiast konstruowane algorytmy i urządzenia rozpoznające nie potrafią tej informacji równie doskonale spożytkować.

Przyjmując zatem, że wymagane jest kilkadziesiąt pasm częstotliwości, w których sygnał będzie rozważany i analizowany, możemy rozpatrywać sposób rozłożenia tych pasm wzdłuż osi częstotliwości. Zasadniczy problem dotyczy wyboru zasady: pasma o stałej, czy o zmiennej szerokości, a w następstwie podjęcia decyzji w tej kwestii wybór jednej z możliwych skal częstotliwości — liniowej lub logarytmicznej. Doświadczenia z wokoderami (por. p. 6.2) oraz tradycja obowiązująca w badaniach akustycznych przemawiają za wyborem skali logarytmicznej i stałej procentowej szerokości pasma. Wydaje się jednak, że są poważne argumenty przemawiające przeciw takiemu podejściu. Sygnał mowy po wprowadzeniu do maszyny poddawany jest dalszemu przetwarzaniu, stosuje się do niego kolejne algorytmy i dokonuje jego parametrycznego opisu, wobec tego liniowa skala częstotliwości okazuje się z reguły wygodniejsza w użyciu. Zresztą przy pokrywaniu przedziału (typowo przyjmowanego) od stu do kilkunastu tysięcy herców za pomocą blisko stu filtrów — większość argumentów przemawiających typowo za stosowaniem skali logarytmicznej staje się nieaktualna. Użycie liniowej skali częstotliwości ma tę dodatkową zaletę, że w przypadku uzyskania dostępu do sprzętu obliczeniowego o większej mocy, po dokonaniu analizy widmowej algorytmem FFT otrzyma się widmo o liniowej skali częstotliwości (por. p. 4.2). W takim przypadku stosowanie w analogowym systemie wstępnego przetwarzania filtrów o skali liniowej gwarantuje możliwość natychmiastowego wykorzystania programu FFT do wszystkich opracowanych wcześniej algorytmów przetwarzania sygnału, wydobywania parametrów, rozpoznawania itd. Użycie na wstępie filtrów o skali logarytmicznej prowadzi do konieczności przerabiania całego oprogramowania w momencie pojawienia się możliwości pełnej „cyfryzacji" systemu.

Sprawa okna czasowego wiąże się z tym, jakie fragmenty mowy bardziej nas interesują. Do analizy głosek szumowych, których charakter jest w dużej mierze przypadkowy, optymalne jest stosowanie okna o dużej szerokości, zapewniającego dłuższe uśrednianie widma i dającego stabilniejszy obraz widma. Natomiast głoski o szybko zmieniającym się widmie — na przykład plozyjne — wymagają okna wąskiego, aby eliminować przy obliczaniu widma wpływ fragmentów poprzedzających i następujących po interesują-

cym tranzjencie, otrzymując w rezultacie wierny i nie zniekształcony obraz procesu przejściowego, którego kształt i parametry decydują zwykle o skutecznym rozpoznaniu głoski. Wybór konkretnej wartości długości okna jest więc kompromisem i jak każdy wybór kompromisowy — ma charakter arbitralny. Literatura niekiedy zaleca okno o szerokości 20 ms, wydaje się to jednak wartością za dużą. Celowe wydaje się rozważenie okna o długości około 10 ms, chociaż do konkretnych badań, koncentrujących uwagę badacza na przebiegach przejściowych, nawet takie okno może się okazać za długie. Co do kształtu okna czasowego, to w całej rozciągłości znajdują tu zastosowanie uwagi przytoczone w p. 4.3. Nie powtarzając przytoczonej tam szczegółowej dyskusji zagadnienia warto wskazać na okno Gaussa jako najkorzystniejsze w sensie braku listków bocznych w charakterystyce częstotliwościowej, ale praktyczna realizacja krzywej Gaussa może napotykać — przy stosowaniu metod czysto analogowych — poważne trudności. Z tego powodu zamiast okna Gaussa bywają stosowane inne typy wymienionych w p. 4.3 okien czasowych — szczególnie okno Hamminga.

W odniesieniu do metod demodulacji sygnału z filtrów często stosowane jest prostowanie dwupołówkowe zamiast poprawniejszego metodologicznie, ale bardzo uciążliwego w realizacji, podnoszenia do kwadratu. Wydaje się, że w zadaniu rozpoznawania mowy, gdzie nie zależy nam na dokładnym pomiarze mocy sygnału w poszczególnych pasmach (do takiego pomiaru podnoszenie do kwadratu amplitud sygnału jest absolutnie konieczne), lecz na przekazaniu do maszyny informacji o kształcie obwiedni widma i jego czasowych zmianach, postępowanie uproszczone, z zastosowaniem prostownika, jest całkowicie wystarczające.

Podobnie mało istotna jest przyjęta reguła uśredniania sygnału. Ze względu na prostą realizację i korzystne własności używania jest zwykle reguła uśredniania z „ważeniem wykładniczym" realizowana za pomocą prostego układu RC. Nie oznacza to jednak, że dysponując odpowiednimi możliwościami nie powinno się dążyć do wykorzystania i zbadania właściwości innych rodzajów uśredniania. Ważny jest natomiast bez wątpienia wybrany czas uśredniania. Obowiązują tu podobne kryteria jak przy wyborze długości okna czasowego. Typowo dla sygnału mowy przyjęło się zakładać czas uśredniania sygnału około 10 ms, co — jak było wyżej pokazane — stanowi kompromis między sprzecznymi wymogami wynikającymi z konieczności analizy segmentów mowy o różnych właściwościach.

Problem wyboru częstości próbkowania sygnałów wyjściowych z filtrów jest bardzo ważny. Z jednej strony bowiem częstość próbkowania jest drugim (po liczbie pasm częstotliwości) podstawowym parametrem wyznaczającym objętość wynikowego zbioru danych wprowadzanych do maszyny cyfrowej dla ustalonego odcinka sygnału mowy. Z tego punktu widzenia korzystne jest stosowanie częstości próbkowania najmniejszej, jak się tylko da. Z drugiej, wiązani jesteśmy częstotliwościami zmian obwiedni sygnałów na wyjściach filtrów, która (por. p. 4.1) musi być mniejsza od połowy przyjętej częstości próbkowania. Naturalnie stosując filtrację dolnoprzepustową sygnału wyjściowego z demodulatorów zamontowanych na wyjściach po-

szczególnych filtrów można bez trudu uzyskać dowolnie małą częstotliwość graniczną — gubiąc jednak bezpowrotnie informacje o tych partiach sygnału mowy, w których widmo zmienia się w sposób szybki. Ponieważ szybkie zmiany widma sygnału odpowiadają szybkim ruchom artykulacyjnym narządów mowy, przeto na ogół niosą one znacznie więcej informacji o wypowiadanych wyrazach niż długotrwałe nawet okresy, w których widmo się nie zmienia, gdyż narządy mowy pozostają nieruchome i trwa artykulacja stanów quasi-ustalonych samogłosek lub spółgłosek szumowych. Wynika z z tego, że przy rozpoznawaniu mowy te właśnie szybkie zmiany widma sygnału będą bardzo przydatne, a ich utrata w momencie wprowadzania sygnału mowy do komputera nie da się później w żaden sposób zrekompensować. Częstotliwość próbkowania widma musi być więc największa, jaka tylko jest wymagana do przeniesienia wszystkich rejestrowanych zmian w widmie. Ponieważ uprzednio przyjęto okno czasowe około 10 ms i tej samej długości czas uśredniania rekomendowano na wyjściach filtrów, przeto sugeruje się, że częstość próbkowania sygnałów wyjściowych z filtrów powinna być około 100 Hz. Oznacza to, że maksymalna częstość zmian widma, która będzie wiernie oddana we wprowadzonym do komputera sygnale, wynosić będzie około 50 Hz. Jest to zakres wystarczający, gdyż badania nad wokoderami pasmowymi wykazały, że graniczne częstotliwości sygnałów w pasmach częstotliwości (zdemodulowanych) wynoszą od 20 do 35 Hz. Trzeba jednak bardzo starannie odfiltrować częstotliwości większe od przyjętej częstotliwości granicznej (Nyquista), gdyż łatwo tu o zakłócające nakładanie się widm — szczególnie, że sygnał na wyjściach filtrów poddawany jest jedynie uśrednianiu i mogą w nim występować tętnienia o częstotliwości odpowiadającej częstotliwości środkowej filtru. Ta filtracja dolnoprzepustowa jest często pomijana w strukturze członów wprowadzających sygnał mowy do komputera, co może być źródłem znacznych zakłóceń rejestrowanego i przetwarzanego sygnału.

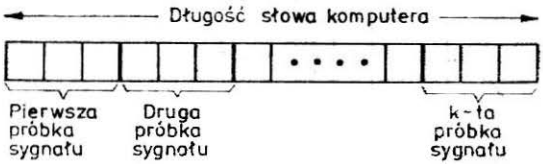
Dokładność amplitudowa przetwarzania analogowo-cyfrowego, następującego w każdym kanale częstotliwościowym oddzielnie lub za pomocą jednego przełączanego przetwornika, odgrywa w sumie mniejszą rolę, niż można przypuszczać. Z pozoru jest to kolejny, trzeci wymiar warunkujący informacyjną objętość wprowadzonego do komputera sygnału. Jednak zakres możliwych zmian tego parametru jest niewielki, gdyż wymagana liczba poziomów jest także (na ogół) bardzo mała. Dynamika sygnałów w poszczególnych pasmach jest niewielka, znacznie mniejsza od dynamiki pełnego sygnału mowy. W dodatku dokładność odwzorowania amplitud sygnału nie ma w zadaniu rozpoznawania mowy tak wielkiego znaczenia, gdyż ważniejsze są relacje między sygnałem w sąsiednich pasmach (na przykład dla poprawnej lokalizacji formantów) niż dokładne wartości. W praktyce oznacza to, że wystarcza przetwarzanie kilkubitowe, przykładowo rekomendować można w tym zastosowaniu przetwornik pięciobitowy, którego zastosowanie gwarantuje (por. p. 4.1) odtworzenie dynamiki sygnału (w pasmach częstotliwości) nie gorsze niż 30 dB, podczas gdy wyniki badań wskazują, że obserwowana dynamika sygnału nie przekracza 20 dB. W prostszych

systemach rozpoznawania mowy stosuje się zresztą przetworniki o mniejszej liczbie bitów — do jednobitowych włącznie. Nie kwestionując przydatności zbinaryzowanego widma stwierdzić jednak należy, że poprawniejsze z punktu widzenia jakości uzyskiwanych wyników jest stosowanie przetwarzania wielobitowego na wejściu i dokonywanie formowania widma na drodze obliczeniowej, z progiem dyskryminacji amplitud dostosowanym do aktualnej (lokalnej) wartości sygnału. Przykładową techniką, która może okazać się przy tym użyteczna, jest technika histogramowa. Określając częstość występowania w pewnym rejonie widma poszczególnych (dyskretnych) wartości amplitud sygnału można łatwo ustalić właściwą dla danego odcinka sygnału wartość granicznego poziomu amplitudy sygnału, powyżej którego odpowiednie fragmenty widma oznaczane będą jako 1, a poniżej jako 0.

Liczba poziomów dyskryminacji amplitudy w stosowanych przetwornikach analogowo — cyfrowych, używanych dla wprowadzania sygnału mowy w postaci widma dynamicznego do systemu komputerowego, jest niewielka — od kilku do 32 (dla przetwornika pięciobitowego). Poziomy te byłoby korzystnie rozmieścić nierównomiernie (por. p. 4.1). W praktyce się tego jednak nie stosuje, gdyż komplikuje się przy tym zarówno budowa przetwornika, jak i struktura algorytmów, wykorzystujących przetworzony sygnał w maszynie cyfrowej. Nieopłacalne jest także — pożądane z merytorycznego punktu widzenia — zróżnicowanie poziomów przetwarzania dla poszczególnych pasm częstotliwości, zgodnie ze znaną regułą, że maksimum energii mieści się dla sygnału mowy w niskich zakresach częstotliwości, ze wzrostem zaś numeru pasma moc sygnału maleje — w przybliżeniu odwrotnie proporcjonalnie do częstotliwości środkowej pasma. Zamiast zróżnicowanych amplitudowo poziomów przetwarzania w poszczególnych pasmach stosuje się więc w praktyce preemfazę, to znaczy wstępne formowanie sygnału wprowadzanego do systemu. Preemfaza podnosi poziom energetyczny składników o dużych częstotliwościach, przy czym dla preemfazy realizowanej w postaci różniczkowania sygnału korekta opadającej charakterystyki naturalnego sygnału mowy jest prawie idealna.

Z niewielką liczbą bitów przeznaczonych do zapisu pojedynczej wartości amplitudy w ustalonym momencie czasu i w określonym pasmie częstotliwości wiąże się kolejny z wymienionych problemów. Chodzi o sposób upakowania informacji akustycznej w pamięci maszyny cyfrowej. Formalnie rzecz ujmując, naturalna reprezentacja sygnału mowy przetworzonego w wyżej omówiony wstępny sposób polega na użyciu tablicy o liczbie kolumn odpowiadających liczbie wyróżnionych pasm częstotliwości i liczbie wierszy zgodnej z liczbą wyróżnionych momentów czasu. Taki sposób reprezentacji jest jednak w najwyższym stopniu rozrzutny: każdy element tablicy odpowiada jednej komórce pamięci używanego komputera, co oznacza, że do jego reprezentacji użytych jest tyle bitów, ile wynosi długość słowa używanego komputera. W praktyce jest to więc 16, 32, a nawet 60 bitów — a w rzeczywistości potrzeby nie przekraczają 5 bitów, gdyż taką dokładność miał przetwornik analogowo-cyfrowy. Aby zaoszczędzić pamięć

komputera, stosuje się więc zabieg „upakowywania" informacji, polegający w uproszczeniu na rozmieszczaniu w jednej komórce pamięci kilku niezależnych informacji, rozmieszczonych odpowiednio na fragmentach słowa używanego komputera (rys. 5-3). „Upakowanie" informacji zmniejsza objętość zajętej przez sygnał mowy pamięci komputera, jednak nie odbywa się to za darmo. Komplikują się odpowiednio wszystkie programy wykorzystujące

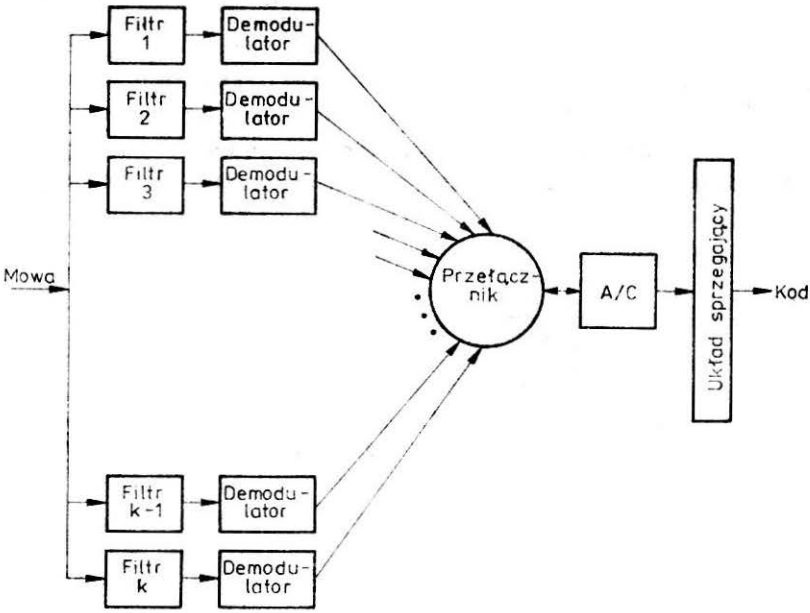



5-3. Sposób upakowania informacji akustycznej w komórkach pamięci komputera. Krotność upakowania, oznaczona na rysunku *k*, może być bardzo duża, jej wartość zależy jednak od długości (liczby bitów) próbki sygnału i od długości słowa maszynowego. Upakowywanie jest nieopłacalne (na ogół) w maszynach o strukturze bajtowej

zgromadzoną głosową informację, gdyż najprostsze nawet procedury przetwarzania wymagają operacji „rozpakowywania" danych i „upakowywania" wyników. Aby czas wykonywania operacji upakowania nie był zbyt duży, konieczne jest pisanie modułów programowych związanych z pakowaniem na poziomie języka wewnętrznego komputera, a to jest kłopotliwe, dlatego upakowanie warto stosować głównie wtedy, gdy jest duża dysproporcja między najmniejszym dostępnym (adresowalnym) kwantem pamięci, a liczbą bitów przeznaczaną do zapisania pojedynczej wartości amplitudy sygnału. Jeśli zysk z upakowania jest niewielki (na przykład kiedy używany komputer lub mikrokomputer ma organizację bajtową), wówczas upakowywanie się nie opłaca i powinno być pominięte.

Problem kodu użytego do zapamiętania pojedynczych kwantów informacji oraz problem liczby bitów na pojedynczy kwant są ze sobą powiązane. Jeśli zamierzamy użyć kodu o specjalnych własnościach lub jeśli przewidujemy możliwość pisania programów przetwarzających informację upakowaną w pamięci w sposób „półrównoległy" (to znaczy bez rozpakowywania), wówczas liczba bitów rezerwowanych w pamięci komputera do zapisania pojedynczej wartości amplitudy musi być większa, niż to wynika z liczby poziomów używanego przetwornika analogowo-cyfrowego. Najczęściej jednak stosowany jest kod BCD, w którym wspomniane efekty nie występują.

Podsumowując przytoczone rozważania można zaproponować strukturę systemu wprowadzania sygnału mowy do maszyny cyfrowej w postaci przedstawionej na rys. 5-4. Jest to — jak wynikało z treści tego podrozdziału — jedna z wielu możliwych koncepcji i struktur, ale jej użycie wydaje się uzasadnione. W kolejnych rozdziałach będziemy rozważali algorytmy operujące na przetworzonym i wprowadzonym do komputera sygnale mowy, które realizować będą kolejne etapy procesu rozpoznawania, wymienione w p. 5.2 i przedstawione na rys. 5-1.



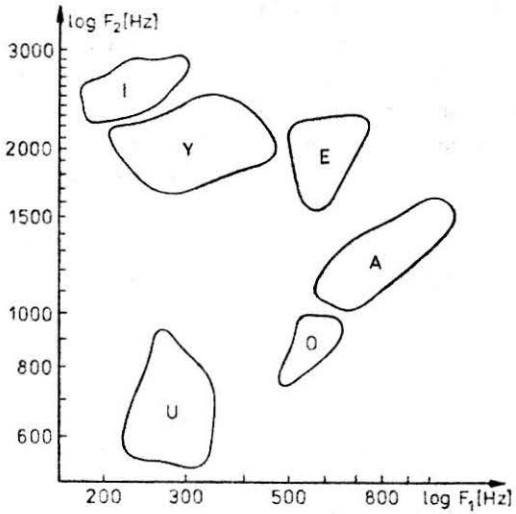

5-4. Struktura układu wstępnego przetwarzania i wprowadzania sygnału mowy do maszyny cyfrowej. Filtry i demodulatory dokonują pasmowej, krótkookresowej analizy widma sygnału, przełącznik tworzy kod szeregowy z wyjść poszczególnych demodulatorów. Przetwornik A/C zwykle jest pojedynczy ze względu na koszt i powtarzalność warunków przetwarzania w poszczególnych kanałach. Układ sprzęgający wprowadza przetworzone próbki sygnału do urządzenia liczącego

## 5.4. Wydzielanie parametrów przydatnych przy rozpoznawaniu

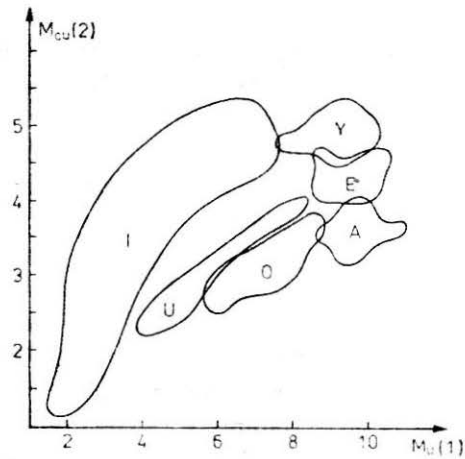
Algorytmy omawianego teraz etapu pełnią rolę przygotowawczą. Sygnał mowy, nawet wstępnie przetworzony i wprowadzony do maszyny cyfrowej, nie stanowi właściwie podstawy dla algorytmów realizujących proces rozpoznawania. Przeszkodą jest tu z jednej strony zbyt duża objętość sygnału, z drugiej zaś jego niedogodna struktura. Jak się okaże z dalszych rozważań, najdogodniejsza forma danych dla algorytmów rozpoznawania, grupowania czy segmentacji polega na stosowaniu wektorów cech. Wektor taki, o ustalonej i na stałe przyjętej wymiarowości, zawiera informacje pozwalające na prawidłowe rozpoznawanie. Najłatwiej ocenić to przy zastosowaniu kryteriów geometrycznych. Wprowadzając przestrzeń cech, w której poszczególne osie odpowiadają wydzielonym oddzielnym cechom (parametrom sygnału) możemy sprecyzować wymagania co do pożądanego wektora cech w następującej postaci. W prawidłowo wybranej przestrzeni cech obiekty*) identyczne pod względem fonetycznym powinny grupować się

*) Pojęcie obiektu, które zostało tu nieformalnie użyte, można sprecyzować w następujący sposób. Każdy ustalony, wydzielony fragment sygnału mowy może być opisany przez zestaw swoich parametrów. Parametry te można uporządkować w formę wektora i w rezultacie każdemu fragmentowi sygnału mowy (na przykład każdej próbce widma dynamicznego) przypisać w rozważanej przestrzeni pewien punkt. Punkt ten, a dokładniej jego położenie względem innych punktów, może być przedmiotem (obiektem) rozpozna-

i skupiać w ustalonym obszarze przestrzeni, możliwie najbardziej odległym od skupisk odpowiadających innym klasom. Zilustrowano to przykładowo na rys. 5-5, na którym obszary skupień odpowiadające poszczególnym samogłoskom języka polskiego na płaszczyźnie (bo tylko taka, dwuwymiarowa przestrzeń daje się narysować), której osie wyznaczają pierwszy i drugi formant. Wyraźne rozdzielenie skupisk odpowiadających samogłoskom i stosunkowo zwarty kształt obszarów, odpowiadających poszczególnym skupiskom dowodzą, że dwa pierwsze formanty niosą bardzo dużo użytecz-




5-5. Obszary lokalizacji poszczególnych samogłosek języka polskiego na płaszczyźnie pierwszego i drugiego formantu. Obszary są rozdzielne, jest więc możliwe rozpoznawanie samogłosek opierając się jedynie na wartościach dwu wskazanych formantów




5-6. Obszary lokalizacji poszczególnych samogłosek języka polskiego na płaszczyźnie momentów widmowych. Rozdzielczość jest tu gorsza, niż dla formantów, ale momenty dostarczają cennych informacji i uzupełniają możliwości rozpoznawania na zbiór spółgłosek — szczególnie szumowych

nej informacji i w kontekście zadania rozpoznawania samogłosek dostarczają wystarczającej informacji do ich rozpoznawania. Nieco gorzej wypadają w tej ocenie niektóre inne parametry. Przykładowo na rys. 5-6 pokazano kształt podobnych obszarów dla poszczególnych samogłosek w przestrzeni wyznaczanej przez momenty widmowe (por. p. 4.4). Widać, że „rozmycie" obszarów poszczególnych klas jest teraz większe, a ich rozseparowanie w proponowanej przestrzeni — gorsze, chociaż nie aż tak złe, by miało to stanowić przesłankę do wnioskowania o złej separowalności i w następstwie do niemożności rozpoznawania samogłosek w tej przestrzeni. Wniosek ten jest zgodny z uwagami, jakie poczyniono przy wprowadzaniu momentów widmowych do parametrycznego opisu sygnału mowy. Nadają się one głów-

wania, wobec tego będziemy używać nazwy obiekt do określenia wydzielonego segmentu mowy, opisanego wytypowanym zestawem parametrów. Każdy obiekt, to punkt w przestrzeni, której bazę definiuje przyjęty zestaw parametrów, a określona zbiorowość obiektów, na przykład wszystkie próbki sygnału określonej głoski, wymawianej przez różne osoby, to skupisko punktów lub podobszar przestrzeni.

nie do opisu głosek szumowych lub jako parametry uzupełniające opis dany innymi parametrami, jednak w przypadku bardzo prostych głosek — a do takich należą bez wątpienia samogłoski — możliwe jest ich rozpoznawanie na podstawie wyłącznie momentów widmowych.

Rysunki 5-5 i 5-6 stanowią przykład sposobu oceny parametrów (cech) na podstawie obrazu rozkładu obiektów poszczególnych rozróżnianych klas w przestrzeni generowanej przez wybrane parametry. Są one bardzo przydatne do tego, aby zrozumieć, jakie koncepcje wiążą się z geometrycznym podejściem do zagadnienia oceny parametrów jako cech, zachowujących minimalną niezbędną ilość informacji, wystarczającą do rozpoznawania i klasyfikacji poszczególnych fragmentów sygnału mowy. Przydatność takich rysunków do rzeczywistej oceny obiektów jest jednak ograniczona, a to z powodu konieczności operowania w przestrzeniach wielowymiarowych. Kryteria skupienia obiektów należących do jednej klasy i rozproszenia poszczególnych klas muszą więc w rzeczywistych zastosowaniach podlegać formalizacji matematycznej, tak aby rozstrzygające znaczenie miała wyliczana matematycznie wartość kryterialna. Formalizacja taka może być stosunkowo łatwo przeprowadzona. Wystarczy określić w przestrzeni cech pojęcie odległości między poszczególnymi punktami oraz zdefiniować odległość punktu od zbioru oraz zbioru od zbioru. Pozostawiając chwilowo sprawę wyboru konkretnej postaci tych odległości możemy stwierdzić, że formułowane wyżej postulaty zwartości obiektów w poszczególnych wyróżnionych klasach oraz dostatecznie dobrego rozseparowania klas w rozważanej przestrzeni cech sformułować można następująco.

Niech $i$-temu wyróżnionemu odcinkowi sygnału mowy odpowiada wektor parametrów $X^i$, którego składowe $x_1^i, x_2^i, \ldots, x_n^i$ oznaczają wartości branych pod uwagę parametrów opisujących sygnał mowy i formujących (w myśl przytoczonych wyżej rozważań) wykorzystywaną przestrzeń cech. Warto zauważyć, że wymiar przestrzeni $n$ (odpowiadający liczbie wyróżnionych parametrów) nie może być zbyt duży, gdyż w przeciwnym przypadku korzyści wynikające ze stosowania parametrycznego opisu mowy stają się problematyczne. Równocześnie jednak w nietrywialnych przypadkach $n > 2$, zatem przydatność rysunków podobnych do 5-5 i 5-6 ogranicza się do orientacyjnego przedstawienia na przekrojach lub rzutach rozkładu obiektów wzdłuż ustalonych płaszczyzn.

Proponowane dalej ujęcie formalne jest wolne od ograniczeń związanych z wymiarem przestrzeni i może znaleźć zastosowanie dla dowolnie dużych $n$, chociaż uciążliwość rachunków w przytoczonych wzorach rośnie w przybliżeniu proporcjonalnie do $n^2$.

Podstawą dalszych rozważań jest pojęcie odległości obiektu opisanej wektorem $X^i$ od obiektu opisanej wektorem $X^j$. Oznaczmy tę odległość $d^{ij}$, przy czym oczywiście $d^{ij} \geqslant 0$ oraz $d^{ij} = d^{ji}$. Rozważając wszystkie obiekty $X^i$ należące do ustalonej, wyróżnionej klasy $i$ (na przykład wszystkie próbki sygnału odpowiadające określonej głosce) możemy określić dla nich $N_k(N_k-1)/2$ odległości (gdzie $N_k$ jest liczbą badanych obiektów należących do klasy $k$). Zakładamy, że mamy ustaloną regułę, zgodnie z którą można

określić jedną wartość odległości, którą uznamy za charakterystyczną dla całej klasy $k$ i będziemy uważali za miarę rozrzutu obiektów wewnątrz tej klasy. Ta pojedyncza miara może być wybrana na wiele sposobów. Może to być średnia odległości wszystkich obiektów klasy, wartość maksymalna tych odległości, maksymalna lub średnia odległość obiektu od „środka ciężkości klasy" itp. W dalszych rozważaniach uważamy, że ta pojedyncza charakterystyczna wartość została dla każdej klasy ustalona i wynosi $d_k$. Przyjmując, że rozważamy $L$ klas obiektów mamy więc już $L$ wartości $d_k$ ($k = 1, 2, \ldots, L$). Wszystkie te wartości będą musiały współuczestniczyć w tworzeniu funkcji kryterialnej, stosowanej do oceny przydatności określonego zbioru parametrów.

Rozważając dalej wszystkie pary klas musimy zaproponować miarę rozsunięcia, rozseparowania, czy — mówiąc krócej — również miarę odległości, ale tym razem całych klas. Rozważając klasę $k$, zawierającą $N_k$ elementów, oraz klasę $m$, zawierającą $N_m$ elementów, możemy określić $N_k N_m$ odległości, które mogą służyć jako tworzywo przy budowie miary odległości klasy $k$ od klasy $m$, którą oznaczymy dalej $D_{km}$. Znowu jest tu do dyspozycji wiele możliwości. Można posłużyć się definicją metryki Hausdorfa, można wybrać arbitralnie odległość maksymalną, minimalną lub średnią. Wybór ten, po jego dokonaniu, będzie rzutował na dokładność końcowego wyniku i na ocenę rozważanych zestawów parametrów.

Niestety, pomimo pozorów zmatematyzowania kryteria naszego wyboru pozostają (częściowo przynajmniej) arbitralne, z uwagi na konieczność wyboru sposobu obliczania $d^{ij}$, $d_k$, $D_{km}$ — i dalsze, również arbitralne wybory. Matematyka nie zawsze bowiem oznacza obiektywizm oceny — chociaż łatwo można o tym zapomnieć, szczególnie posługując się komputerem.

Wybrawszy jedną z wymienionych (lub dowolną inną) ewentualność możemy przystępować do próby oceny. Na tym etapie rozważań mamy miary rozrzutu obiektów we wszystkich klasach $d_k$ oraz miary odstępu między wszystkimi klasami $D_{km}$ ($k, m = 1, 2, \ldots, L$). Ocena powinna być w sumie tym wyższa, im większe będą $D_{km}$ i im mniejsze będą $d_k$. Ostateczna postać formuły matematycznej, określającej funkcję kryterium

$$Q = Q(D_{12}, D_{13}, \ldots, D_{L-1,L}, d_1, d_2, \ldots, d_L) \tag{5.1}$$

musi być wybrana przez badacza zgodnie z jego preferencjami. Sugerować można jedynie w charakterze przykładowego, sprawdzonego w działaniu rozwiązania, wzór postaci

$$Q = \frac{\min\limits_{i,j} D_{ij}}{\max\limits_{i} d_i} \tag{5.2}$$

Stosując wzór (5.2) można uznać rozważany zestaw cech za zadowalający, jeśli $Q > 1$. Zazwyczaj jednak trzeba się zadowalać gorszymi wynikami, które wszakże bynajmniej nie muszą oznaczać gorszej jakości rozpoznawania. Istota rzeczy bowiem polega na tym, aby dla tych klas, których rozrzut jest niepokojąco duży, zapewnić dostatecznie duży dystans od sąsiednich

klas w celu możliwości prawidłowej separacji. Natomiast dla tych klas, które mają mały rozrzut, dopuszczalny jest też mniejszy dystans od klas sąsiednich. Wzór (5.2) ma więc charakter asekurancki, zbyt surowy i może niekiedy sugerować celowość odrzucenia zestawów cech w istocie bardzo przydatnych przy rozpoznawaniu. Niestety, trudno jednak proponować inne miary, gdyż zawsze da się dobrać przykład tak dobranych danych, że utworzone przez te dane struktury w przestrzeni cech będą źle separowalne przy poprawnych wartościach funkcji kryterialnej. Użycie wzoru (5.2) można zatem traktować jako zło konieczne do czasu opracowania doskonalszych kryteriów.

Wracając do konkretów trzeba stwierdzić, że zaproponowane kryterium (przy wszystkich jego mankamentach) pozwala na wybór parametrów, za których pomocą będziemy opisywać sygnał mowy przed jego rozpoznawaniem, wcześniejsze rozważania zaś (por. p. 4.4 i ewentualnie 4.5) pozwalają na generację takich parametrów. Zresztą, co warto podkreślić, w tym zakresie jest jeszcze wiele do zrobienia i tu właśnie najłatwiej można wnieść nowy, znaczący wkład do badań nad rozpoznawaniem mowy.

Tymczasowo jednak skupimy się na koncepcjach znanych i uznanych. Poza dyskusją wydaje się być celowość włączenia formantów do zbioru parametrów przydatnych do rozpoznawania. Ich znaczenie w procesie artykulacji i naturalnej percepcji mowy, a także liczne potwierdzone w praktyce pozytywne próby rozpoznawania mowy z wykorzystaniem formantów, stanowią tu argumenty, z którymi trudno polemizować. Tak więc pierwsze trzy składowe proponowanego wektora cech, używanego dalej przy rozpoznawaniu, są identyczne z wartościami trzech pierwszych częstotliwości formantowych

$$x_i = F_i \quad i = 1, 2, 3 \tag{5.3}$$

przy czym w razie braku (lub niemożliwości wykrycia) określonego $i$-tego formantu odpowiednia wartość $x_i = 0$ na zasadzie definicji. Formanty nie wystarczają jednak do rozpoznawania niektórych ważnych klas głosek, na przykład głosek szumowych (trących) i dlatego wymiar wektora cech musi być rozszerzony. Trudno przesądzać, które parametry są najbardziej predestynowane do tego, aby zająć kolejne pozycje, wydaje się jednak, że korzystne własności w tych przypadkach, dla których zawodzą formanty, wykazują momenty widmowe (por. p. 4.4) i z tego względu można przyjąć, że kolejne dwa elementy wektora cech $X$ powinny mieć postać

$$x_4 = M_u(1) \tag{5.4}$$

$$x_5 = M_{cu}(2) \tag{5.5}$$

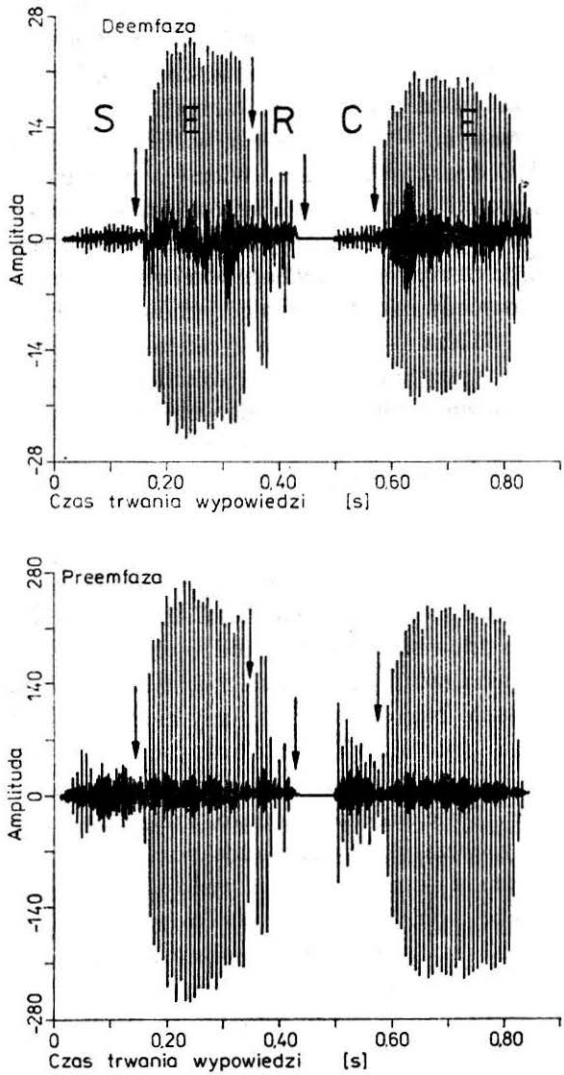
Oczywiście istnieje bardzo wiele innych parametrów, które mogą okazać się przydatne w zadaniach rozpoznawania mowy. Ich definicje i sposoby praktycznego obliczania są przedmiotem licznych publikacji, a poszczególni Autorzy mogą na bardzo przekonywających przykładach udowodnić, że ich propozycje doskonale nadają się do rozpoznawania pewnych, ustalonych klas elementów mowy. Podstawową zaletą wektora cech $X$ powinna jednak być zwięzłość. Wymagając krótkiego i zwartego opisu nie możemy mnożyć

parametrów w nieskończoność, gdyż wkrótce może się okazać, że objętość informacyjna parametrycznego opisu sygnału jest porównywalna z objętością źródłowego sygnału, co jest efektem bez wątpienia w najwyższym stopniu niekorzystnym. Poprzestaniemy zatem na pięciowymiarowym ($n = 5$) wektorze cech, opisującym wybrane próbki sygnału mowy. Rozpoznawanie, klasyfikacja, segmentacja sygnału — wszystkie te operacje będą prowadzone w pięciowymiarowej przestrzeni, co obok innych zalet ma jeszcze tę, że stwarza przesłanki do przetwarzania i rozpoznawania sygnału w czasie rzeczywistym. Na problem ten, wzmiankowany już w p. 4.4, warto teraz zwrócić dodatkowo uwagę. Wiemy z poprzedniego podrozdziału, że kolejne próbki widma sygnału wprowadzane są do komputera rozpoznającego mowę w odstępach czasowych około 10 ms. Zachowanie warunków pracy w czasie rzeczywistym wymaga w tej sytuacji takiej definicji wektora cech $X$ oraz opracowania takich algorytmów określania jego składowych, aby proces obliczania wartości wszystkich $x_i$ nie trwał dłużej niż 10 ms. Tylko w takim przypadku program określania parametrów będzie nadążał za strumieniem napływających danych, przetwarzając je na bieżąco do dogodnej dla dalszych etapów rozpoznawania formy. Samo rozpoznawanie może trwać nieco dłużej, gdyż po zakończeniu wypowiedzi (komendy) człowiek jest skłonny zaczekać pewien czas na reakcję maszyny — chociaż i tu obowiązują ograniczenia czasowe. Jeśli czas reakcji systemu będzie się nadmiernie wydłużał, to człowiek może się zdekoncentrować z fatalnym skutkiem dla sterowanego procesu. Aby więc nie dopuścić do utraty ciągłości dialogu człowieka z maszyną, odpowiedź (lub wymagana reakcja w postaci wykonania zadanej czynności) musi nastąpić nie później, niż po około dwu sekundach. Niekorzystne jest, gdy następuje szybciej (człowiek czuje się wówczas „poganiany" przez maszynę i ten dyskomfort odbija się niekorzystnie na efektywności pracy), fatalne jednak jest, jeśli następuje znacznie później. Dwie sekundy to dużo, komputer może w tym czasie wykonać miliony operacji, jednak biorąc pod uwagę złożoność algorytmów rozpoznawania — to mało, zbyt mało, by z tego czasu „pożyczać" część na proces wydobywania parametrów sygnału. Tak więc graniczny czas około 10 ms określa możliwości stosowania do opisu sygnału mowy liczniejszych i bardziej wyrafinowanych parametrów. Pozostaniemy więc przy wektorze cech, opisanym wzorami (5.3) ÷ (5.5), gdyż — jak wykazuje doświadczenie — łatwo dostępne w Polsce komputery zapewniają przy takim wektorze cech warunki pracy w czasie rzeczywistym, zawartość potrzebnej informacji zaś w tak określonym zbiorze parametrów wydaje się być wystarczająca do skutecznego rozpoznawania większości elementów mowy polskiej.

## 5.5. Problem segmentacji ciągłego sygnału mowy

Jak wspomniano wyżej (por. p. 5.2), jednym z zasadniczych problemów, jakie musi podjąć i rozstrzygnąć badacz zajmujący się rozpoznawaniem mowy, jest wybór elementów podlegających rozpoznawaniu. Wybór ele-

mentu zbyt dużego (na przykład wyrazu lub sylaby) wiąże się z koniecznością przechowywania w pamięci systemu bardzo dużej liczby wzorców*). Wybór elementu małego i wygodnego, jakim bez wątpienia jest fonem, wiąże się z koniecznością segmentacji. W ciągłym sygnale mowy granice między fone-



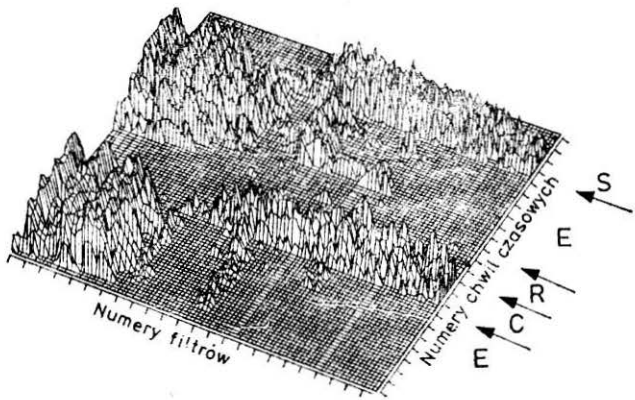

5-7. Przebieg czasowy wypowiedzi *serce*. Człowiekowi dość łatwo jest dokonać segmentacji tego przebiegu, a wydzielone segmenty w dość naturalny sposób odpowiadają fonemom. Granice między segmentami zaznaczono strzałkami

mami są zatarte. Co więcej — o czym już wcześniej była również mowa — dla poprawnej identyfikacji fonemu mogą być potrzebne informacje leżące poza jego teoretycznymi granicami (na przykład głoski zwarte można łatwo

*) Można przyjąć, że do pełnego rozpoznawania mowy polskiej trzeba zgromadzić słownik przynajmniej 500 tysięcy odmiennych form wyrazów — chyba że system wyposażymy w obszerny moduł gramatyczny, zdolny do uporania się z fleksyjnością mowy polskiej, gdyż wówczas wystarczy zapamiętać około 120 tys. wzorców wyrazów dla rozpoznawania dowolnych wypowiedzi i około 30 tys. dla specjalnych podzbiorów języka, względnie konieczne jest pamiętanie ponad 2500 wzorców sylab.

rozpoznawać jedynie wtedy, gdy możliwe jest śledzenie zmian formantów w poprzedzających i następujących po nich samogłoskach). Niemniej problem segmentacji istnieje i wymaga rozwiązania.

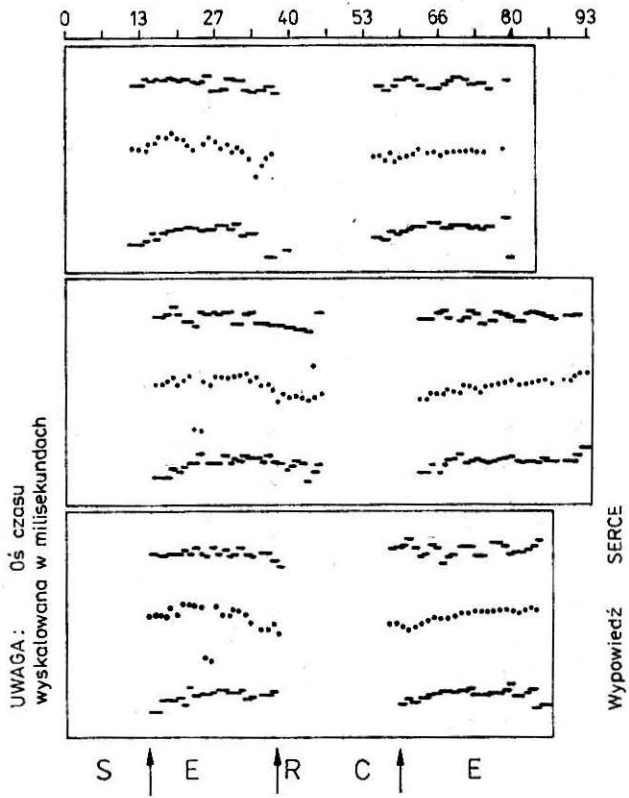
Większość proponowanych w literaturze metod rozpoznawania granic między segmentami opiera się na stwierdzeniu, że sygnał na granicy zmienia istotnie swój charakter. Istotnie, rozważając przedstawiony na rys. 5-7, przebieg wypowiedzi *serce* łatwo można zauważyć, że w punktach oznaczonych strzałkami charakter sygnału zmienia się dość radykalnie, co w większości przypadków jest charakterystycznym elementem pozwalającym wykryć granice między fonemami. Równie dobitnie można to prześledzić na rys. 5-8, obrazującym widmo dynamiczne rozważanego sygnału. Tu także



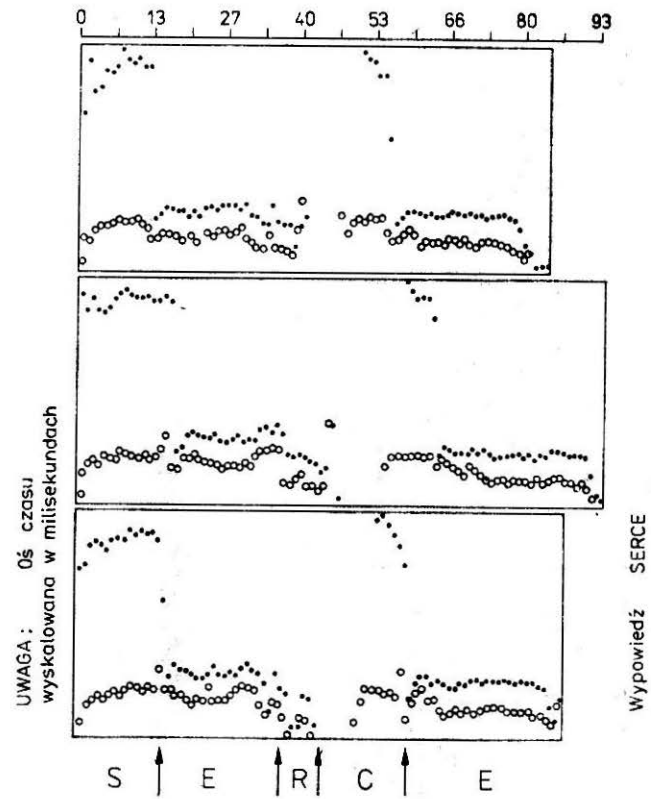

5-8. Segmentacja jest również możliwa na podstawie obrazu widma dynamicznego, którego zmiany (zaznaczone strzałkami) odpowiadają granicom segmentów

granice między fonemami cechują się wyraźną zmianą charakteru sygnału. Wszystkie te rozważania prowadzą jednak do stwierdzeń mało przydatnych w praktyce komputerowego rozpoznawania mowy: „zmiana charakteru" sygnału, łatwo zauważalna przez człowieka i dość oczywista dla nawet mało doświadczonego obserwatora, jest bardzo trudna do opisania w kategoriach algorytmu komputerowego. Jak bowiem ująć ilościowo „charakter" sygnału, jak wyrazić jego zmianę i na jakiej podstawie wyrokować, że jedna zmiana jest „wyraźna", a inne nie? Jak w dodatku zapewnić stosowną do potrzeb szybkość tego procesu, aby segmentacja nie stała się „wąskim gardłem" procesu rozpoznawania? Na te pytania niełatwo udzielić odpowiedzi, a propozycje, znajdujące się w literaturze, podzielić można na trzy grupy: albo autorzy prac unikają problemu segmentacji przyjmując, że rozpoznawanymi obiektami są całe wyrazy (w dodatku wyraźnie rozdzielane przy wymawianiu), albo prowadzi się badania nad rozpoznawaniem elementów wydzielonych z ciągłego sygnału mowy „ręcznie" (przez odpowiednio kwalifikowanego operatora), albo wreszcie proponowane są algorytmy rozdzielania i segmentacji sygnału mowy — ale tak skomplikowane i pracochłonne, że ich zastosowanie w praktyce wydaje się problematyczne.

W tej sytuacji celowe jest poszukiwanie rozwiązań niekonwencjonalnych. Zamiast szukać metod segmentacji sygnału mowy na elementy, przyjęte dla tego sygnału w sposób sztuczny, lepiej poszukiwać elementów, na które




5-9. Segmentacji można dokonywać także na podstawie wartości parametrów. Prezentowana na rysunku mapka formantów wypowiedzi *serce* pokazuje, że granice segmentów mogą być tu także wykryte — chociaż nie wszystkie — na przykład granica między *r* a *c*




5-10. Granice segmentów są również widoczne na mapce momentów widmowych. W tym przypadku możliwe było zlokalizowanie wszystkich granic. Łącznie z obrazem formantów momenty dają prawie stuprocentową gwarancję poprawnej segmentacji. (Zaznaczone granice dotyczą najniżej zlokalizowanej mapki wypowiedzi).

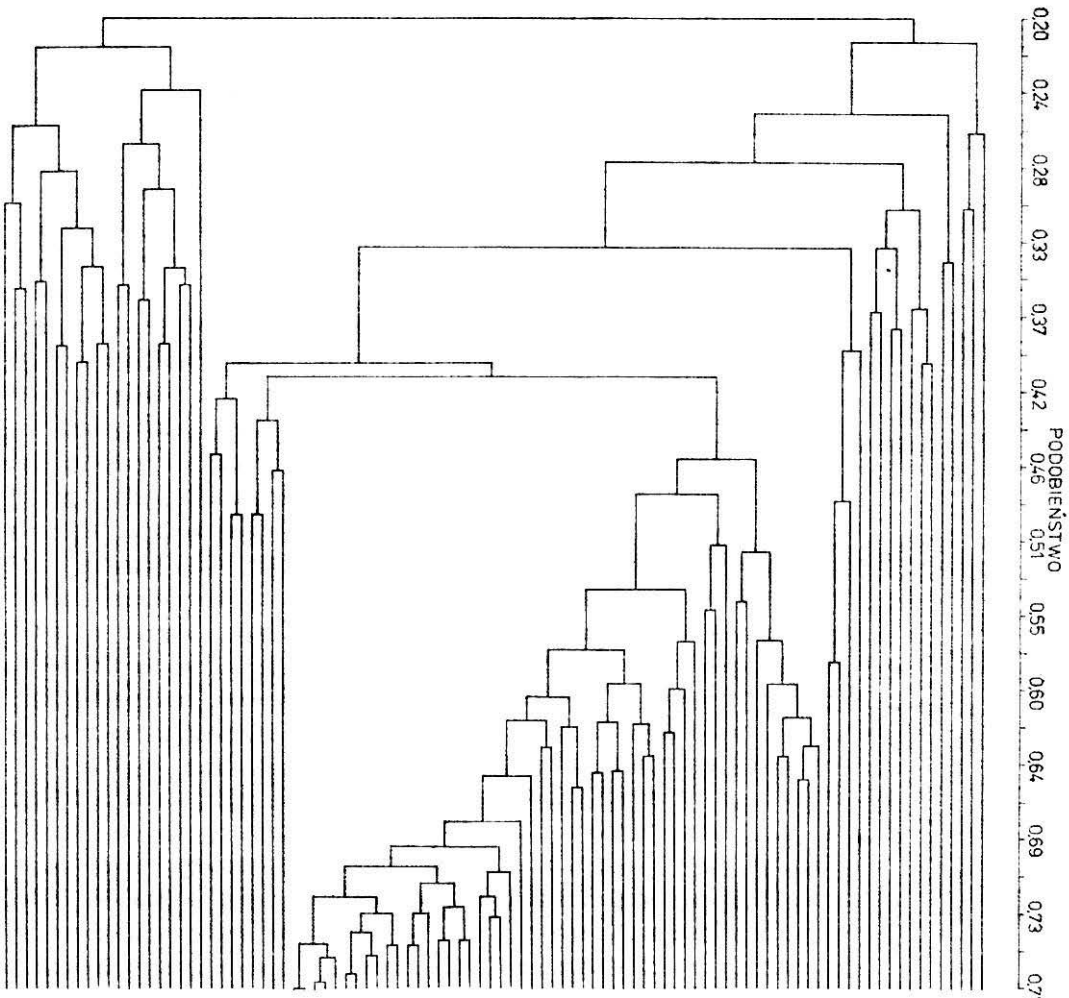
sygnał mowy dzieli się w sposób naturalny, na zasadzie podobieństwa jego struktur. Oprzemy się przy tym na parametrach i cechach, wydzielonych przez algorytmy omówione w poprzednim podrozdziale. W przestrzeni parametrów granice między segmentami są bowiem również zauważalne (por. rys. 5-9 i 5-10, na których pokazano „mapki" zamian formantów i momentów widmowych dla tego samego sygnału, co na rys. 5-7 i 5-8), a prowadzenie segmentacji jest znacznie mniej pracochłonne. Zauważmy przy tym, że pewna, bardzo arbitralna i nie związana z jakimikolwiek naturalnymi strukturami w sygnale mowy, ale realna, segmentacja dokonywana jest już w momencie wprowadzania sygnału mowy do komputera. Istotnie, proces próbkowania czasowego kwantuje widmo — dzieli na odcinki nazywane niekiedy widmami chwilowymi. Można sobie wyobrazić sytuację, że właśnie widma chwilowe będziemy traktować jako podlegające rozpoznawaniu segmenty, które roboczo można nazwać mikrofonemami. Składnik „mikro" w proponowanej nazwie sugeruje, że podlegająca rozpoznawaniu jednostka jest mniejsza od fonemu, że fonem może być zdefiniowany jako określona sekwencja takich podjednostek oraz że rozmiary (czasowe) podlegającego rozpoznawaniu elementu są najmniejsze z możliwych.

Powstaje jedynie problem, z jakimi wzorcami porównywać mikrofonemy, jak je klasyfikować i jak rozpoznawać. Nie ulega wątpliwości, że wzorców mikrofonemów musi być więcej niż wzorców fonemów. Wynika to z faktu, że między fonemami w mowie ciągłej występują stany przejściowe, które zresztą bywają bardzo użyteczne z punktu widzenia procesu rozpoznawania. Jeśli więc rozważamy najprostszy zestaw fonemów, na przykład wyraz *As*, to możemy w nim oczekiwać co najmniej pięciu wzorców mikrofonemów: segmentu odpowiadającego narastaniu głoski *a* (segment przejściowy typu „cisza-a"), segmentu ustalonego głoski *a*, segmentu przejściowego pomiędzy *a* i *s*, segmentu ustalonego głoski *s* oraz segmentu zanikania głoski *s* (przejście „s — cisza"). Rozpatrując to zagadnienie w podobny do podanego sposób możemy oczekiwać kilkuset mikrofonemów (przy 40 fonemach liczba oczekiwanych mikrofonemów sięga 820 wzorców, co bynajmniej nie wyczerpuje wszystkich możliwości, gdyż niektóre fonemy nawet w swoim „stanie ustalonym" prezentują na przemian kilka wzorców widma — na przykład głoska *r* — inne zaś fonemy mają kontekstowo zależne odmiany brzmieniowe o zróżnicowanym kształcie widma i rozmaitych wartościach używanych do opisu parametrów).

Liczba ta jest stanowczo zbyt duża, aby rozważaną grupę wzorców traktować jako dobry zbiór rozpoznawanych elementów. Sposobów redukcji wskazanej grupy wzorców najdogodniej poszukiwać metodami opierającymi się na wzajemnej bliskości obiektów w przestrzeni cech. W tym celu każde widmo chwilowe rozważanego zbioru próbek sygnału mowy (im obszerniejszy ten zbiór próbek — tym lepiej) traktujemy jako punkt w przestrzeni cech. W poprzednim podrozdziale ustalono i ustalenie to obowiązuje nadal, że przestrzeń cech ma pięć wymiarów, wynikających z pięciu mierzonych dla każdego widma parametrów sygnału: trzech formantów i dwu momentów widmowych. W tej pięciowymiarowej przestrzeni dokonuje się następnie

grupowania obiektów, wykorzystując do tego celu znane algorytmy analizy skupień (ang. *cluster analysis*). Algorytmy te dokonują łączenia bliskich sobie (w sensie określonej miary) obiektów tworząc skupienia, możliwe w dalszych rozważaniach do zastąpienia przez pojedynczych reprezentantów.

Proces tworzenia skupień można prowadzić generalnie na dwa sposoby. Pierwszy, nazywany aglomeracyjnym, polega na kolejnym łączeniu bliskich sobie obiektów (pierwotnych, pochodzących z rozważanego zbioru danych, lub wtórnych — będących reprezentantami wcześniej utworzonych skupień), aż do uzyskania pożądanej liczby skupień lub do utworzenia skupień o wymaganych własnościach. Drugi sposób, nazywany podziałowym, polega na traktowaniu pierwotnie całej zbiorowości dostępnych danych jako jednego



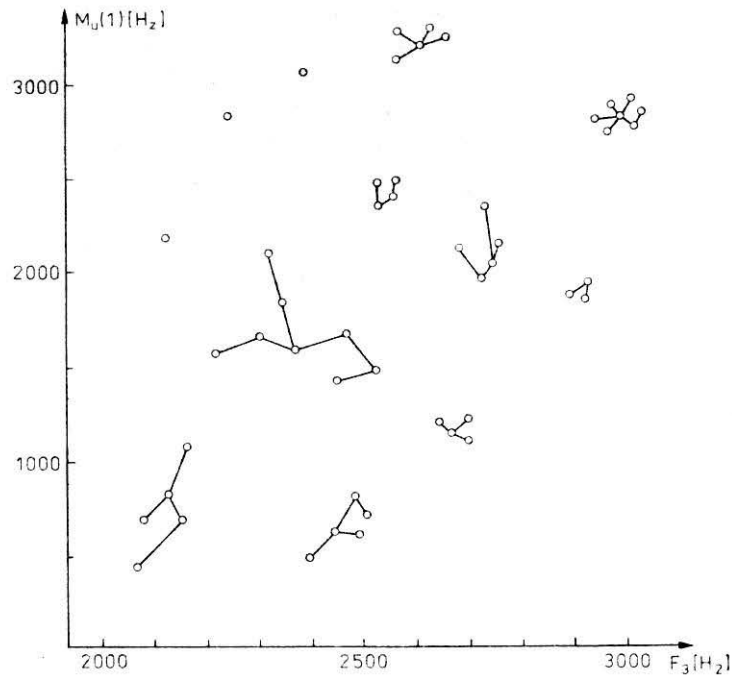

5-11. Dendrogram obrazujący kolejność łączenia próbek sygnału mowy w procesie grupowania jednym z algorytmów analizy skupień. Po lewej stronie rysunku naniesiono skalę stopnia podobieństwa. Algorytmy podziałowe funkcjonują na zasadzie przechodzenia wskazanego grafu od dołu do góry, algorytmy aglomeracyjne odpowiadają przechodzeniu od góry do dołu; w obydwu przypadkach ważny jest moment przerwania procesu grupowania, decydujący o przydatności zbudowanych skupień

dużego skupienia, które następnie jest kolejno dzielone na mniejsze skupienia metodą rozcinania najdłuższego połączenia (dzięki temu podziały separują obiekty lub skupienia najbardziej oddalone od siebie w przestrzeni cech).

Ponieważ obok rozróżnienia na metody podziałowe i aglomeracyjne istnieje wiele dalszych rozróżnień i podziałów wykorzystywanych algorytmów analizy skupień, przeto ostateczny wynik zastosowania dyskutowanych metod zależy po części od charakteru dzielonych danych, po części jednak również od użytej metody — ma więc znowu poniekąd arbitralny charakter. Kolejny element dowolności wprowadza kryterium zatrzymania używanego algorytmu. Istotnie, bez ustalenia konkretnych warunków zatrzymania, programy działające według metody podziałowej nie zatrzymują się, dopóki nie utworzą osobnego „skupienia" z każdego analizowanego obiektu, algorytmy aglomeracyjne będą zaś działały do momentu utworzenia jednego wielkiego skupienia ze wszystkich obiektów połączonych razem. Pożądany wynik leży oczywiście między tymi skrajnościami i trzeba go „wyłowić" zadając odpowiedni warunek „stopu". Nie jest to proste, a w dodatku — jak wspomniano — wnosi do rozważań kolejny element arbitralności.

Wynik pracy algorytmów grupowania najlepiej rozważać w formie dendrogramu pokazującego, jakie obiekty i z jakim stopniem podobieństwa (bliskości) zostały w kolejnych krokach algorytmu połączone lub rozdzielone. Na przykład, na rys. 5-11 pokazano dendrogram procesu grupowania widm chwilowych sygnału mowy, stanowiący podstawę do wytypowania pewnej ustalonej grupy wzorców przy rozpoznawaniu. Taka forma prezentacji jest w praktyce jedyną możliwą, ponieważ stosowane niekiedy w podręcznikach prezentacje w postaci obrazów rozkładu obiektów w przestrzeni cech oraz połączeń między nimi są tu nieprzydatne, gdyż przestrzeń, w której dokonuje się grupowania, jest pięciowymiarowa. Jedynie do demonstracji można wybrać dowolne dwa z pięciu używanych wymiarów przestrzeni i pokazać powiązania przykładowego zbioru obiektów — zrzutowane na wybraną płaszczyznę. Na rysunku 5-12 pokazano układ punktów odpowiadających wybranym obiektom zrzutowany na płaszczyznę trzeciego formantu i pierwszego momentu widmowego. Zaznaczone powiązania między obiektami odpowiadają pewnemu początkowemu etapowi grupowania metodą aglomeracyjną. Na dalszych etapach kreślenie podobnej mapki jest utrudnione ze względu na liczne, wielokierunkowe powiązania między obiektami i ich grupami, co zaciemnia obraz. Podkreślić należy, że rysunki 5-11 i 5-12 stanowią jedynie ilustracje pokazujące istotę zastosowanych metod grupowania, ponieważ — niezależnie od sposobu prezentacji — przytoczenie wyników rzeczywistych badań nastręcza poważne trudności ze względu na liczbę obiektów, na podstawie których określa się grupy. Rzeczywistym celem grupowania jest bowiem, przypomnijmy, znalezienie wzorców mikrofonemów, które mogą pełnić rolę punktów odniesienia przy rozpoznawaniu. Badania prowadzone w Zakładzie Biocybernetyki AGH w Krakowie przez dra Andrzeja Izworskiego pozwoliły na ustalenie następujących prawidłowości. W wyniku zastosowania analizy skupień do ponad 30 tysięcy widm

chwilowych stanowiących próbki rzeczywistego sygnału mowy (pojedynczy głos męski, nagranie w komorze bezechowej) wyróżniono i zlokalizowano metodami aglomeracyjnymi początkowo 973 wzorce widm. Następnie prowadzono (na podstawie macierzy odległości tych wzorców widm) proces tworzenia skupień do momentu pojawienia się mikrosegmentów łączących ewidentnie różne z fonetycznego punktu widzenia elementy sygnału mowy. W momencie przerwania programu analizy skupień (po około $70 \cdot 10^3$ s




5-12. Rzutowanie skupień elementów sygnału mowy na jedną z możliwych płaszczyzn w przestrzeni cech (w prezentowanym przykładzie jest to płaszczyzna trzeciego formantui drugiego momentu)

obliczeń komputera Cyber 72) wyróżnionych było 270 mikrosegmentów, spośród których większość (203 segmenty) stanowiła mało liczne (poniżej 10 widm chwilowych) skupienia widm „zakłóceniowych" (nietypowe obrazy widma poszczególnych głosek, przebiegi przypadkowe, zniekształcone zapisy itp.). Analizie poddano wyłącznie pozostałe 67 segmentów, wśród których stwierdzono istnienie 52 segmentów odpowiadających pojedynczym fonemom (najczęściej stanom ustalonym fonemów), 15 skupień zaś grupowało typowe przebiegi transjentowe (stany przejściowe między ustalonymi fonemami). Niestety, mimo znacznej liczebności zbioru wyróżnionych mikrosegmentów (mikrofonemów) nie dla każdej głoski udało się zidentyfikować odpowiadające jej skupienie. Są wprawdzie głoski, dla których udało się określić jedno lub — częściej — kilka skupień widm (przykładowo dla wszystkich samogłosek, sylabicznych i niesylabicznych, głoski *r*, spółgłosek trących dźwięcznych *v*, *z*, ʒ), są jednak niestety i takie skupienia, które odpowiadają kilku różnym głoskom (na przykład spółgłoski zwarte łączą

się wszystkie w jedno skupienie, podobnie wspólne skupienia mają głoski trące i zwarto-trące).

Jak z tego wynika, rozpoznawanie oparte na koncepcji mikrofonemów nie może (w chwili obecnej, gdyż możliwy jest w tej dziedzinie postęp przy wykorzystaniu bardziej adekwatnych zbiorów cech) gwarantować całkowicie poprawnego rozpoznania wszystkich elementów mowy. Jeśli jednak uwzględni się dodatkową informację wnoszoną przez segmenty transjentowe, a także jeśli wykorzysta się możliwości korygowania rozpoznania wynikające z uwzględnienia szerszego kontekstu, wówczas użyteczność mikrofonemów wydaje się prawdopodobna, a ich zalety — głównie w postaci możliwości eliminowania kłopotliwego procesu segmentacji sygnału przed rozpoznawaniem oraz szansy uproszczenia procesu rozpoznawania — mogą skłaniać do prób praktycznego wykorzystania przytoczonych wyników.

W celu pełniejszej orientacji w tablicy 3 przytoczono pełną listę wyłonionych w trakcie badań mikrofonemów wraz z podaniem nazw fonemów, którym te mikrofonemy odpowiadają. Dla uniknięcia sporów terminologicznych mikrofonemy wyłącznie numerowano, nie nadając im nazw. Mikrofonemy o numerach wyższych od 52 odpowiadają transjentom, co zaznaczono podając w objaśnieniu kolejnych wierszy tabeli pary fonemów połączonych łącznikiem „—". Zwraca uwagę, że transjenty tworzą wyłącznie samogłoski, przy czym transjent może być charakterystyczny dla spółgłoski poprzedzającej samogłoskę lub następującej po niej.

Dzięki wprowadzeniu koncepcji mikrofonemu można uniknąć kłopotów związanych z segmentacją sygnału mowy. Wydatnie wzrasta przy tym liczba podlegających rozpoznawaniu klas (zamiast 40 fonemów blisko 70 mikrofonemów), co jest zjawiskiem niekorzystnym. W dodatku rozpoznanie nie ma charakteru ostatecznego, gdyż na podstawie sekwencji mikrofonemów trzeba dopiero „odgadywać" sekwencję fonemów — co zajmuje czas i komplikuje algorytmy rozpoznawania. Rozwiązania opierające się na koncepcji mikrofonemów nie są więc optymalnym rozwiązaniem problemu segmentacji, lecz są właściwie „trikiem" o chwilowym zastosowaniu — zanim nie zostaną opracowane naprawdę skuteczne i szybkie metody segmentacji. Zresztą — być może segmentacja okaże się w przyszłości zbyteczna. Istnieją poglądy — i jest w nich zapewne wiele prawdy — że ludzie percepują mowę bez dokonywania segmentacji, umiejętność zaś rozłożenia wyrazu na elementy składowe (na przykład głoski) jest czymś wtórnym w stosunku do rozpoznawania. Łatwo się o tym przekonać słuchając nieznanego tekstu w obcym języku. Niemożliwe okazuje się nie tylko zrozumienie wypowiadanych wyrazów, ale także ich zapisanie (fonetyczne rzecz prosta). Dopiero wielokrotne wysłuchanie wyrazu pozwala na jego analizę i wyróżnienie występujących w nim fonemów. I dotyczy to człowieka, którego możliwości w zakresie rozpoznawania mowy — co wielokrotnie podkreślano — wielokrotnie przewyższają możliwości najdoskonalszych komputerów. Zatem może segmentacja jest w istocie zbędna? Niestety, na razie w systemach automatycznego rozpoznawania nie można się bez niej obejść. Jednak w miarę rozwoju elektroniki, w miarę udostępniania coraz większych i tańszych

Tablica 3. Mikrofonemy i odpowiadające im głoski, grupy głosek lub transjenty (dynamiczne stany przejściowe pomiędzy głoskami)

| Numer mikrofonemu | Odpowiadające mu fonemy lub transjenty | Numer mikrofonemu | Odpowiadające mu fonemy lub transjenty |
|---|---|---|---|
| 1 | (i) | 35 | (f), (s), (t͡s), (d͡z) |
| 2 | (ɨ) | 36 | (f), (v) |
| 3 | (e) | 37 | (f), (ʒ) |
| 4 | (e) | 38 | (v) |
| 5 | (e) | 39 | (v) |
| 6 | (e) | 40 | (v) |
| 7 | (a) | 41 | (s), (x) |
| 8 | (a) | 42 | (z) |
| 9 | (a) | 43 | (z) |
| 10 | (a) | 44 | (z) |
| 11 | (a) | 45 | (ʃ), (ʒ), (ɕ), (t͡ʃ), (d͡ʒ), (t͡ɕ) (d͡ʑ) |
| 12 | (a) | 46 | (ʒ) |
| 13 | (o) | 47 | (ʒ) |
| 14 | (o) | 48 | (ʑ) |
| 15 | (o) | 49 | (dʑ) |
| 16 | (o) | 50 | (p), (b), (t), (d), (c), (ɟ), (k), |
| 17 | (o) | | (g) |
| 18 | (u) | 51 | (k) |
| 19 | (u) | 52 | (k) |
| 20 | (u) | 53 | (a) — (k) (transjent) |
| 21 | (j) | 54 | (o) — (n) (transjent) |
| 22 | (w) | 55 | (o) — (n) (transjent) |
| 23 | (w) | 56 | (o) — (n) (transjent) |
| 24 | (r) | 57 | (o) — (m) (transjent) |
| 25 | (r) | 58 | (o) — (l) (transjent) |
| 26 | (l), (m) | 59 | (m) — (o) (transjent) |
| 27 | (l), (n) | 60 | (a) — (s) (transjent) |
| 28 | (m), (ɲ) | 61 | (k) — (a) (transjent) |
| 29 | (m), (ɲ) | 62 | (k) — (a) (transjent) |
| 30 | (m) | 63 | (k) — (a) (transjent) |
| 31 | (n), (ŋ) | 64 | (s) — (a) (transjent) |
| 32 | (n) | 65 | (t͡s) — (e) (transjent) |
| 33 | (m), (n), (ɲ), (ŋ) | 66 | (ʑ) — (e) (transjent) |
| 34 | (ɲ) | 67 | (i) — (l) (transjent) |

pamięci, potężniejszych mocy obliczeniowych, procesorów macierzowych — kto wie? Może już wkrótce argumenty przemawiające za koniecznością segmentacji, przytoczone na początku tego podrozdziału, będą miały znaczenie — jedynie historyczne?

## 5.6. Rozpoznawanie elementów mowy

Mając wydzielone podlegające rozpoznawaniu elementy, a także mając zdefiniowane parametry, za których pomocą zamierzamy te elementy rozpoznawać, możemy teraz zastanowić się nad wyborem metody rozpozna-

wania, optymalnej dla zadania rozpoznawania mowy. Zadanie, które przed nami stoi, można na płaszczyźnie formalnej rozważać w następującej postaci.

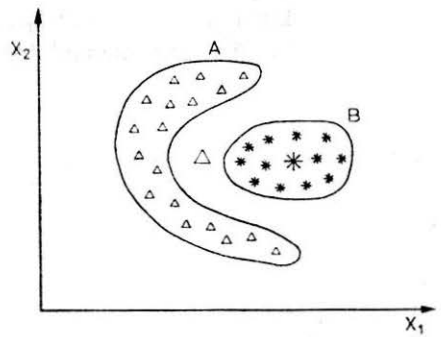
Mamy $L$ klas obiektów, z których każdy charakteryzowany jest przez zespół 5 cech, tworzących wektor $X = (x_1, x_2, x_3, x_4, x_5)$ i może być reprezentowany jako punkt w pięciowymiarowej przestrzeni. Reguły rozkładu obiektów różnych klas w przestrzeni cech nie są znane, jedyne informacje na ten temat, jakimi badacz może się posłużyć, pochodzą z tak zwanego zbioru uczącego. Jest to zbiór obiektów (próbek sygnału mowy), których przynależność do wybranych klas jest ustalona a priori. Zadanie polega na ustaleniu na podstawie ciągu uczącego — reguły rozpoznawania klasy, do której należą nowe, nieznane obiekty, a następnie na efektywnym rozpoznawaniu tych nieznanych obiektów. Możliwe jest kilka podejść do rozwiązania tak sformułowanego zadania, przy czym ich szczegółowa dyskusja wykracza poza ramy książki i ponownie, jak za każdym podobnym razem, odeślemy Czytelnika do wyliczonych na końcu pozycji literatury. Podane zostaną teraz zasadnicze wyniki i podstawowe wnioski, mające zastosowanie w zadaniu rozpoznawania mowy — gdyż większość publikacji dotyczących rozpoznawania odwołuje się do problemów rozpoznawania obrazów.

Najprostsza idea przy rozpoznawaniu polega na mierzeniu odległości między nieznanym obiektem a wszystkimi obiektami, których przynależność (poprawna identyfikacja) jest znana. Oznacza to w praktyce konieczność zapamiętania w urządzeniu rozpoznającym wszystkich obiektów ciągu uczącego, czyli wszystkich próbek elementów sygnału mowy, których rozpoznanie jest znane. Zwykle jest to bardzo dużo wzorców i jest to jedna z głównych wad rozważanej metody. Zaletą jej jest natomiast prostota i skuteczność działania. Decyzję podejmuje się na tej zasadzie, że nieznany obiekt zostaje zaliczony do tej samej klasy, do której zaliczony był jego najbliższy sąsiad, czyli obiekt ciągu uczącego, którego odległość od rozpoznawanego obiektu była w przestrzeni cech minimalna. Podana zasada jest prosta w realizacji i efektywna, a także intuicyjnie zrozumiała. Ma ona także wiele wspólnego z chętnie stosowanym przez ludzi rozumowaniem przez analogię. Jeśli tylko obiekty ciągu uczącego dobrano właściwie, to znaczy jeśli są one reprezentatywne (obejmują wszystkie ważne możliwe warianty rozważanych próbek sygnału), a także jeśli nie zawierają błędów, czyli obiektów o błędnie podanej lokalizacji lub omyłkowo przypisanej przynależności, wówczas opisana metoda — nazywana w literaturze algorytmem NN — jest bardzo skuteczna. Omyłki lub braki rozpoznania zdarzają się przy jej użyciu jedynie sporadycznie, a wiarogodność rozpoznawania jest najwyższa z możliwych. Niestety, omówione zalety metody okupione są zasadniczymi wadami: wspomnianą wcześniej dużą zajętością pamięci (wynikającą z konieczności pamiętania wszystkich obiektów ciągu uczącego) oraz małą szybkością działania, wynikającą z tej samej przyczyny (policzenie odległości od nieznanego obiektu do wszystkich znanych musi trwać, nawet przy szybkim komputerze, dość długo, a operacja wyszukiwania odległości najmniejszej też do najszybszych nie należy). Wadę tę można starać się częściowo usunąć stosując

ograniczoną reprezentację wszystkich klas w postaci sztucznie tworzonych wzorców. Na przykład jeśli kształt obszaru wyznaczonego w przestrzeni cech przez obiekty pewnej interesującej klasy zbliżony jest do kuli, wówczas te same reguły decyzyjne otrzymuje się stosując rozpoznawanie na podstawie mierzenia odległości nieznanego obiektu od wszystkich elementów ciągu uczącego jak w przypadku, kiedy całą klasę reprezentuje pojedynczy obiekt zlokalizowany w środku kuli. Wyznaczenie takiego reprezentatywnego obiektu (centroidu), jest w rozważanym przypadku łatwe. Współrzędne centralnego punktu mogą być wyznaczone jako wartości średnie wyliczone ze współrzędnych wszystkich punktów wchodzących w skład ciągu uczącego dla rozważanej klasy.

Opisane postępowanie może być w określonych przypadkach uzasadnione. Z jednej strony w wyniku wyliczenia wartości średniej ze współrzędnych punktów ciągu uczącego otrzymuje się — rozważając to z geometrycznego punktu widzenia — współrzędne punktu odpowiadającego „środkowi ciężkości" zbiorowości punktów danej klasy. Na ogół przy regularnym rozkładzie punktów ich środek ciężkości istotnie najlepiej może je wszystkie zastępować. Z drugiej strony proces uśredniania jest rutynowo stosowany do eliminacji skutków przypadkowych zakłóceń, zatem przyjmując wyobrażeniowo pewien model procesu artykulacji mowy możemy twierdzić, że opisany sposób postępowania jest (przy przyjęciu podanych niżej założeń) zbliżony do optymalnego. Model procesu artykulacji, do którego odwoływano się w ostatnim zdaniu, sprowadza się do przypuszczenia, że istnieją pewne wzorce idealnej artykulacji określonych elementów sygnału mowy (na przykład poszczególnych fonemów). Każda rzeczywista realizacja procesu artykulacji, a więc każda próbka zarejestrowanego sygnału mowy, stanowi niedoskonałą, zakłóconą próbę reprodukcji tego wzorca. Przyjmując (co jest zresztą bardzo wątpliwe), że zniekształcenia wzorca można rozważać jako dodawanie do sygnału o idealnych parametrach składowej losowej, której parametry mają zerową wartość oczekiwaną i ograniczoną wariancję, otrzymujemy model artykulacji, dla którego idealnym (optymalnym) sposobem odtworzenia zakłóconego wzorca wypowiedzi jest — właśnie uśrednianie. Oczywiście proces uśredniania nie zawsze musi dostarczać „wzorca" dla klasy, który będzie poprawnie ją reprezentował w procesie rozpoznawania. Jeśli kształt obszaru, odpowiadającego rozważanej klasie elementów sygnału mowy, odbiega w przestrzeni parametrów od kuli (por. rys. 5-13), to wówczas położenie punktu środka ciężkości może być zupełnie przypadkowe. Punkt o współrzędnych pochodzących z uśredniania nie może w takim przypadku pełnić roli „wzorca" klasy, co jednak na ogół nie dyskwalifikuje samej koncepcji zastąpienia wszystkich punktów ciągu uczącego przy rozpoznawaniu — ich skróconą reprezentacją. Na ogół jednak w przypadkach, kiedy geometria obszarów w przestrzeni cech odbiega od kształtów zbliżonych do kuli, zachodzi potrzeba znalezienia dla każdej klasy kilku (a nie pojedynczego) reprezentatywnych wzorców. Istnieją i są opisane w literaturze metody określania liczby niezbędnych wzorców, a także sposoby ich lokalizacji i wydzielania.

Opisane metody rozpoznawania: algorytm najbliższego sąsiada NN, w którym rozważane są odległości nieznanego obiektu od wszystkich obiektów o znanej przynależności, i algorytm wykorzystujący pojęcie wyliczanego „wzorca" (lub „wzorców") dla każdej klasy — opierają się na pojęciu odległości w przestrzeni cech. Pojęcie to było zresztą używane także wcześniej (por. rozdz. 5.4) i będzie przydatne także w dalszych rozważaniach. Warto więc teraz, przed skrótowym przedyskutowaniem przynajmniej niektórych metod rozpoznawania, nie wymagających odwoływania się do pojęcia odległości, poświęcić nieco uwagi zagadnieniu sposobu definiowania odległości w przestrzeni cech. Warto bowiem mieć świadomość związków, jakie istnieją między wyborem odpowiedniej metryki przestrzeni cech a własnościami omówionych wcześniej algorytmów rozpoznawania. Wszak nawet kształt kuli w przestrzeni cech jest zależny od przyjętej definicji odległości, a przy niektórych definicjach kształt ten może daleko odbiegać od znanego nam. Zatem dyskusja dotycząca przedmiotu przydatności lub nieprzydatności metody uśredniania do wyznaczania wzorców musiała być prowadzona z licznymi — dla prostoty pomijanymi wyżej — zastrzeżeniami na temat używanego pojęcia odległości. Również efektywność algorytmów rozpoznawania może w zasadniczy sposób wiązać się z użytym pojęciem metryki. Nie jest bowiem obojętne, nawet przy stosowaniu największych i najszybszych komputerów, jakie działania arytmetyczne trzeba wykonać, żeby wyliczyć wartość odległości. Minimalne nawet oszczędności w definicji pojęcia odległości mogą dawać znaczące oszczędności w czasie obliczeń, skoro odległość musi być obliczana dla tysięcy punktów ciągu uczącego przy każdej próbie rozpoznawania. W tej sytuacji mikrosekundy oszczędności na pojedynczym obiekcie mogą oznaczać całe minuty obliczeń dla całego zbioru danych i dla pełnego rozpoznania interesującej wypowiedzi.




5-13. Element będący środkiem ciężkości obiektów ciągu uczącego może być używany jako wzorzec całej klasy jedynie dla obiektów, których rozkład w przestrzeni kształtów ma prosty i stosunkowo regularny kształt (na przykład gwiazdki). Natomiast dla obiektów klasy A, oznaczonych trójkątami, środek ciężkości leży poza obszarem danej klasy (większy trójkąt) i nie może pełnić roli wzorca.

Przystępując do dyskusji pojęcia odległości w przestrzeni cech możemy na wstępie przedyskutować najczęściej stosowaną i z pozoru „oczywistą" miarę odległości Euklidesa. Przypomnijmy, że dla dwu obiektów $X^i = (x_1^i, x_2^i, ..., x_5^i)$ oraz $X^j = (x_1^j, x_2^j, ..., x_5^j)$ ich odległość, oznaczana już

wcześniej $d_{ij}$ (por. p. 5.4), wyliczana jest przy zastosowaniu metryki euklidesowej ze wzoru

$$d_{ij} = \sqrt{\sum_{k=1}^{5} (x_k^i - x_k^j)^2} \tag{5.6}$$

Już pobieżna analiza przytoczonej formuły pozwala zauważyć jej wady. Po pierwsze operacja podnoszenia do kwadratu i następującego po niej wyciągania pierwiastka jest czasochłonna, czyli niedogodna obliczeniowo. Aby to zlikwidować, używa się niekiedy innej metryki, nazywanej „uliczną" (bo odpowiada ona odległości, jaką musi przebyć turysta w mieście o wytyczonych równolegle i prostopadle ulicach, uniemożliwiających chodzenie na skróty od punktu $X^i$ do punktu $X^j$). Metryka ta wiązana jest niekiedy z nazwiskiem Hamminga ze względu na formalne podobieństwo z wprowadzoną przez tego badacza miarą odległości ciągów kodowych, służącą do badania nadmiarowości kodów i stopnia ich zabezpieczenia przed przekłamaniami. W metryce tej odległość punktów $X^i$ oraz $X^j$ wyraża się wzorem

$$d_{ij} = \sum_{k=1}^{5} |x_k^i - x_k^j| \tag{5.7}$$

Widać, że uciążliwe obliczeniowo operacje zostały w tej metryce niemal w całości wyeliminowane i zastąpione prostymi i szybko wykonywanymi działaniami. Warto przy tym zauważyć, że przyjęcie metryki danej wzorem (5.7) powoduje, że kulę w przestrzeni cech zastępuje sześcian. Może on niekiedy lepiej pasować do kształtów obszarów w przestrzeni cech, co może być (ale nie musi) dodatkową zaletą metryki (5.7).

Obie przytoczone definicje odległości mają wspólną wadę. Jeśli zakres zmienności któregoś z rozważanych parametrów okaże się większy niż dla innego parametru, to odpowiednie składniki w sumie będą dominowały nad pozostałymi, co szczególnie dotkliwie może dać się zauważyć we wzorze (5.6), ze względu na operację podnoszenia do kwadratu. Zwróćmy uwagę, że na przykład zakres zmienności trzeciego formantu jest znacznie większy niż odpowiedni zakres dla pierwszego formantu. Czyli miara odległości silniej będzie zależała od różnic w wartościach trzeciego formantu niż od różnic w wartościach formantu pierwszego. Analogicznie zakresy zmian momentów widmowych są większe od zakresu zmienności formantów — wartości momentów będą silniej wpływały na rozważane odległości niż formanty.

Wszystkie omówione zróżnicowania są odwrotne w stosunku do tendencji, jakie powinny mieć miejsce z punktu widzenia rzeczywistej ważności odpowiednich parametrów. Wszak formanty są generalnie bardziej wartościowe, z punktu widzenia identyfikacji elementów sygnału mowy, niż momenty widmowe, a wśród formantów pierwszy wnosi więcej informacji i jest ważniejszy niż trzeci. Wynika z tego, że celowe jest wprowadzenie poprawki do formuły (5.6) i (5.7), polegającej na wprowadzeniu współczynników „wagowych", zróżnicowujących w pożądany sposób udział po-

szczególnych składników w odpowiednich sumach. Dokonując poprawek przechodzimy od metryk zwykłych do tzw. uogólnionych, których formuły są następujące:

$$d_{ij} = \sqrt{\sum_{k=1}^{5} \lambda_k^2 (x_k^i - x_k^j)^2} \tag{5.8}$$

względnie

$$d_{ij} = \sum_{k=1}^{5} \lambda_k |x_k^i - x_k^j| \tag{5.9}$$

Współczynniki wagowe $\lambda_k$ mogą być dobrane tak, aby równe były odwrotnościom zakresów zmienności poszczególnych cech (co odpowiada w istocie wstępnej normalizacji zakresów zmienności wszystkich cech do stałego przedziału [0,1]), względnie można w wartościach współczynników $\lambda_k$ zawrzeć dodatkowo informację o względnej wartości czy też ważności poszczególnych cech. Warto przy tym zauważyć, że zastosowanie wzorów (5.8) i (5.9) do opisu pojęcia odległości w przestrzeni cech prowadzi do dalszej deformacji „kul" w tej przestrzeni. Dla metryki (5.8) przybierają one formę hiperelipsoid o osiach równoległych do osi układu współrzędnych, a dla wzoru (5.9) mamy do czynienia z hiperrównoległościanami.

Z punktu widzenia prostoty obliczeń i dobrych wyników „samonormalizacji" poszczególnych cech dobre własności ma metryka Camberra:

$$d_{ij} = \sum_{k=1}^{5} \frac{|x_k^i - x_k^j|}{|x_k^i + x_k^j|} \tag{5.10}$$

której charakterystyczne, ruchome kule łatwo na ogół dopasowują się do rzeczywistych kształtów obszarów poszczególnych mikrofonemów w przyjętej przestrzeni cech. Badania nad rozpoznawaniem ograniczonego zbioru elementów mowy potwierdziły przydatność metryki (5.10) do opisu obszarów w tej przestrzeni.

Wszystkim rozważanym metrykom można postawić dodatkowo jeden zarzut. Otóż metryka euklidesowa opiera się de facto na twierdzeniu Pitagorasa, przeto może być wykorzystywana do układów współrzędnych, których osie są prostopadłe. Tymczasem układ współrzędnych zaproponowanej przestrzeni pięciowymiarowej nie jest prostokątny. Na pozór wydaje się, że jest to kwestia zupełnie dowolnej decyzji, czy zbuduje się układ współrzędnych prostokątny, czy dowolny inny. Tymczasem sprawa nie jest taka prosta. Ortogonalny (prostokątny) układ współrzędnych może służyć do odkładania wartości całkowicie niezależnych parametrów. Natomiast rzeczywiste parametry opisujące sygnał mowy są ze sobą skorelowane. Jest to logiczne. Na przykład, formanty nie mogą być niezależne, skoro wszystkie razem są kształtowane w tym samym procesie artykulacji przez ruchy tych samych narządów mowy i przez współdziałanie jednego zespołu wnęk rezonansowych. Analogiczne rozważania można przeprowadzić dla momentów widmowych, wykazując ich wzajemne powiązania oraz związki (słabsze na

ogół) pomiędzy grupą parametrów formantowych i momentami. Innymi słowy w sensie statystycznym wszystkie parametry opisujące sygnał mowy są ze sobą powiązane. Skoro tak, to po wybraniu dowolnego spośród zespołu pięciu wytypowanych parametrów, można próbować oszacowywać jego wartości z równania regresji, będącego liniową kombinacją pozostałych parametrów. Jeśli jednak wybrany parametr może być (w przybliżeniu, co prawda) przedstawiony jako liniowa kombinacja pozostałych, to geometrycznie fakt ten interpretując leży on na płaszczyźnie (dokładniej — w podprzestrzeni liniowej) wyznaczonej przez te parametry. A to oznacza, że nie powinniśmy wskazanego parametru odkładać na osi prostopadłej, do pozostałych parametrów, wszak leży on nieomal w ich płaszczyźnie! Powtarzając podobne rozumowanie dla wszystkich parametrów upewniamy się, że układ współrzędnych naszej przestrzeni jest nieortogonalny, skośnokątny. Nie wolno więc stosować metryk opartych na twierdzeniu Pitagorasa.
Właściwym rozwiązaniem jest stosowanie metryki Mahalanobisa, w której wykorzystuje się macierz kowariancji cech $S$ dla korekty efektów skośności układu współrzędnych. Odpowiedni wzór jest następujący:

$$d_{ij} = (X^i - X^j)^T S^{-1} (X^i - X^j) \tag{5.11}$$

gdzie dla uproszczenia zapisu wykorzystano notację wektorowo-macierzową: $S^{-1}$ oznacza macierz odwrotną do macierzy kowariancji cech $S$, odejmowanie należy traktować jako odejmowania wektorowe, $T$ oznacza zaś transpozycję (macierzy lub wektora). Można łatwo wykazać, że forma kwadratowa dana wzorem (5.11) jest zawsze dodatnia z wyjątkiem przypadku $X^i = X^j$, kiedy przyjmuje wartość zero. Może zatem być użyta jako miara odległości w przestrzeni cech. Trudniej udowodnić inne prawdziwe własności metryki (5.11). Między innymi pożądaną z punktu widzenia zadań rozpoznawania mowy własność dekorelacji cech. W uproszczeniu można więc przyjąć, że odległość liczona ze wzoru (5.11) odpowiada odległości euklidesowej („uogólnionej" zgodnie ze wzorem (5.9)) w zdekorelowanej, „wyprostowanej" przestrzeni cech. W dodatku „kule", zadane miarą odległości (5.11) mają szczególnie korzystny kształt: są to hiperelipsoidy, których rozmiary i kierunek przestrzennej orientacji osi są zgodne z kształtami, jakie w przestrzeni cech przyjmują obszary odpowiadające poszczególnym klasom. Te korzystne własności okupione są jednak dużą złożonością obliczeniową wzoru (5.11). Nakład pracy związany z obliczeniami według reguły (5.11) jest bez porównania większy niż przy uprzednio wprowadzanych metrykach, a w dodatku operacja odwracania macierzy kowariancji $S$ bywa uciążliwa obliczeniowo ze względu na bliski zera wyznacznik główny.
W sumie więc większość badaczy docenia zalety metryki Mahalanobisa „teoretycznie": chwali ją w publikacjach i nie stosuje w praktyce.
W uzupełnieniu dyskusji grupy metod rozpoznawania, w których podstawą procesu rozpoznawania jest wyliczanie odległości od nieznanego, rozpoznawanego obiektu do wszystkich, lub tylko niektórych, wybranych obiektów ciągu uczącego, warto odnotować jeszcze jedną zaletę zaproponowanych w poprzednim podrozdziale mikrofonemów. Otóż w niektórych metodach

grupowania, wykorzystywanych do tworzenia skupień będących wzorcami mikrofonemów, obok innych możliwości wyliczane są także „centroidy" wszystkich rozważanych klas, co zasadniczo ułatwia stosowanie algorytmów rozpoznawania, odwołujących się do tych uogólnionych wzorców klas. W ten sposób, wykorzystując mikrofonemy, możemy swobodnie wykorzystywać zalety metod opartych na podejściu „minimalnoodległościowym" bez konieczności akceptowania ich największej wady, wiążącej się z dużym zajęciem pamięci oraz długim czasem liczenia.

Wady metod minimalnoodległościowych mogą być ominięte także i na innej zasadzie. Obszerna grupa dyskutowanych w literaturze metod rozpoznawania odwołuje się do pojęcia tak zwanych funkcji przynależności. Postuluje się w tych metodach — stosując rozmaite nazewnictwo, różną notację i zróżnicowane algorytmy obliczeniowe — właściwie stale podobną koncepcję. Oto najważniejsze, typowo występujące jej elementy. Niech istnieje $L$ funkcji argumentu wektorowego $X$

$$g_i(X) = \sum_k \varphi_k(X) \cdot w_k(i); \quad i = 1, 2, \ldots, L \tag{5.12}$$

gdzie funkcje $\varphi_k$ stanowią odpowiednio dobraną rodzinę funkcji wektorowych (o jej doborze będzie jeszcze dalej mowa), współczynniki zaś wagowe $w_k(i)$ różnicują poszczególne funkcje $g_i(X)$ i są ustalane na podstawie ciągu uczącego w procesie iteracyjnym, nazywanym zwykle uczeniem. Podstawowe założenie i zasadniczy wymóg, jaki można postawić funkcjom $g_i(X)$, jest następujący. Jeśli wektor $X$ jest wektorem parametrów (cech) obiektu należącego do pewnej ustalonej klasy $n$, to wówczas funkcja $g_n(X)$ przyjmuje w punkcie $X$ wartości większe niż wszystkie inne funkcje $g_i(X)$ dla $i \neq n$. Wynika z tego prosta reguła rozpoznawania, oparta na wartościach funkcji $g_i(X)$. Jeśli mamy rozpoznać nieznany fragment sygnału mowy, to po określeniu dla niego wszystkich parametrów i skompletowaniu wektora $X$ oblicza się wartości wszystkich funkcji $g_1(X)$, $g_2(X)$, ..., $g_L(X)$, gdzie $L$ jest liczbą rozpoznawanych klas. Jedna z tych wartości będzie większa od pozostałych, załóżmy, że jest to wartość dla numeru $n$:

$$g_n(X) \geqslant g_i(X) \quad i = 1, 2, \ldots, L \tag{5.13}$$

W takim przypadku poprawne rozpoznanie odpowiadać będzie klasie $n$, co oznacza, że przy podejmowaniu decyzji wystarczy kontrolować, która z funkcji $g_i(X)$ przyjmuje wartość największą, a rozpoznanie utożsamiać z numerem klasy, której funkcja jest większa od pozostałych. Ze względu na swoje własności funkcje $g_i(X)$ są nazywane funkcjami przynależności, ponieważ ich wartości określają stopień przynależności nieznanego obiektu $X$ do odpowiednich klas $i$. Naturalne jest przy tym rozpoznanie tej klasy, dla której wspomniany stopień przynależności okaże się największy.

Przytoczony sposób sformułowania zadania rozpoznawania przesuwa środek ciężkości problemu ze sfery podejmowania decyzji do sfery obliczeń arytmetycznych, co jest korzystne z punktu widzenia realizacji tego zadania z wykorzystaniem elektronicznej maszyny cyfrowej. Równocześnie jednak pojawia się problem zbudowania funkcji przynależności $g_i(X)$ w ten spo-

sób, aby mogły spełniać swoje zadanie zgodnie ze sformułowanymi wyżej postulatami. Kłopot polega przy tym na braku jakiejkolwiek konkretnej informac na temat kształtu i przebiegu poszukiwanych funkcji $g_i(X)$, gdyż jedyna informacja, jaką dysponuje badacz, zawarta jest w ciągu uczącym, to znaczy w zbiorze wybranych obiektów, dla których znane są zarówno charakteryzujące je parametry (składowe wektora $X$), jak i poprawna przynależność (numer klasy, do której obiekty te należą). Zwróćmy uwagę, że mimo pewnych podobieństw zadanie tu sformułowane jest odmienne od zadania aproksymacji funkcji. W zadaniach aproksymacji funkcji znane muszą być w wybranych punktach zarówno argumenty (w naszym przypadku wektor $X$), jak i wartości funkcji. Wówczas można zastosować liczne znane i bardzo efektywne obliczeniowo metody aproksymacji — na przykład metodę najmniejszych kwadratów, wzmiankowaną wyżej przy metodach liniowej predykcji (por. p. 4.5). Niestety jednak w zadaniu rozpoznawania w ustalonych punktach znane są jedynie relacje między wartościami funkcji przynależności, a nie same wartości. Istotnie, jeśli wiemy, że dla obiektu opisanego wektorem $X$ właściwe rozpoznanie odpowiada klasie $n$, wówczas jedyny wniosek, jaki z faktu tego można wyciągnąć, ma postać zbioru nierówności typu (5.13), określających stosunki między wartościami funkcji przynależności — ale nie same wartości.

Metoda postępowania, która gwarantuje rozwiązanie postawionego zadania, jest następująca. W pierwszym kroku zakłada się, że funkcje przynależności można przedstawić w postaci rozwinięcia na szereg, z ustalonym zbiorem funkcji bazowych $\varphi_k(X)$. Funkcje $\varphi_k$ są jednakowe we wszystkich funkcjach przynależności dla wszystkich klas $i$, natomiast współczynniki rozwinięcia $w_k$ są zależne od tego, do której klasy ma być stwierdzona przynależność przy wykorzystaniu danej funkcji, co we wzorze (5.12) odnotowano formalnie zapisując je jako $w_k(i)$.

Takie postawienie zadania stwarza podwójnie dogodną sytuację. Po pierwsze, proces uczenia polegający na formowaniu funkcji przynależności dla poszczególnych klas sprowadza się dzięki takiemu postawieniu sprawy do określenia zbioru wartości współczynników $w_k(i)$ dla każdej klasy — a to jest zadanie prostsze. Po drugie, zadanie zgromadzenia w pamięci komputera wyników procesu nauczania jest tanie, przy podanym sposobie budowy funkcji przynależności. W rezultacie procesu uczenia określone zostają wartości wszystkich współczynników $w_k(i)$ dla wszystkich $k$ oraz dla wszystkich $i$. Na ogół objętość pamięci potrzebna do zapamiętania wartości tych współczynników jest niewielka, o wiele mniejsza od objętości wymaganej przy metodach minimalnoodległościowych, a ponadto objętość ta jest stała, niezależnie od liczby obiektów ciągu uczącego, branych pod uwagę w trakcie procesu uczenia.

Podstawową sprawą do dalszych rozważań jest zakres sumowania (przedział wartości, w obrębie którego zmienia się $k$) w rozwinięciu danym wzorem (5.12). Z jednej strony korzystne jest, aby zakres ten był możliwie mały. Dzięki temu zmniejsza się objętość pamięci koniecznej do zapamiętania współczynników $w_k(i)$, a także upraszczają się i przyspieszają obliczenia.

Z drugiej jednak korzystne jest, aby ten zakres był duży, istnieją bowiem teoretyczne przesłanki do tego, aby sądzić, że prawdopodobieństwo poprawnego rozpoznawania będzie rosło ze wzrostem zakresu sumowania we wzorze (5.12). Uzasadnienie tej tezy jest dość złożone, jeśli wymagany jest dowód formalny. Intuicyjnie jednak jest to dość oczywiste. Jeżeli wzór (5.12) będziemy traktować jako przybliżenie za pomocą rozwinięcia na szereg nieznanej funkcji $g_i(X)$, to oczywiste jest, że przybliżenie to jest tym dokładniejsze, im więcej członów rozwinięcia branych jest pod uwagę, zatem jakość wypełniania przez funkcję $g_i(X)$ roli funkcji przynależności zależna jest od liczby składników sumy. Można przypuszczać, biorąc pod uwagę teorię rozwinięć funkcyjnych, że najlepsze wyniki osiągnie się przy nieskończonym przedziale zmienności $k$, co jest jednak w sposób oczywisty nie do przyjęcia w praktyce. Sprzeczność, występującą między interesem sprawności obliczeniowej a wymogami dokładności rozpoznawania, można częściowo godzić dobierając odpowiednio rodzinę funkcji $\varphi_k(X)$. Gdyby na przykład przez odpowiedni dobór funkcji $\varphi_k(X)$ można było zapewnić warunek $w_k(i) = 0$ dla wszystkich $i$ oraz dla wszystkich $k > m$, gdzie $m$ jest ustaloną, dostatecznie małą liczbą, to wówczas pogodzenie warunków dokładności i efektywności byłoby proste. Jak jednak tego dokonać, skoro o podlegających aproksymacji funkcjach $g_i(X)$ prawie nic nie wiadomo? Istotnie, problem wyboru funkcji bazowych $\varphi_k(X)$ do łatwych nie należy, szczególnie dlatego, że muszą to być funkcje argumentu wektorowego. Znane z poradników i podręczników rodziny funkcji używanych w rozwinięciach na szeregi są zwykle opracowywane dla argumentów skalarnych (funkcje trygonometryczne, wielomiany Czebyszewa itp.).

Pozostawiając dyskusję konkretnych metod doboru własności funkcji bazowych $\varphi_k(X)$ dla konkretnych zadań rozpoznawania do ewentualnych samodzielnych studiów Czytelników (odpowiednie pozycje literatury cytowane są na końcu książki) przedstawimy teraz kilka interesujących z teoretycznego punktu widzenia i przydatnych praktycznie rodzin funkcji $\varphi_k(X)$. Niewątpliwie najczęściej dyskutowany w literaturze, najbardziej przydatny z dydaktycznego punktu widzenia i wysoce użyteczny praktycznie jest przypadek funkcji liniowej. Poszczególne funkcje $\varphi_k(X)$ są w tym przypadku równe kolejnym składowym wektora $X$, całe równanie (5.12) zaś przyjmuje uproszczoną postać:

$$g_i(X) = \sum_{k=1}^{5} w_k(i) x_k + w_0(i) \tag{5.14}$$

Przyjęcie funkcji o tej postaci oznacza, że obszary odpowiadające poszczególnym klasom w przestrzeni cech rozgraniczane będą hiperpłaszczyznami. Istotnie, rozważając dowolne dwie klasy: $i$ oraz $j$ możemy stwierdzić, że w pobliżu granicy rozdzielającej ich obszary w przestrzeni cech podejmuje się decyzje o rozpoznaniu obiektu klasy $i$, gdy $g_i(X) > g_j(X)$, decyzję zaś o przynależności do klasy $j$ podejmuje się przy $g_j(X) > g_i(X)$. Równanie powierzchni rozgraniczającej te dwa obszary ma więc postać:

$$g_i(X) - g_j(X) = 0 \tag{5.15}$$

a po uwzględnieniu postaci funkcji przynależności danej wzorem (5.14) równanie powierzchni granicznej staje się równaniem hiperpłaszczyzny

$$\sum_{k=1}^{5} [w_k(i) - w_k(j)] x_k + [w_0(i) - w_0(j)] = 0 \tag{5.16}$$

Oznacza to, że przydatność rozważanej metody ograniczona jest do przypadku, kiedy topografia obszarów w przestrzeni cech pozwala na ich rozgraniczanie hiperpłaszczyznami. Położenie tych hiperpłaszczyzn zależne jest od wartości współczynników $w_k(i)$ oraz $w_k(j)$. Zatem opisany iteracyjny proces „uczenia" zmierzający do ustalenia optymalnych wartości współczynników wagowych $w_k$ dla poszczególnych klas traktowany być może jako przesuwanie i obracanie hiperpłaszczyzn granicznych w ten sposób, aby umieścić je dokładnie w lukach między obszarami poszczególnych klas. Jeśli to jest tylko możliwe (jeśli pomiędzy obszarami klas da się zmieścić hiperpłaszczyznę), proces uczenia doprowadzi do takiego ustawienia granicznej powierzchni, aby separacja dokonywana była w sposób doskonały, przynajmniej dla dostępnych danych z ciągu uczącego. Jeśli graniczna powierzchnia ma kształt bardziej złożony i nie może być sensownie przybliżona hiperpłaszczyzną, to proces uczenia prowadzi do takiego ustawienia granicy, aby błędy wynikające z niedopasowania kształtu brzegów obszarów i rozgradzającej je hiperpłaszczyzny były minimalne.

Jak wynika z przytoczonych uwag, kluczową rolę dla całego rozpoznawania ma w omawianej grupie metod proces uczenia. Zasada tego uczenia jest zaskakująco prosta, a efekty — nadspodziewanie dobre. Przypomnijmy, że podstawą procesu uczenia jest ciąg uczący, to znaczy zbiór obiektów, dla których znana jest poprawna ich klasyfikacja. Zapiszmy ciąg uczący jako zbiór par:

$$U = \{\langle X^n, i^n \rangle, n = 1, 2, \ldots, N\} \tag{5.17}$$

gdzie $X^n$ jest wektorem cech $n$-tego obiektu, a $i^n$ — numerem klasy, do której obiekt ten należy. Wówczas regułę uczenia można zapisać w sposób następujący:

$$w_k(i^n) = w_k(i^n) + x_k^n \quad k = 1, \ldots, 5 \tag{5.18}$$

$$w_0(i^n) = w_0(i^n) + 1 \tag{5.19}$$

$$w_k(j^n) = w_k(j^n) - x_k \quad k = 1, \ldots, 5 \tag{5.20}$$

$$w_0(j^n) = w_0(j^n) - 1 \tag{5.21}$$

Powyższe zapisy należy traktować podobnie jak instrukcje podstawienia w językach programowania, a nie jak równania. To znaczy, że odpowiednie wartości po prawej stronie znaku równości traktować należy jako wartości przed dokonaniem korekty, wynikającej z pokazania $n$-tego obiektu ciągu uczącego, a te same wartości po lewej stronie znaku równości odpowiadają wartościom skorygowanym, po dokonaniu elementarnego kroku procesu uczenia. Formalnie wartości te należało rozróżniać, pisząc na przykład znak „prim" ($w_k'(i^n) = w_k(i^n)$ i tak dalej), zaniechano jednak tego, aby nie komplikować i tak złożonego zapisu.

Przy analizie wzorów (5.18)÷(5.21) kluczowe jest ustalenie znaczenia numeru klasy $j^n$. Opisowo można stwierdzić, że jest to ten numer klasy, który byłby wskazany jako rozpoznany, gdyby rozpoznawanie powierzono „nienauczonej" procedurze. Dokładniej można powiedzieć w sposób następujący: w momencie pokazywania $n$-tego obiektu ciągu uczącego procedura rozpoznająca ma już zapamiętane wszystkie wartości współczynników $w_k(i)$ — chociaż są to wartości niedokładne, jako że proces uczenia jeszcze trwa. Jeśli jednak obliczyć, posługując się wzorem (5.14) i tymi niedokładnymi wartościami współczynników $w_k$, wartości funkcji przynależności, to wówczas jedna z nich będzie miała największą wartość i zgodnie z regułą (5.13) byłaby podana jako rozwiązanie — gdyby to był etap roboczego rozpoznawania, a nie jeszcze uczenie. Ten właśnie numer klasy, który byłby rozpoznany przez „niedouczoną" procedurę dla obiektu pokazanego na $n$-tym kroku procesu uczenia, oznaczono $j^n$. Oczywiście $g_{j^n}(X^n) \geqslant g_i(X^n)$ dla wszystkich $i$, a ponadto na ogół $j^n \neq i^n$, gdyż niedouczona procedura popełnia błędy. Jak widać ze wzorów (5.18)÷(5.21), istota procesu uczenia polega na tym, że zwiększane są współczynniki wagowe tej klasy, która powinna była być rozpoznana (wzory (5.18) i (5.19)), zmniejszane są zaś współczynniki dla tej klasy, która została omyłkowo rozpoznana (wzory (5.20) i (5.21)). Warto zauważyć, że w przypadku, kiedy zaproponowane przez procedurę rozpoznającą „prowizoryczne" rozpoznanie jest poprawne ($j^n = i^n$), korekty dane wzorami (5.18) i (5.19) oraz (5.20) i (5.21) znoszą się wzajemnie. W rezultacie wartości współczynników $w_k(i)$ pozostają nie zmienione.

Fakt biernego reagowania procedury uczącej na poprawnie klasyfikowane obiekty ciągu uczącego bywa wykorzystywany do określania momentu zatrzymania procesu uczenia. Istotnie, jeśli obserwując proces uczenia widzimy, że kolejne obiekty nie wywołują zmian wartości współczynników $w_k(i)$ — czyli prowizoryczne klasyfikacje okazują się prawidłowe — to możemy przypuszczać, że klasyfikacja będzie poprawna także i dla innych obiektów. Przypuszczenie takie jest tym bardziej wiarygodne, im więcej bezbłędnych rozpoznań zaobserwujemy. Jednak z podjęciem decyzji o zaprzestaniu procesu uczenia nie można czekać w nieskończoność, bo zużywa się bezproduktywnie czas wykorzystywanego komputera. Konieczna jest więc kompromisowa decyzja — po ilu poprawnie sklasyfikowanych obiektach ciągu uczącego można już uznać, że procedura jest zadowalająco nauczona? W literaturze pojawiają się teoretyczne oszacowania — bądź oparte na analizie zbieżności procesu uczenia, bądź na wnioskowaniu typu statystycznego. W pierwszym przypadku wykorzystuje się wniosek z twierdzenia o silnej zbieżności (w skończonej liczbie kroków) procesu uczenia zadanego wzorami (5.18)÷(5.20). W dowodzie tego twierdzenia wyznacza się formułę określającą maksymalną liczbę możliwych pokazów, przy których mogą wystąpić błędy. W drugim podejściu można oszacować prawdopodobieństwo popełnienia omyłki przy podejmowaniu decyzji o przerwaniu uczenia. Jedno i drugie podejście jest jednak niepraktyczne, gdyż wyliczone ilości pokazów są bardzo duże, znacznie większe od rzeczywiście niezbędnych. Praktyczne podejście może być więc zaproponowane w następującej

postaci. Ciąg uczący jest ograniczony, a operacje zadawane wzorami (5.18) i (5.21) są dość proste. Celowe jest więc pokazywanie ciągu uczącego cyklicznie: po pokazaniu (i wykorzystaniu do celów nauki) ostatniego obiektu, można ponownie pokazać pierwszy obiekt — gdyż wcale nie ma gwarancji, że po korektach współczynników $w_k(i)$ danych omawianym zespołem wzorów, wszystkie pokazane obiekty będą już prawidłowo rozpoznawane, a w dodatku poprawiając rozpoznania jednych obiektów można psuć rozpoznania innych. Przy cyklicznym pokazywaniu ciągu uczącego warunek zaprzestania uczenia jest trywialny. Można przerwać uczenie, jeżeli prawidłowo sklasyfikowany jest cały jeden zbiór obiektów, a więc w rezultacie wszystkie dostępne obiekty ciągu uczącego. Oczywiście takiego rezultatu można niekiedy w ogóle nie osiągnąć — w przypadku kiedy granice między obszarami poszczególnych klas nie mogą być hiperpłaszczyznami. Wówczas proces uczenia musi być przerwany pomimo wciąż pojawiających się błędów przy klasyfikacji obiektów ciągu uczącego, przy czym użyć trzeba dodatkowego kryterium — na przykład licznika „obiegów" przy cyklicznym przeglądaniu ciągu uczącego.

Równania (5.18)÷(5.21) jako metoda uczenia, równanie (5.14) jako kryterium podejmowania decyzji oraz przytoczone wyjaśnienia dotyczące reguł stosowania wzorów (5.18)÷(5.21) do ciągu uczącego (5.17) kończą w zasadzie opis jednej ze skuteczniejszych metod rozpoznawania — metody funkcji liniowych. Warunkowy tryb ostatniego zdania wynikał z faktu, że pominięto (celowo) drobny szczegół procesu uczenia, który wart jest, aby go teraz dodatkowo omówić. Otóż reguły dane wzorami (5.18)÷(5.21) podają sposób poprawiania wartości współczynników $w_k(i)$ w trakcie procesu uczenia. Od czego jednak zacząć, to znaczy, jakie wartości ma mieć zbiór współczynników $w_k(i)$ dla wszystkich $k$ oraz wszystkich $i$ — przed pokazaniem pierwszego obiektu ciągu uczącego? Że trzeba tu arbitralnie przyjąć pewne wartości — to jest oczywiste. Wymaga tego zarówno postać równań (5.18)÷(5.21) — w których dla $n = 1$ również po prawej stronie wystąpić muszą „poprzednie" wartości $w_k(i)$, a także pragmatyka procesu uczenia, w którym muszą być podejmowane „prowizoryczne" próby rozpoznawania, dzięki którym określa się wartości numeru „błędnej" klasy $j^n$ (w rozważanym przypadku $-j^1$). Wybór wartości $w_k(i)$ dla $k = 0, 1, \ldots, 5$ oraz $i = 1, 2, \ldots, L$ może być dowolny. Odpowiednie twierdzenia w teorii uczenia maszyn rozpoznawania obrazów wskazują, że przy dowolnym wyborze wartości $w_k(i)$ w momencie rozpoczęcia uczenia możliwe jest osiągnięcie docelowego, optymalnego zestawu parametrów, po odpowiednio długim procesie uczenia. W praktyce jednak nie jest obojętne, jakie wartości początkowe zostaną wybrane, gdyż długość procesu uczenia może zależeć od tego wyboru w sposób zasadniczy. Z tego względu należy starać się wybierać wartości $w_k(i)$ w chwili początkowej możliwie bliskie wartościom oczekiwanym jako docelowe, korzystne jest tu bowiem wykorzystanie każdej posiadanej a priori przesłanki. Przykładowo, jeśli wiadomo, że dana głoska, której funkcję przynależności rozważamy, charakteryzuje się dużą średnią częstotliwością (przykładowo może to dotyczyć głoski $s$), wówczas współ-

czynnik $w_4$ dla danej głoski (stojący przy pierwszym momencie widmowym, zgodnie z przyjętą numeracją wektora cech) należy wybrać duży i dodatni. Oczywiście tego typu decyzje są zawsze przybliżone, a ich wynik nie ma zastąpić procesu uczenia, lecz jedynie go przyspieszyć. Jednak wykorzystanie podobnych sugestii jest prawie zawsze celowe. W przypadku braku rozsądnych przesłanek dla innego wyboru celowe jest przyjmowanie wartości $w_k(i)$ równych zeru.

Jak wspomniano, przypadek funkcji postaci (5.14) rozważać trzeba jako przypadek szczególny ogólniejszej formuły (5.12). Zakładając, że funkcje $\varphi_k(X)$ zapewniają spełnienie warunku $w_k(i) = 0$ dla wszystkich $i = 1, 2, \ldots, L$, w przypadku $k > m$, można zapisać

$$g_i(X) = \sum_{k=1}^{m} \varphi_k(X) w_k(i) \tag{5.22}$$

Przyjęcie wzoru (5.22) w ogólnym przypadku zwiększa szanse na poprawne rozpoznawanie, jednak powstaje przy tym problem, jak prowadzić w takim przypadku proces uczenia, w celu ustalenia wartości współczynników $w_k(i)$. Zadanie to staje się prostsze przy zauważeniu możliwości przedstawienia formuły (5.22) jako złożenia przekształcenia przestrzeni cech $X$ w nową przestrzeń wektorów $Y$ zgodnie ze wzorami

$$y_k = \varphi_k(X) \quad k = 1, 2, \ldots, m \tag{5.23}$$

oraz funkcji liniowej postaci

$$g_i(X) = \hat{g}_i(Y) = \sum_{k=1}^{m} w_k(i) y_k \tag{5.24}$$

Dla funkcji (5.24) reguła uczenia jest prosta i oczywista:

$$w_k(i^n) = w_k(i^n) + y_k^n \quad k = 1, \ldots, m \tag{5.25}$$

$$w_k(j^n) = w_k(j^n) - y_k^n \quad k = 1, \ldots, m \tag{5.26}$$

gdzie $y_k^n = \varphi_k(X^n)$.

Warto zwrócić uwagę na pewną interpretację przekształcenia (5.23). Otóż funkcje przynależności postaci (5.22) stosuje się w przypadku, gdy granice między obszarami poszczególnych klas są w przestrzeni cech zbyt złożone, by można je było przybliżać hiperpłaszczyznami (5.16). Przekształcenie (5.23) sprawia jednak, że w przestrzeni wektorów $Y$ ($m$-wymiarowej) możliwe jest stosowanie formuły (5.24) — sprowadzającej granice ponownie do hiperpłaszczyzn — w dodatku przechodzących przez początek układu współrzędnych w przestrzeni $Y$. Ze względu na tę interpretację przestrzeń $Y$ bywa nazywana przestrzenią prostującą, a przekształcenie (5.23) — przekształceniem prostującym, gdyż w przestrzeni po transformacie (5.23) krzywoliniowe uprzednio granice stają się proste. Interpretacja ta może być wykorzystana do oceny przydatności określonej rodziny funkcji $\varphi_k(X)$ we wzorze (5.22). Najczęściej stosowane funkcje $\varphi_k(X)$ mają postać wielomianów

$$\varphi_k(X) = \sum_{\nu=0}^{k} a_\nu \prod_{\mu=1}^{\nu} x_\mu \tag{5.27}$$

można jednak rozważać celowość użycia funkcji $\varphi_k(X)$ dowolnej innej postaci.

Klasa funkcji przynależności opisana wzorem (5.22) jest potencjalnie bardzo duża. Zalety omówionej metody, polegające głównie na małych wymaganiach pamięciowych (trzeba rezerwować pamięć jedynie dla $L \cdot m$ współczynników $w_k(i)$ niezależnie od liczby elementów ciągu uczącego $N$), wskazują na celowość brania pod uwagę omówionych metod w zadaniach rozpoznawania mowy — także i z tego powodu, że dotychczasowe prace z tej dziedziny stroniły od omawianego podejścia. Jednak dla kompletności rozważań przytaczanych w tym podrozdziale celowe jest wzmiankowanie o jeszcze jednej grupie metod. Mowa o tak zwanym podejściu probabilistycznym, którego algorytmy nazywane bywają także Bayesowskimi (ze względu na wykorzystywany wzór opisujący prawdopodobieństwo a posteriori) względnie znane są pod nazwą analizy dyskryminacyjnej. Istota wspomnianych metod polega na wykorzystaniu w procesie rozpoznawania informacji statystycznych dotyczących rozpoznawanych obiektów i służących do rozpoznawania cech. Warto od razu podkreślić, że sens stosowania omawianej teraz grupy metod ograniczony jest do przypadku, kiedy wspomniane informacje statystyczne są dostępne i określone, zanim przystąpi się do prób rozpoznawania. Formułowane niekiedy w literaturze przypuszczenia, że ewentualnie brakującą informację statystyczną można estymować na podstawie ciągu uczącego, dowodzą nieznajomości problemu. Wymagane w metodach Bayesowskich informacje statystyczne są tego rodzaju, że dla ich poprawnego określenia niezbędne są tysiące dobrze opracowanych obserwacji. Zakładanie, że odpowiednie parametry zbierze się „przy okazji", prowadzi do bardzo pracochłonnych programów lub do bardzo miernych efektów rozpoznawania — często zresztą „osiąga się" obydwa wymienione niekorzystne skutki.

Nie należy jednak metod probabilistycznych pochopnie odrzucać. W badaniach nad sygnałem mowy i jego parametrycznym opisem zebrano duży materiał statystyczny, porządnie opracowany i zamieszczony w dziesiątkach opracowań raportów i publikacji. Gdyby to bogactwo danych zebrać, uporządkować, uzupełnić i powtórnie opracować — powstałaby bardzo użyteczna baza wiedzy nad wyraz przydatna przy rozpoznawaniu. Chwilowo jednak potrzebne informacje są niekompletne i rozproszone, co pozwala na podejmowanie jedynie ograniczonych prób stosowania probabilistycznych metod — z dość miernymi skutkami.

A oto podstawowe wiadomości na temat dyskutowanego podejścia i stosowanych w nim metod. Podstawę, jak wspomniano, stanowią dane statystyczne. Potrzebne są mianowicie prawdopodobieństwa występowania poszczególnych (podlegających rozpoznawaniu) klas:

$$p_i \quad i = 1, 2, \ldots, L \tag{5.28}$$

Te są z reguły łatwe do uzyskania (por. na przykład przytoczoną w rozdz. 4 tabl. 2, dostarczającą potrzebnych wartości prawdopodobieństw dla fonemów). Trudniej natomiast określić inne niezbędne dane, a mianowicie warunkowe gęstości prawdopodobieństwa występowania wektorów cech $X$ dla poszczególnych klas $i$

$$f_i(X) \quad i = 1, 2, \ldots, L \tag{5.29}$$

Zauważmy, że funkcje $f_i(X)$ muszą być dane w całej przestrzeni cech, co wyklucza dogodne i łatwo dostępne ujęcie numeryczne. Równocześnie przy zakładanej w książce strukturze przestrzeni cech oraz przy wybranym zbiorze rozpoznawanych obiektów wymaganie dane wzorem (5.29) oznacza konieczność określenia ponad stu funkcji pięciu zmiennych — co jest w w ogólnym przypadku zadaniem bardzo trudnym. Zadanie to dodatkowo komplikuje fakt, że poszczególne składowe wektora $X$ są skorelowane. Wyklucza to produktową technikę tworzenia wielowymiarowego rozkładu z rozkładów jednowymiarowych. W dodatku łatwo sprawdzić, że rozkłady (5.29) nie powinny być aproksymowane znanymi i łatwymi w użyciu rozkładami teoretycznymi. W szczególności rutynowy zabieg, jaki stosują autorzy podręczników rozpoznawania obrazów, polegający na założeniu normalnej postaci rozkładów (5.29) i sprowadzeniu całego problemu do prostszej obliczeniowo problematyki wyznaczenia parametrów rozkładu (wektorów średnich i macierzy kowariancji) — nie potwierdza w praktyce swojej przydatności. W większości interesujących przypadków z zakresu rozpoznawania mowy rozkłady (5.29) nie są normalne, a co więcej w wielu przypadkach można wykazać ich nieunimodalność. W sumie — wyznaczenie funkcji $f_i(X)$ jest na tyle złożone, że przydatność omówionej niżej procedury rozpoznawania może być rozważana jedynie pod warunkiem posiadania niezbędnej apriorycznej wiedzy statystycznej — bez konieczności jej pozyskiwania wyłącznie do rozpoznawania.

Jeśli jednak pominąć wskazane niedogodności z pozyskaniem danych początkowych, to metody probabilistyczne można ocenić bardzo pozytywnie. Po pierwsze, można udowodnić, że metody te są optymalne z punktu widzenia wielkości strat ponoszonych na skutek błędów w procesie rozpoznawania. Z tego powodu klasyfikator Bayesowski jest nazywany optymalnym i stosowany jako „punkt odniesienia" przy ocenie innych algorytmów rozpoznawania. Po drugie, proces rozpoznawania jest prosty i szybki, a uczenia może wcale nie być — z oczywistą korzyścią dla prostoty budowy i strojenia aparatury rozpoznającej. Po trzecie, istnieje teoretyczna możliwość „uczulania" algorytmu rozpoznającego, opartego na metodzie probabilistycznej, na pewne konkretne rozróżnienia. Można bowiem założyć „ceny" różnego rodzaju błędów, oczywiście wyższe tam, gdzie pomyłka jest bardziej dotkliwa w skutkach, mniejsze zaś dla błędów mało wpływających na końcowe rezultaty. W przypadku rozpoznawania sygnału mowy takie rozróżnienie jest bardzo pożądane, gdyż niektóre fonemy mają znacznie większe znaczenie dla rozpoznania całego wyrazu niż inne. Przykładowo znacznie więcej informacji przenoszą generalnie spółgłoski niż samogłoski. Minimalizując łączną „cenę" błędów otrzymuje się w omawianych metodach program rozpoznający, przywiązujący szczególną wagę do tych najistotniejszych rozpoznań.

Pomijając jednak z braku miejsca bardziej szczegółowe rozważania oraz odsyłając bardziej dociekliwych Czytelników do literatury w celu prześledzenia uzasadnień podanego dalej algorytmu, można stwierdzić, że w pod-

stawowym wariancie rozważanej metody funkcja przynależności obiektu opisanego wektorem cech $X$ do określonej klasy $i$ wyraża się za pomocą danych (5.28) i (5.29) wyjątkowo prosto:

$$g_i(X) = p_i f_i(X) \quad i = 1, 2, \ldots, L \tag{5.30}$$

Oczywiście pozorna prostota tego wzoru ukrywa fakt, że funkcję gęstości prawdopodobieństwa $f_i(X)$ trzeba wyrazić analitycznie, co prowadzi do niekiedy bardzo złożonych formuł. Jedynie w przypadku, kiedy $f_i(X)$ można traktować jako rozkład normalny, funkcje (5.30) sprowadzają się do form kwadratowych, a w szczególnie dogodnym przypadku jednakowych macierzy kowariancji dla wszystkich klas i funkcja przynależności daje się wyrazić jako funkcja liniowa.

## 5.7. Pozostałe elementy systemu rozpoznającego

Omówione algorytmy, pozwalając rozpoznawać elementy mowy polskiej, mogą stanowić domknięcie zadania identyfikacji sygnału, rozumianego jako zadanie badawcze. Z punktu widzenia celów praktycznych, a w szczególności ze względu na możliwość wykorzystania w systemach automatyki, układ rozpoznający jedynie elementy sygnału mowy musi być jednak uznany za niekompletny. Konieczne są przecież dalsze etapy analizy, kończące się rezultatem, który w p. 5.2 nazwano umownie rozumieniem mowy. Niestety, aktualna wiedza na temat tych etapów jest nader niekompletna i fragmentaryczna. Ponadto wiele uzyskanych rezultatów dotyczy języka naturalnego wprowadzanego do maszyny na drodze alfanumerycznej, zatem mimo oczywistych związków między ortograficznym zapisem wypowiedzi a jej dźwiękową formą — wyniki te można wiązać z hasłem „sygnał mowy" jedynie z uwzględnieniem całego szeregu zastrzeżeń. Z tych powodów wszystkie — nader złożone zresztą — procesy i operacje dokonywane w celu „rozumienia" mowy istniejącej już w maszynie w postaci rozpoznanych (być może z błędami!) elementów zbierzemy w tym podrozdziale i omówimy nader skrótowo.

Podstawowy problem, jaki wyłania się przy próbie scalania elementarnych rozpoznanych elementów, wiąże się z niejednakowym, zmiennym tempem emisji sygnału mowy. Istotnie, jeśli wybrano określone jednostki podlegające rozpoznawaniu (na przykład fonemy lub mikrofonemy), to wzorzec wypowiedzi wyższego rzędu (wyrazu, krótkiego, kilkuwyrazowego hasła lub całego zdania) — wyrażony oczywiście w postaci sekwencji takich samych, jak rozpoznane elementów sygnału mowy — będzie bez wątpienia zawierał tych jednostkowych elementów mniej lub więcej niż sygnał rozpoznawany, gdyż bardzo mało jest prawdopodobne, aby tempo artykulacji wzorca i rozpoznawanego wyrazu było takie samo. W dodatku nierównomierności tempa wypowiedzi mogą być różne w różnych jej częściach, w związku z czym niemożliwa jest prosta normalizacja typu przeskalowania osi czasu. Cóż pozostaje? Trzeba znaleźć metodę dopasowania elementów wzorca do elementów rozpoznawanej wypowiedzi, przy czym trzeba to zrobić dla

wszystkich rozpatrywanych wzorców. Oznacza to, że metoda dopasowania musi być niezawodna i szybko działająca. Z licznych prób i koncepcji rozwiązania tego zagadnienia na uwagę zasługuje najpopularniejsza ostatnio metoda oparta na technice programowania dynamicznego. Metoda programowania dynamicznego służy zasadniczo do rozwiązywania wieloetapowych zadań optymalizacji, gdyż z myślą o takich problemach została opracowana przez Ryszarda Bellmana w latach sześćdziesiątych i spopularyzowana wśród automatyków i ekonomistów na całym świecie. Okazuje się jednak, że metoda ta może znaleźć zastosowanie w zadaniu nieliniowej normalizacji skali czasu przy rozpoznawaniu mowy. Przedstawimy dalej podstawowe koncepcje wykorzystania programowania liniowego do identyfikacji wzorca wypowiedzi opierając się na obszernej pracy dra Stefana Grocholewskiego.

Zacznijmy od przykładu. Niech rozpoznawane słowo ma postać

*ossaa*

zaś wzorzec, z którym chcemy je porównać, niech ma postać

*oossa*

Jak widać, różnica jest nieznaczna i bez wątpienia chodzi o jedno i to samo słowo, jednak porównanie elementów zajmujących w obydwu wypowiedziach te same pozycje daje wynik negatywny: zgodnych jest jedynie trzy spośród pięciu pozycji. Jaka jest rada? Zacznijmy od ilościowego wyrażenia miary niezgodności. Wprowadźmy w tym celu pojęcie odległości*) elementów wzorca i analizowanej wypowiedzi. Załóżmy dla ustalenia uwagi, że odległość danego fonemu od niego samego wynosi 1, samogłoski są od siebie odległe o 2, spółgłoska zaś odległa jest od samogłosek o 3 umowne jednostki:

$$d(o, o) = d(s, s) = d(a, a) = 1 \tag{5.31}$$

$$d(a, o) = 2 \tag{5.32}$$

$$d(o, s) = d(a, s) = 3 \tag{5.33}$$

Na podstawie tych ustaleń można określić mapę odległości między elementami wzorca a elementami podlegającej analizie wypowiedzi:

$$\begin{array}{cccccc} a & 2 & 3 & 4 & \mathbf{1} & \mathbf{1} \\ s & 3 & \mathbf{1} & \mathbf{1} & 3 & 3 \\ s & 3 & \mathbf{1} & \mathbf{1} & 3 & 3 \\ o & \mathbf{1} & 3 & 3 & 2 & 2 \\ o & \mathbf{1} & 3 & 3 & 2 & 2 \\ & o & s & s & a & a \end{array} \tag{5.34}$$

*) Warto zauważyć, że dla dyskutowanych uprzednio mikrofonemów zastosowanie analizy skupień dostarcza miar odległości między poszczególnymi skupieniami. Wynik ten można tu bezpośrednio wykorzystać. Jednak wprowadzając miarę odległości elementów wzorca i rozpoznawanej wypowiedzi można w ogóle pominąć etap rozpoznawania elementów. Opisana dalej procedura może funkcjonować na próbkach sygnału bez ich wcześniejszej klasyfikacji i może być traktowana jako uogólnienie metody najbliższego sąsiada. W istocie, często wykorzystuje się programowanie dynamiczne wprost do rozpoznawania.

Optymalne dopasowanie wzorca do wypowiedzi nastąpi wówczas, kiedy znaleziona zostanie taka linia przebiegająca od lewego dolnego rogu mapy do prawego górnego, na której suma elementów (odległości) będzie minimalna. W tabeli — wzór (5.34) — możliwych jest 321 różnych linii, optymalna z nich została wytłuszczona i jak łatwo się przekonać — odpowiada ona najbardziej logicznemu przyporządkowaniu elementów wzorca elementom rozpoznawanej wypowiedzi.

Rozważany przykład był prosty i poszukiwane połączenia łatwo było znaleźć. Aby algorytm mógł działać niezależnie od stopnia złożoności, trzeba go nieco sformalizować. W tym celu przerobimy nieco tabelę ze wzoru (5.34), zastępując nazwy elementów wzorca i wypowiedzi numerami (zamiast *o* będzie 1, zamiast końcowego *a* będzie 5), zaś zamiast odległości elementów *i* oraz *j*, oznaczanej $d(i,j)$ wpiszemy do tabeli minimalne skumulowane odległości od punktu końcowego (5, 5), oznaczone przez $D(i, j)$ i wyliczane ze wzoru

$$D(i,j) = d(i,j)+\min[D(i+1,j), D(i+1,j+1), D(i,j+1)] \qquad (5.35)$$

Jak widać reguła (5.35) jest rekursywna i musi być wyliczana w ustalonej kolejności, poczynając od $D(5, 5)$, które oczywiście wynosi z definicji 1. Rezultat tych obliczeń podano we wzorze (5.36)

| | | | | | |
|---|---|---|---|---|---|
| 5 | 10 | 8 | 5 | **2** | **1** |
| 4 | 7 | 4 | **3** | 4 | 4 |
| 3 | 7 | **4** | 4 | 7 | 7 |
| 2 | **5** | 7 | 7 | 9 | 9 |
| 1 | **6** | 10 | 10 | 11 | 11 |
| | 1 | 2 | 3 | 4 | 5 |

(5.36)

Łatwo zauważyć, że optymalna trajektoria jest teraz wyznaczona dokładnie: wytyczają ją (poczynając od prawego górnego rogu) punkty o najmniejszych wartościach. Właściwy cel całej procedury zawarty jest jednak w punkcie początkowym, o współrzędnych (1,1). Jego wartość (w rozważanym przypadku 6) jest miarą stopnia zróżnicowania rozważanej wypowiedzi i badanego wzorca. Określając analogicznie wartości $D(1,1)$ dla innych wzorców możemy bez trudu wybrać ten wzorzec, do którego rozważany (rozpoznawany) wyraz jest najbardziej podobny, niezależnie od ewentualnych różnic skali czasu wzorców i rozpoznawanej wypowiedzi.

Omówiona procedura postępowania ma wiele odmian, przy czym zasadnicze ulepszenia, które wprowadzają do rozważanego schematu poszczególni autorzy, polegają na zmniejszeniu pracochłonności obliczeń. W ogólnym przypadku nie ma bowiem potrzeby wyliczania wszystkich wartości $D(i, j)$ i zapełniania matrycy postaci (5.36) w całości, gdyż optymalna trajektoria zawsze przebiega w pobliżu głównej przekątnej matrycy — tak jak to było pokazane na przykładzie. Zamiast wyliczać wszystkie wartości $D(i, j)$ wystarczy więc analizować wyłącznie „pas" o ustalonej szerokości, biegnący wzdłuż głównej diagonali, natomiast metody dyskutowane przez różnych autorów sprowadzają się do tego, by efektywnie ten pas wytyczać. Bliższe szczegóły tego procesu znaleźć można w cytowanej literaturze, warto jedynie

dodać, że jakość rozpoznawania, osiągana przy metodach stosowania programowania dynamicznego wprost do źródłowego sygnału, jest bardzo wysoka (ponad 90% poprawnych rozpoznań). Można oczekiwać, że w połączeniu ze wstępną identyfikacją elementów jakość rozpoznawania znacznie wzrośnie.

W odniesieniu do zagadnień wyższego poziomu, niż omówiony poziom leksykalny, zastosowanie mają pozycje literatury dotyczące metod analizy i przetwarzania języka naturalnego w formie tekstów alfanumerycznych. Problematyka ta ma obecnie bogatą literaturę i wydaje się, że przynajmniej część uzyskanych na tej drodze wyników znajdzie zastosowanie w technice rozpoznawania mowy. Wszelkie pogłębione analizy i bardziej kategoryczne sądy są tu jednak przedwczesne: zarówno dziedzina przetwarzania języka naturalnego nie dorosła jeszcze do tego, aby dostarczać pewnych i uniwersalnych algorytmów, jak również dyskutowana tu problematyka rozpoznawania sygnału mowy nie dopracowała się metod identyfikacji dostarczających tekst o porównywalnej wierności, jak terminale alfanumeryczne. Warto w uzupełnieniu dodać, że największe osiągnięcia w dziedzinie komputerowego przetwarzania tekstów języka naturalnego odnotowano w zakresie języka angielskiego i japońskiego. Są to języki o tak odmiennej strukturze i gramatyce w stosunku do języka polskiego, że korzystanie z zagranicznych osiągnięć ograniczać się musi do ogólnej inspiracji oraz podstawowych pomysłów. Podstawa systemu analizy i rozpoznawania tekstów języka polskiego musi powstać w pracowniach polskich badaczy. Prace na ten temat trwają od lat, przy czym największe osiągnięcia ma, jak się wydaje, grupa docenta Leonarda Bolca z Uniwersytetu Warszawskiego. Wybrane publikacje tej grupy, cytowane na końcu książki, stanowią znacznie lepsze wprowadzenie do tego zagadnienia niż zamieszczony tu, skrótowy z konieczności, opis.

W najbardziej uproszczonym zarysie wspomniane etapy przetwarzania wyników rozpoznawania sygnału mowy mogą się przedstawiać następująco. Po zidentyfikowaniu wyrazów, których wzorce zostały uzgodnione w omówiony sposób z elementami rozpoznanymi w sygnale mowy, możliwe jest także określenie (z pewną dokładnością) formy gramatycznej wyrazu oraz jego roli w zdaniu. Równocześnie przechowywany w pamięci komputera zbiór schematów spodziewanych zdań oraz wykaz słów, odgrywających kluczową rolę przy identyfikacji sensu wypowiedzi pozwala na sformułowanie hipotezy co do treści zadania. Uzupełniające informacje mogą pochodzić z systemu analizującego intonację wypowiedzi. Wiadomo, że śledząc zmiany częstości tonu krtaniowego oraz obrys czasowy i amplitudowy wypowiedzi można wyróżnić typ zadania, akcentowane elementy i ogólny schemat wypowiedzi. Wszystkie te elementy łącznie składają się na identyfikację sensu wypowiedzi, który w istocie jest — w zakresie założonego kontekstu rozmowy — identyfikowany dość pewnie i precyzyjnie. Zwraca uwagę w tym schemacie jego sztywność, wynikająca z użycia wzorców spodziewanych zdań i znaczeń słów kluczowych, na których opiera się identyfikację. Schemat taki funkcjonuje bardzo dobrze, jeśli rozmowa mieści się wewnątrz

założonego scenariusza, zawodzi natomiast całkowicie w przypadku opuszczenia tych ram. Wynika to z faktu odmiennej pozycji, w jakiej występuje przy rozmowie komputer i biorący udział w dialogu człowiek. Przy rozmowie pomiędzy ludźmi niewyobrażalnie dużą rolę odgrywa wspólne rozmówcom dziedzictwo kulturowe, wiedza o świecie, doświadczenie. Tego wszystkiego maszyna nie ma i dlatego każda próba dialogu będzie zawsze napotykała przeszkody — chyba że nauczymy się przekazywać komputerom naszą wiedzę o świecie, a maszyny staną się zdolne z wiedzy tej korzystać. Problem inżynierii wiedzy — jak się podkreśla z najnowszych pracach z zakresu sztucznej inteligencji — jest problemem kluczowym dla bardzo wielu zagadnień: od omawianych tu problemów rozpoznawania mowy i analizy języka naturalnego, do automatycznego tłumaczenia z jednego języka naturalnego na inny, systemów odpowiadających na pytania (systemów ekspertowych) i sterowania robotów włącznie. Na jego rozwiązanie łoży się na całym świecie ogromne sumy, być może zatem jego zasadnicze elementy zostaną rozpracowane na tyle w bliskiej przyszłości, że odegrają istotną rolę w zagadnieniach rozpoznawania mowy. Bardziej prawdopodobny jest jednak odwrotny przypadek: to właśnie pokonanie barier związanych z rozpoznawaniem mowy wpłynie na takie upowszechnienie językowych kontaktów z maszynami, że problem rozpoznawania treści wypowiedzi, gromadzenia wiedzy i jej wykorzystywania rozwiązany zostanie niejako przy okazji. W chwili obecnej jest to jednak futurologia. Dzień dzisiejszy to w najlepszym przypadku systemy działające zgodnie z przytoczonym wyżej schematem: słów kluczowych, scenariuszy i szablonów zdań. Jest to prymitywne, ale przy konkretnie wyznaczonych celach systemu rozpoznawania — skuteczne. Na to nas dzisiaj stać. A jutro ...?

# 6

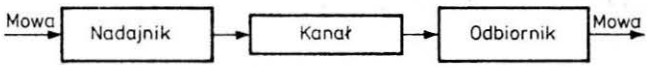
# Sygnał mowy w telekomunikacji

## 6.1. Sygnał mowy w kanale telekomunikacyjnym

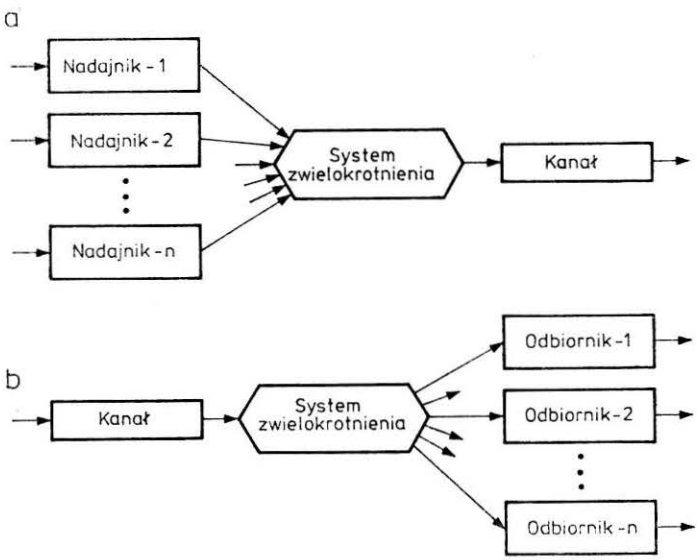
W poprzedzających rozdziałach analizowano sygnał mowy w warunkach sztucznych i nienaturalnych: w laboratorium, gdzie bada się jego czasowe, częstotliwościowe i parametryczne własności, względnie w systemie automatyki, gdzie mowa jest wykorzystywana do przekazywania informacji pomiędzy personelem a podlegającym nadzorowi systemem sterowania obiektu. Tymczasem sygnał mowy powstał i najczęściej do dziś jest wykorzystywany po prostu jako środek komunikowania się pomiędzy ludźmi. Jeśli komunikacja ta zachodzi bezpośrednio, wówczas technik nie ma w tym żadnego udziału i rozważanie takiego przypadku w niniejszej książce mija się z celem. Jeśli jednak komunikacja głosowa za pomocą mowy odbywa się na dużą odległość, wówczas muszą w niej uczestniczyć urządzenia techniczne, a zainteresowanie sygnałem mowy, który ma być przesyłany, rośnie proporcjonalnie do kosztów jego przesyłania.

Wzrost odległości przesyłania sygnału z jednej strony i wzrost liczby osób zainteresowanych telekomunikacją (jako jej użytkownicy, a nie jako twórcy systemów) — z drugiej strony, warunkują ustawiczny wzrost stopnia zainteresowania środkami i metodami przesyłania dużej liczby rozmów na duże odległości — przy minimalnych kosztach. Rozważając najprostszy sche-

mat łącza telekomunikacyjnego (rys. 6-1) możemy stwierdzić, że koszty przesyłania informacji — w szczególności sygnału mowy — przez to łącze rozkładają się nierównomiernie na poszczególne elementy. Koszt nadajnika i odbiornika nie różnią się w istotny sposób dla transmisji sygnału na małe i na duże odległości, są to zresztą prawie zawsze koszty niewielkie. Nato-




6-1. Ogólny schemat przesyłania mowy przez kanał telekomunikacyjny. Na ogół koszt kanału jest większy niż koszty nadajnika i odbiornika, co skłania do poszukiwania takich rozwiązań, które za cenę wzrostu złożoności nadajnika i odbiornika dają możliwość lepszego wykorzystania kanału (przesłania nim większej liczby rozmów)




6-2. Część nadawcza (u góry) i część odbiorcza (u dołu) systemu telekomunikacyjnego ze zwielokrotnionym wykorzystaniem toru. Za pomocą odpowiedniego podziału sygnału w dziedzinie czasu lub w dziedzinie częstotliwości można ten sam tor tejekomunikacyjny wykorzystywać do przesyłania dużej liczby rozmów

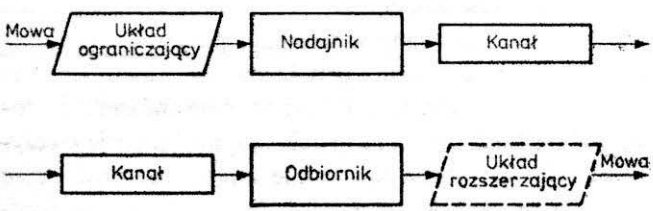
miast koszty kanału przesyłania informacji rosną ze wzrostem odległości przesyłania i ze wzrostem wymogów stawianych jakości sygnału na odbiorczym końcu łącza, przy czym wspomniany wzrost kosztów jest znacznie szybszy niż proporcjonalny. Innymi słowy, przesłanie sygnału na dwukrotnie większą odległość z reguły kosztuje więcej niż dwukrotna cena przesłania na małą odległość, w czym głównie partycypują koszty budowy, utrzymania i konserwacji linii telekomunikacyjnych. Zestawienie małych kosztów nadajnika i odbiornika z bardzo dużymi kosztami linii skłania do poszukiwania możliwości zwiększenia efektywności wykorzystania linii przez przyłączenie do jednej linii większej liczby nadajników i odbiorników (rys. 6-2) oraz przez użycie urządzeń wielokrotnego wykorzystania toru. Nie ma potrzeby dyskutowania tu możliwych metod wielokrotnego wykorzystania toru, jest ich bowiem bardzo wiele: z podziałem czasowym, częstotliwościowym itd. Problematyka ta ma zresztą własną, obszerną literaturę. Z punktu widzenia analizy i przetwarzania sygnału bardziej interesujące są natomiast środki, które można podejmować w celu zmniejszenia informacyjnej

objętości sygnału mowy w kanale telekomunikacyjnym. Niezależnie bowiem od tego, jak zbudowany jest kanał i jaka obowiązuje organizacja przesyłania w nim informacji dźwiękowej: ze zwielokrotnieniem lub bez niego, z podziałem w dzidzinie czasu czy częstotliwości, cyfrowa czy analogowa, zawsze konfrontowane są ze sobą dwie wartości — przepustowość kanału i objętość informacyjna transmitowanego sygnału. Objętość sygnału musi być mniejsza od przepustowości kanału, gdyż w przeciwnym przypadku pojawią się niemożliwe do skorygowania zniekształcenia i przekłamania sygnału, wiążące się z niekontrolowaną utratą informacji. Równocześnie możliwości poszerzenia przepustowości kanału są ograniczone i bardzo kosztowne. W sumie rozwiązaniem jest więc jedynie ograniczenie objętości sygnału — im bardziej radykalne, tym korzystniejsze.

Ograniczenie takie jest możliwe, co bezpośrednio wynika z rozważań przeprowadzanych w poprzednich rozdziałach. W szczególności w p. 3.3 uzasadniono tezę, że proces percepcji mowy wydobywa z pełnego sygnału docierającego do błony bębenkowej, zawierającego znaczne ilości informacji, tylko niektóre jego własności. Wykazano przy tym, że na jakość percepcji, zrozumiałość i wyrazistość mowy wpływają wszystkie jej cechy, lecz nie wszystkie w jednakowym stopniu. Przykładowo — o czym będzie mowa w p. 3.3 — nawet bardzo znaczne ograniczenie pasma częstotliwości analizowanego sygnału mowy nie powoduje zauważalnego zmniejszenia jej zrozumiałości. Fakt ten zresztą jest już wykorzystywany w telefonii, ponieważ przekazywane pasmo sygnału jest ograniczone do przedziału $350 \div 3400$ Hz, bez wpływu na jakość transmisji i skuteczność przekazywania informacji. Podobne rozważania wiązać można z amplitudową skalą sygnału, gdzie rzeczywisty zakres dynamiki sygnału można radykalnie zwęzić nie powodując dużej straty zrozumiałości przekazywanej mowy, chociaż jej jakość (oceniana subiektywnie przez odbiorców) bardzo znacznie pogarsza się — trudno rozpoznać charakterystyczne cechy indywidualnego głosu, a słuchanie przekazywanych informacji staje się męczące i nieprzyjemne. Znaczne rezerwy tkwią także w czasowej strukturze sygnału. Zarówno w szumowych, jak i w quasi-periodycznych fragmentach sygnału można wyróżnić charakterystyczne, krótkie (kilkumilisekundowe) fragmenty sygnału, których repetycja dostarcza takiej samej informacji, jak transmisja całej głoski bez ograniczeń czasowych. Na tej zasadzie funkcjonują niektóre spośród syntezatorów mowy, omawianych w p. 2.3, w których możliwość zastępowania pełnego przebiegu sygnału w całym czasie jego trwania skróconą reprezentacją zasadniczo ogranicza obszar niezbędnej pamięci i umożliwia efektywną syntezę, przy ograniczonym zbiorze wzorców. Ten sam fakt usiłowano wykorzystywać w teletransmisji, jakkolwiek bez powodzenia, ze względu na duże trudności wydobywania takich „charakterystycznych" fragmentów z sygnału mowy w czasie jego trwania oraz nieproporcjonalnie (do uzyskiwanych efektów) rozbudowany układ odbiornika, odtwarzającego zrozumiały dla człowieka sygnał mowy. A jednak redukcja ilości informacji zawartej w sygnale mowy, niezbędnej do jego przesłania i bezbłędnego odtworzenia w odbiorniku, jest stale aktualnym zadaniem badawczym i technicznym.

Skala potencjalnych możliwości w tym zakresie oszacowana została w p.4.6. Okazuje się, że stosując odpowiednio efektywne metody kompresji sygnału można — na razie jedynie teoretycznie niestety — przesłać przez ten sam system kanałów i łączy teletransmisyjnych kilkadziesiąt razy więcej rozmów, niż to ma miejsce obecnie. Jest więc o co się starać, chociaż uzyskać można to kosztem rozbudowy nadajnika i odbiornika sygnałów. Wspomniano jednak, że już obecnie koszty urządzeń nadawczych i odbiorczych są wielokrotnie niższe niż linii przesyłowych. Nawet w przypadku wprowadzenia wysokowydajnych łączy światłowodowych prawidłowość ta się utrzyma, między innymi ze względu na systemy wielkiej skali integracji analogowe i cyfrowe, obniżające koszt układów elektronicznych o rzędy wielkości w ciągu niewielu lat.

Reasumując można powiedzieć, że zarówno systemy linii jednokrotnie wykorzystywanych (rys. 6-1), jak i systemy ze zwielokrotnieniem wykorzystania toru (rys. 6-2) mogą wiele zyskać, jeśli w konstrukcji nadajników i odbiorników uwzględni się możliwości redukcji objętości informacyjnej sygnału mowy przed jego przesłaniem, a na odbiorczym końcu łącza dokona się odtworzenia pełnej formy sygnału w celu przekazania go w dogodnej do słuchania postaci odbiorcy informacji. Możliwe są przy tym różne podejścia. W najprostszym przypadku (rys. 6-3) sygnał mowy przed przekaza-

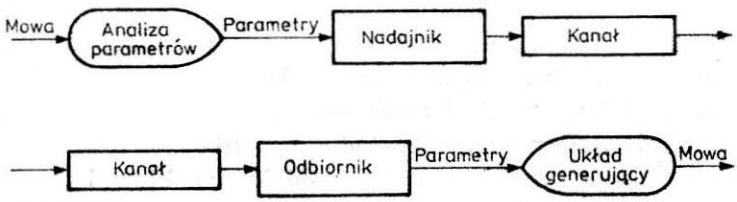



6-3. Układ ograniczający informacyjną objętość sygnału mowy w kanale pozwala używać tańszego kanału lub zwiększać stopień zwielokrotnienia wykorzystania w kanale doskonalszym. Ograniczenie może dotyczyć amplitudowego, czasowego lub częstotliwościowego wymiaru sygnału. Rozszerzenie sygnału w odbiorniku (odtworzenie w naturalnej postaci) nie zawsze bywa konieczne

niem może podlegać ograniczeniu w zakresie swoich podstawowych parametrów, wyznaczających jego informacyjną objętość. Można więc rozważać sygnał ograniczony w dziedzinie częstotliwości, w zakresie amplitud oraz — w omówiony wyżej sposób — w czasie. Ze wszystkiego, co uprzednio zostało powiedziane, wynika, że możliwości tkwiące we wskazanym podejściu są ograniczone i w zasadzie zostały już wykorzystane w istniejących systemach telefonii rozmównej. Nowych rozwiązań poszukiwać trzeba na innej drodze.

W podrozdziale 4.4 wykazano, że sygnał mowy może być opisany za pomocą parametrów, przy czym osiągana jest na ogół znaczna oszczędność w zakresie ilości informacji zawartej w sygnale. Na tej zasadzie możliwe jest skonstruowanie układu ograniczającego objętość przesyłanej informacji (rys. 6-4). W nadajniku z sygnału są wydobywane parametry potrzebne do jego poprawnego odtworzenia na odbiorczym końcu łącza. Parametry te są przesyłane do urządzenia odbiorczego i tam następuje synteza mowy.

Parametrów można wyróżnić wiele: widmowych, czasowych, liniowej predykcji itp. W zależności od tego, jakie parametry są wydobywane, prostsza lub bardziej złożona jest budowa układu nadawczego i odbiorczego, większa lub mniejsza jest osiągana kompresja, a także lepsza lub gorsza jest jakość przekazywanej mowy. Urządzenia służące do ograniczania objętości sygnału mowy metodą wydzielania i przesyłania jej parametrów nazywane są typowo wokoderami. Technika wokoderów liczy już sobie kilkadziesiąt lat, chociaż nadal daleko do tego, aby można było istniejące konstrukcje uznać

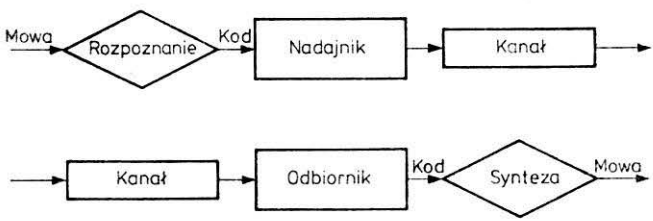



6-4. Parametryczny sposób kompresji mowy. W nadajniku dokonywana jest analiza sygnału i wydobywane są jego parametry. Przesyłanie parametrów angażuje kanał w znacznie mniejszym stopniu niż przesyłanie sygnału. W odbiorniku układ generujący sterowany parametrami odtwarza sygnał mowy o akceptowalnej jakości

za optymalne i ostateczne rozwiązanie postawionego problemu, podobnie jak odległa jest jeszcze chwila, kiedy technika wokoderów będzie powszechnie stosowana w telefonii użytkowej. Istnieje kilka technik, w których typowo buduje się wokodery, zatem pomimo omówienia zasadniczych faktów w p. 4.4 poświęcimy dodatkowo kolejny, następny podrozdział, aby omówić nieco dokładniej używane w teletransmisji metody parametrycznego opisu sygnału mowy, a także podać nieco szczegółów na temat budowanych wokoderów.

Zanim to jednak nastąpi, przedyskutujemy inne możliwe — chociaż na obecnym etapie futurystyczne — podejścia do zadania kompresji sygnału mowy w kanale telekomunikacyjnym. W podrozdziale 4.6 podano oszacowania wielkości nadmiarowości sygnału mowy oraz przedyskutowano źródła tej nadmiarowości. Nie powtarzając przytoczonych tam argumentów należy stwierdzić, że nadmiarowość akustycznych struktur sygnału, usuwana (nie do końca zresztą) przez stosowanie nawet najdoskonalszych wokoderów, stanowi jedynie część nadmiarowości całego sygnału. Znacznie dalej posunięta redukcja informacyjnego nadmiaru, a zatem bez porównania większe oszczędności kosztów przesyłania sygnału mowy możliwe są w przypadku zastosowania rozwiązania, przedstawionego schematycznie na rys. 6-5. Jak widać, istota koncepcji polega na dokonaniu w odbiorniku próby rozpoznania sygnału mowy. Następnie kody rozpoznanych elementów zostają przesyłane przez kanał teletransmisyjny (praktycznie całkowicie bez nadmiarowych elementów, jeśli tylko nie obawiamy się zakłóceń), natomiast na odbiorczym końcu łącza syntezator mowy, sterowany nadchodzącymi sygnałami, dokonuje odtworzenia sygnału. Zalety takiego systemu są oczywiste: daje on możliwość zredukowania nadmiarowych informacji niemal do zera, a transmisji podlega wyłącznie niezbędna, merytoryczna treść przekazywanych wiadomości. W dodatku pod wieloma względami

układy zastosowane w schemacie na rys. 6-5 mogą być prostsze i tańsze niż układy omawiane w p. 2.3 oraz 5.2÷5.7. Proces rozpoznawania sygnału przed jego nadaniem nie musi bowiem obejmować elementów analizy leksykalnej, syntaktycznej i semantycznej — wystarcza identyfikacja elementów. Zresztą jakość rozpoznawania nie musi też być najwyższej jakości, gdyż na odbiorczym końcu łącza słucha wypowiedzi człowiek — myślący, rozumiejący swojego rozmówcę, mogący odtwarzać sobie brakujące elementy sygnału na podstawie kontekstu — obejmującego zarówno dany wyraz, całe zdanie lub kilka sąsiednich zdań. Wreszcie — zawsze można poprosić o powtórzenie niezrozumiałego fragmentu. W przypadku kiedy odbiorcą rozpoznawanych poleceń był komputer (por. rozdz. 5) wymagania musiały być pod każdym względem bardziej rygorystyczne. Zmienne tempo mowy nie stanowi w rozważanym przypadku żadnej przeszkody — przeciwnie, sterowany w naturalnym tempie syntezator na odbiorczym końcu łącza zachowa — przynajmniej w części — intonację i dynamikę wypowiedzi, co niwelować będzie częściowo przykrą własność przedstawionego na rys. 6-5




6-5. Hipotetyczna koncepcja kompresji mowy przy przesyłaniu jej metodą rozpoznawania w nadajniku, przesyłania zakodowanej treści wypowiedzi kanałem i syntezy mowy w odbiorniku na podstawie nadesłanego kodu. Taki system nie jest jeszcze dziś możliwy do skonstruowania, ale miałby on najlepsze parametry w sensie oszczędnego wykorzystania kanału

hipotetycznego systemu komunikacji, a mianowicie brak możliwości słuchania naturalnego brzmienia głosu rozmówcy. Synteza mowy jest zresztą w omawianym przypadku także prostsza i łatwiejsza do realizacji, niż w przypadku niezależnego od człowieka produkowania mowy przez urządzenie techniczne (por. p. 2.3), gdyż znaczna część parametrów procesu syntezy (na przykład czas trwania poszczególnych elementów mowy) zadawana jest przez łącze telekomunikacyjne w naturalnym następstwie procesu analizy mowy.

Reasumując można stwierdzić, że koncepcja ograniczenia informacyjnej objętości sygnału mowy na drodze rozpoznawania jej w nadajniku i syntezowania w odbiorniku jest realna, chociaż jeszcze nigdzie nie realizowana. Koncepcja ta stanowi równocześnie łącznik między rozważaniami przytoczonymi w tym rozdziale a wynikami zawartymi w rozdziałach poprzednich, okazuje się bowiem, że proces rozpoznawania mowy wykorzystać można nie tylko w zadaniach automatyki i informatyki, ale również może on znaleźć zastosowanie w telekomunikacji. Idąc jeszcze dalej tym samym tropem można twierdzić, że rozwiązanie problemu rozpoznawania mowy i udostępnienie efektywnych, tanich i skutecznie działających systemów rozpoznają-

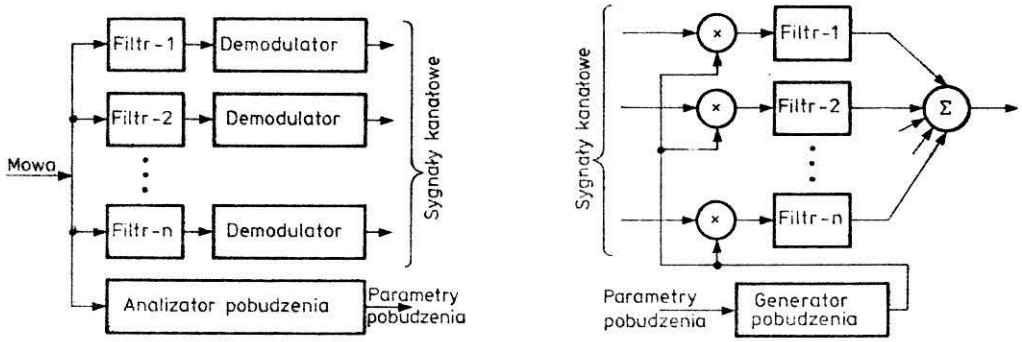
cych do powszechnego użytku, uczyni zbyteczne wszystkie specjalizowane metody kompresji sygnału mowy — ze wszystkimi typami wokoderów. Proces syntezy mowy jest bowiem — co wielokrotnie podkreślano— znacznie łatwiejszy i w zasadzie już obecnie całkowicie opanowany. Obiegowo wymienia się już nawet nazwę „wokodery fonemowe" jako właściwy termin dla urządzeń działających na zasadzie podanej na rys. 6-5. Nazwa ta nie jest jednak najbardziej trafna, biorąc pod uwagę znacznie głębszy stopień przetworzenia sygnału mowy w tego typu urządzeniach w stosunku do typowej techniki wokoderowej, a ponadto nie jest bynajmniej sprawą przesądzoną, czy rozpoznawanymi i kodowanymi elementami mają być właśnie fonemy. Być może bardziej celowe będzie użycie i tutaj proponowanych w rozdziale 5 mikrofonemów, a może przeciwnie — korzystne będzie użycie diad, triad, sylab — czy nawet całych wyrazów. Potrzebne są dodatkowe badania, analizy porównawcze i doświadczenia praktyczne, którym — na razie — stoi na przeszkodzie brak efektywnych algorytmów, metod i systemów rozpoznawania mowy.

## 6.2. Metody kompresji sygnału mowy

Przyjmując do wiadomości futurystyczną wizję, naszkicowaną pod koniec poprzedniego podrozdziału, musimy jednak zająć się metodami, które znajdują zastosowanie już obecnie — chociaż niezbyt często. Mowa o typowych wokoderach, urządzeniach, które w ogólnym przypadku wydobywają ustalone parametry nadawanego sygnału mowy, kodują je i przesyłają do odbiornika, który na ich podstawie dokonuje odtworzenia sygnału o zadanych parametrach — czyli w przybliżeniu zadanego sygnału mowy. Oczywiście kluczowym problemem jest przy takim postawieniu sprawy wybór parametrów, użytych do opisu sygnału. Problem ten pojawiał się w książce i był dyskutowany już wcześniej, dlatego zostanie tu potraktowany skrótowo.

Najprostsze i najczęściej stosowane w praktyce są parametry widmowe. W nadajniku określa się widmo sygnału (oraz zazwyczaj dodatkowy parametr, sygnalizujący, czy rozważany fragment sygnału ma charakter dźwięczny, czy szumowy). W odbiorniku odtwarza się widmo sygnału za pomocą zestawu generatorów lub (częściej) zestawu filtrów o regulowanych charakterystykach, których działanie jest wymuszane przez generator tonu lub/i szumu o charakterystykach zbliżonych do naturalnego źródła tonu krtaniowego i szumu spółgłosek szumowych. Dawniej analiza widma przed przesłaniem go do urządzenia odbiorczego była dokonywana na drodze analogowej za pomocą zestawu filtrów pasmowych lub niekiedy jednego filtru heterodynowo przestrajanego w zadanym przedziale częstotliwości. Obecnie analogiczne wyniki uzyskuje się zazwyczaj na drodze cyfrowej, używając metod krótkookresowej analizy Fouriera oraz algorytmu FFT (por. p. 4.2 i 4.3). Technika nie ma jednak istotnego znaczenia, gdyż zasada działania pozostaje nie zmieniona. Ważne jest, że w widmie sygnału wydziela się pewną liczbę dyskretnych pasm częstotliwości, a następnie wydo-

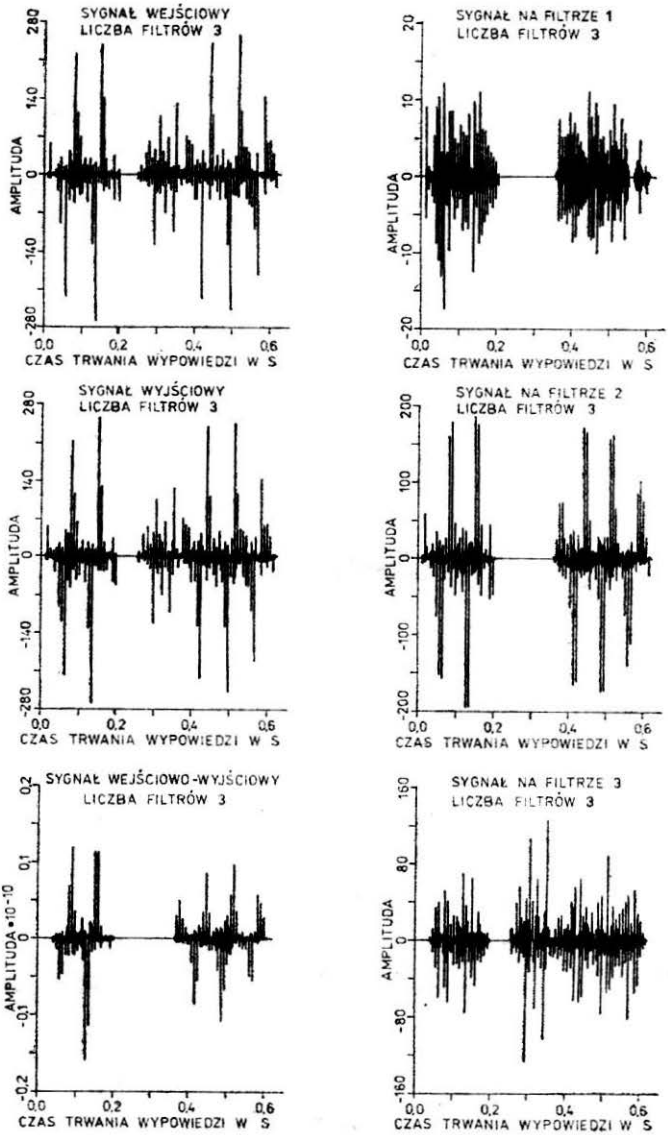
bywa (w ten lub inny sposób) wolnozmienną obwiednię amplitud sygnału w tych pasmach, którą to obwiednię, szeregowo lub równolegle, przesyła się do odbiornika (rys. 6-6). Ze względu na kluczową rolę, jaką odgrywa w omówionej metodzie podział sygnału na pasma częstotliwości, odpowiednie wokodery nazywa się p a s m o w y m i. Naturalnie im mniej pasm, tym oszczędniejszy wokoder, gdyż wymaga przesłania mniejszej liczby




6-6. Struktura nadajnika (po lewej) i odbiornika (po prawej) wokodera pasmowego. Nadajnik wydziela z ciągłego sygnału mowy parametry pobudzenia (ton lub szum, ewentualna częstotliwość) oraz określa za pomocą filtrów i demodulatorów niskoczęstotliwościową obwiednię widma sygnału. Informacje te, przesłane przez kanał, wykorzystywane są w odbiorniku do syntezy sygnału mowy. Sygnał z generatora pobudzenia, sterowanego parametrami wydzielonymi w odbiorniku, podawany jest na *n* filtrów o takich samych parametrach jak w nadajniku. Sygnały kanałowe sterują intensywnością tonu lub szumu z generatora, docierającego do odpowiedniego pasma nadajnika. Przez zsumowanie sygnałów wyjściowych z filtrów powstaje sygnał zrozumiały jako sygnał mowy

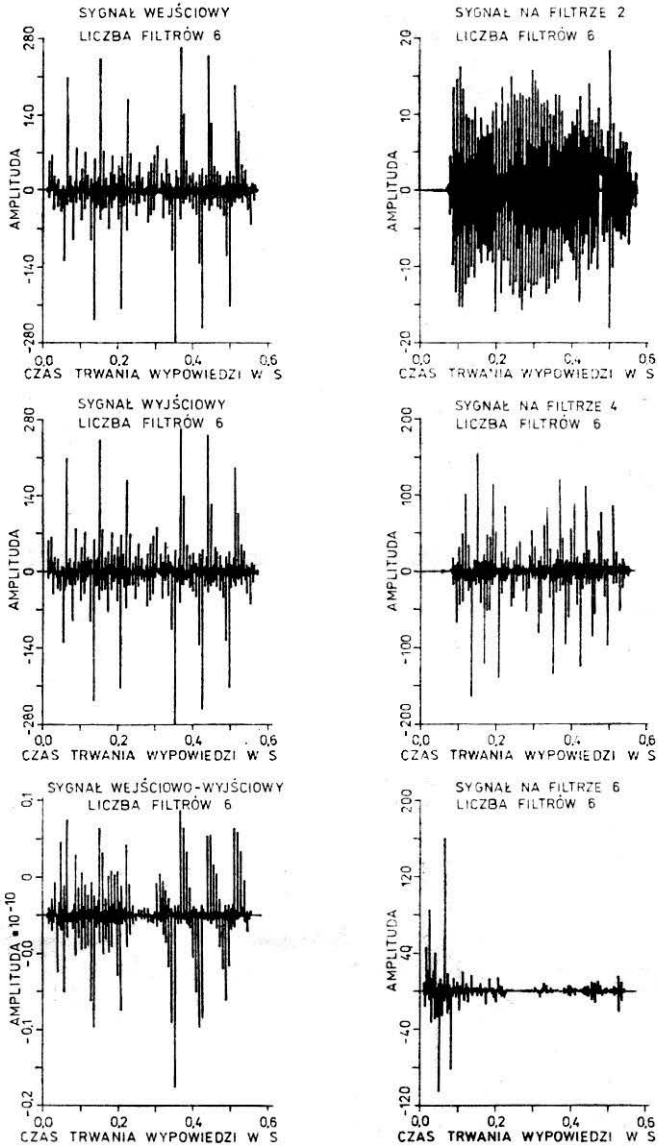
informacji, równocześnie jednak maleje dokładność odtworzenia sygnału na odbiorczym końcu łącza. Przykładowo na rysunkach 6-7, 6-8 i 6-9 przedstawiono przebiegi sygnałów w wokoderach pasmowych (symulowanych komputerowo) o odpowiednio trzech, sześciu i ośmiu pasmach. Pokazano przebieg sygnału wejściowego (lewy górny róg każdego rysunku), przebieg sygnału odtworzonego na wyjściowym łączu wokodera (poniżej), różnicę między sygnałem wejściowym i wyjściowym (na dole z lewej) oraz przykładowe trzy przebiegi w wybranych trzech pasmach wokodera (sygnały kanałowe) — po prawej stronie rysunku.

Wniosek z podobnych badań, a także z eksperymentów polegających na ocenianiu przez ludzi jakości mowy odtwarzanej przez wokoder, jest następujący: zrozumiałość i wyrazistość mowy polepsza się ustawicznie wraz ze wzrostem liczby pasm przesyłanych, jednak wzrost ten jest najszybszy przy małej liczbie filtrów, poczynając od około dziesięciu pasm dalszy przyrost jakości przesyłanej mowy wraz ze wzrostem liczby użytych pasm staje się na tyle wolny, że nie zawsze może być oceniony jako opłacalny. Z tego względu można przyjąć, że optymalna liczba pasm częstotliwości używanych w wokoderach pasmowych wynosi około dziesięciu, minimalna natomiast (taka, poniżej której jakość przesyłanej mowy jest niedopuszczalnie zła) — około pięciu. W użyciu praktycznym są jednak również wokodery zawierające kilkadziesiąt filtrów pasmowych, gdyż nawet przy tak dużej liczbie użytych pasm stosowanie wokodera jest opłacalne: suma sygnałów wszyst-




6-7. Wybrane przebiegi sygnałów w symulowanym komputerowo modelu wokodera pasmowego. Po lewej stronie rysunku — idąc od góry — pokazano kolejno: przebieg sygnału wejściowego (wypowiedź *praca*), przebieg sygnału wyjściowego (syntezowanego w odbiorniku) oraz różnicę między sygnałem wejściowym i wyjściowym — stanowiącą wiadomy obraz błędów transmisji. Po prawej stronie pokazano przebiegi sygnałów w wybranych kanałach wokodera. Przesyłaniu podlega oczywiście sygnał zdemodulowany, czyli obwiednia podanych po prawej stronie rysunku przebiegów. Prezentowane przebiegi odpowiadają wokoderowi trzypasmowemu

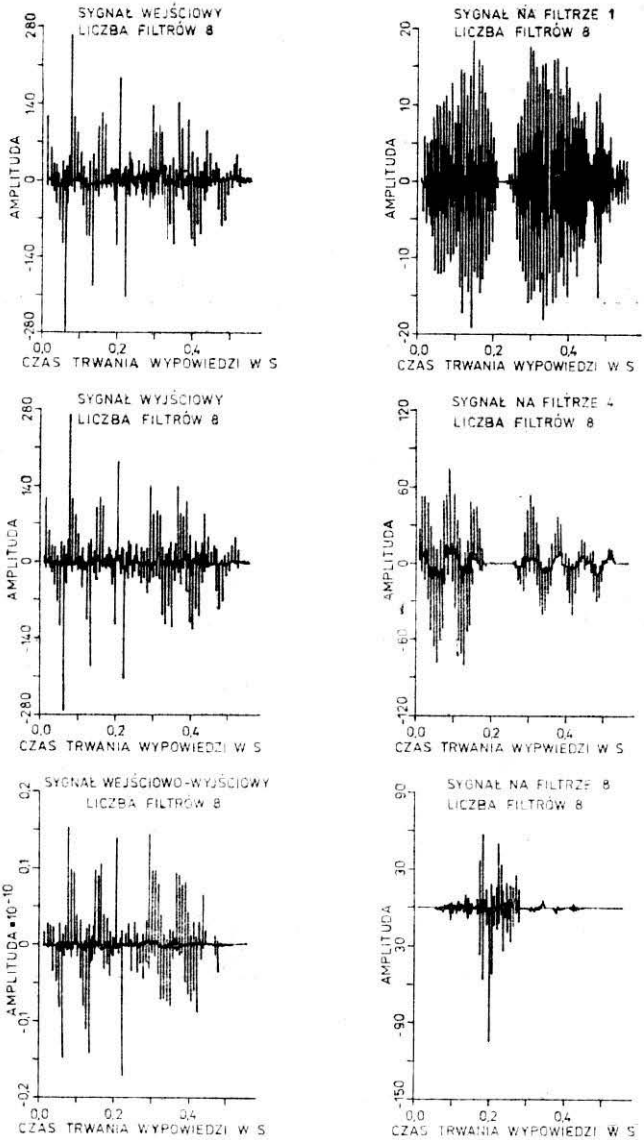
kich kanałów, powiększona o dodatkową informację na temat charakterystyki pobudzenia (ton krtaniowy/szum) zajmuje w linii telekomunikacyjnej znacznie mniej miejsca niż oryginalny przebieg sygnału mowy. Wynik ten jest osiągany głównie dlatego, że sygnały na wyjściach filtrów wokodera (rys. 6-10) zmieniają się wolno (częstotliwość graniczna sygnałów obwiedni jest na poziomie kilkunastu do trzydziestu herców maksymalnie), a także mają znacznie mniejszą dynamikę niż oryginalny sygnał. Przesyłanie wszystkich sygnałów kanałowych zajmuje więc znacznie węższe pasmo w linii niż sygnał oryginalny. Zatem stosując zarówno metody zwielokrotnienia w dziedzinie częstotliwości, jak i metody zwielokrotnienia w dziedzinie czasu — można w tym samym łączu teletransmisyjnym „zmieścić" znacznie




6-8. Wybrane przebiegi sygnałów w wokoderze pasmowym o sześciu pasmach. Struktura rysunku identyczna jak na rys. 6-7. Wypowiedź *cena*, głos męski

więcej sygnałów przetworzonych wokoderowo niż sygnałów w naturalnej formie.

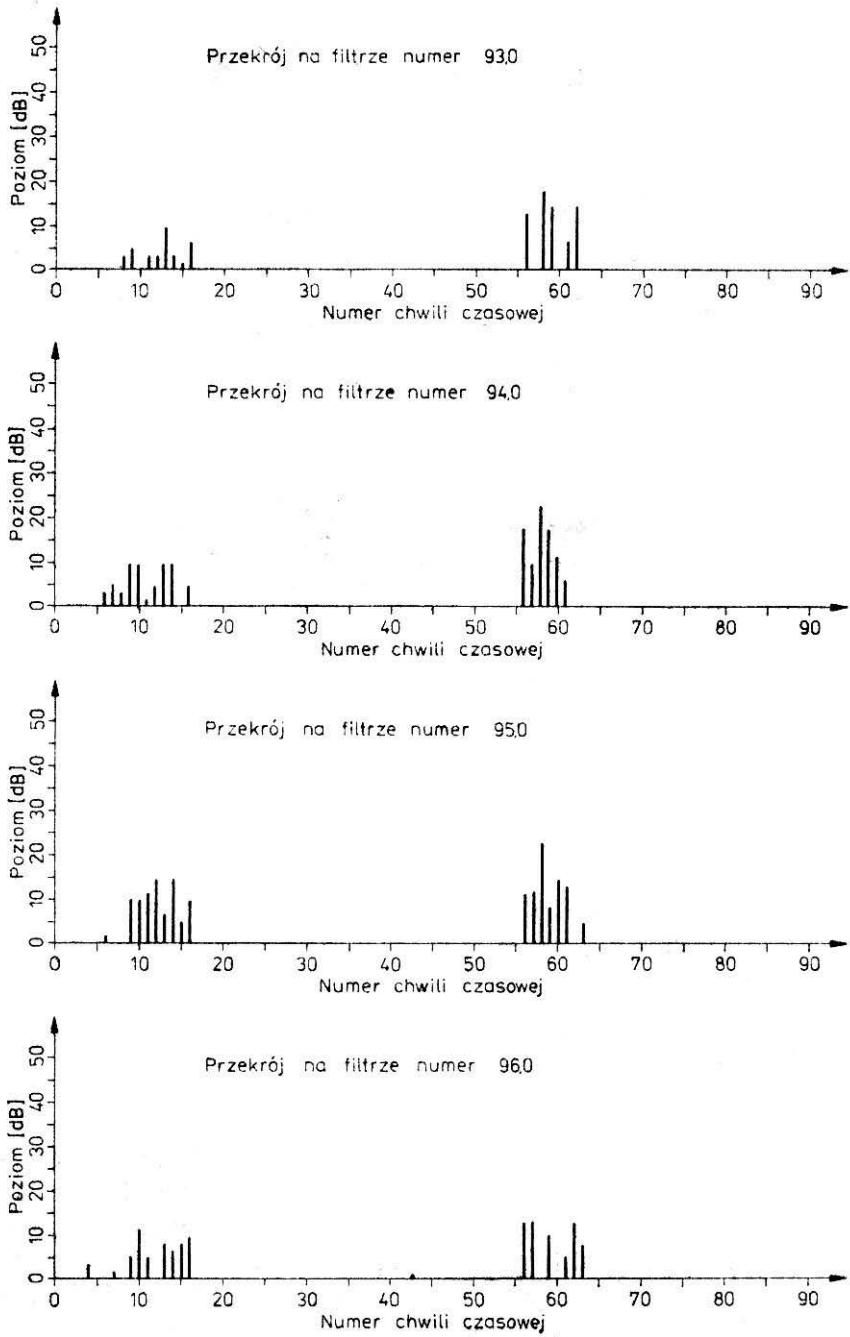
Zaletami wokodera pasmowego (rys. 6-6) są: prosta budowa oraz dobre powiązanie struktury wokodera z naturalnymi procesami artykulacji i percepcji mowy (por. rozdz. 2 i 3). Istotnie, urządzenie analizujące wokodera ma postać zestawu filtrów lub — w przypadku najnowocześniejszej, cyfrowej realizacji — ogranicza się do zastosowania algorytmu FFT i procedur uśredniających. Odbiornik ma również prostą budowę: generator tonu, generator szumu, zestaw filtrów o takich samych charakterystykach jak w nadajniku (często, przy łączności dwukierunkowej są to fizycznie te same filtry) oraz układy mnożące, wytwarzające sygnały w poszczególnych pas-




6-9. Wybrane przebiegi sygnałów w wokoderze pasmowym o ośmiu pasmach. Struktura identyczna jak na poprzednich rysunkach. Wypowiedź *Azor*, głos męski

mach. Bardziej wyrafinowane realizacje przewidują dodatkowe użycie analizatora zmian częstotliwości podstawowej (tonu krtaniowego $F_0$) i stosowną modulację funkcji generatora w nadajniku. Wbrew pozorom nie służy to jedynie zwiększeniu naturalności odbieranej mowy, która zachowuje dzięki temu elementy naturalnej intonacji i brzmi milej dla ucha, ale sprzyja to także większej zrozumiałości odbieranego sygnału, gdyż monotonny, pozbawiony elementów intonacyjnych sygnał jest rozumiany słabo, a zmęczeni słuchacze często popełniają błędy przy interpretacji nadawanych wypowiedzi.

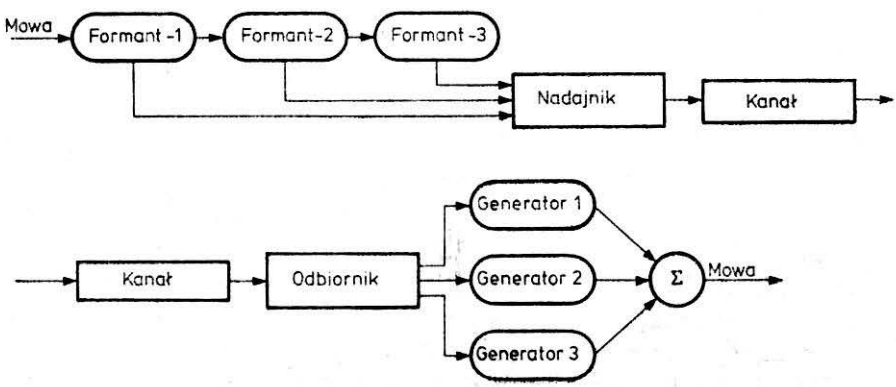
Ostatnim zagadnieniem, o którym warto wspomnieć w związku z budową wokodera pasmowego, jest problem sposobu rozmieszczenia wybranych




6-10. Przebiegi sygnałów kanałowych w wielokanałowym wokoderze pasmowym. Podany przykład dotyczy symulacji cyfrowej wokodera, stąd dyskretny charakter sygnałów kanałowych (wypowiedź *serce*, głos męski). Widoczna jest powolna zmienność sygnałów kanałowych, umożliwiająca ich oszczędne przesyłanie w kanale telekomunikacyjnym

pasm częstotliwości. Zagadnienie to było dyskutowane uprzednio w innym kontekście, a mianowicie z punktu widzenia wprowadzania sygnału mowy do maszyny celem rozpoznawania go i automatycznego rozumienia. W przypadku wokodera wnioski są jednak odmienne od uprzednio przytoczonych. Przy przesyłaniu sygnału mowy i przy tak małej liczbie wyróżnianych pasm, jak to typowo ma miejsce w odniesieniu do wokodera, celowe jest stosowanie logarytmicznej skali częstotliwości i rozmieszczanie częstotliwości środkowych poszczególnych pasm i ich szerokości zgodnie z regułą postępu geometrycznego. Niekiedy rekomendowana jest również tak zwana skala subiektywna (Königa), w której w zakresie do 1 kHz pasma są rozmieszczone równomiernie w skali liniowej (daje to korzystny efekt „zagęszczenia" analizy dolnych pasm częstotliwości, szczególnie istotnych dla rozpoznawania mowy), a powyżej wspomnianej częstotliwości — skala staje się logarytmiczna. Wszystkie wspomniane nierównomierności skali łatwo jest osiągnąć w przypadku stosowania zarówno metod analogowych, jak i cyfrowych (por. 4.2 i 4.3). Natomiast stosowanie algorytmu FFT zmusza do korzystania z dodatkowych programów „przeskalowujących" widmo, gdyż oryginalne widmo uzyskane na drodze obliczeniowej zawsze dane jest, w tym przypadku, w skali liniowej. W przypadku cyfrowej realizacji wokodera można natomiast osiągnąć stabilniejszą jego pracę, lepsze parametry filtrów kanałowych, funkcji okna czasowego, generatorów odtwarzających sygnał oraz pewniejsza jest (bardziej odporna na zakłócenia) transmisja sygnałów między nadajnikiem i odbiornikiem.

Obok wokodera pasmowego, który może być rozpatrywany jako konstrukcja o ustalonej renomie, ale już stosunkowo mało nowoczesna, pojawiły się liczne koncepcje innych wokoderów. Naturalnie na pierwszym miejscu pojawiają się tu konstrukcje działające z wykorzystaniem formantów. Koncepcja wokodera formantowego doczekała się dziesiątków skutecznie działających modeli laboratoryjnych i wydaje się nadal bardzo obiecująca, mimo „konkurencji" ze strony nowocześniejszych podejść, między innymi opartych na metodach liniowej predykcji. Strukturę wokodera formantowego przedstawiono na rys. 6-11. W podanym na rysunku schemacie nie wpro-

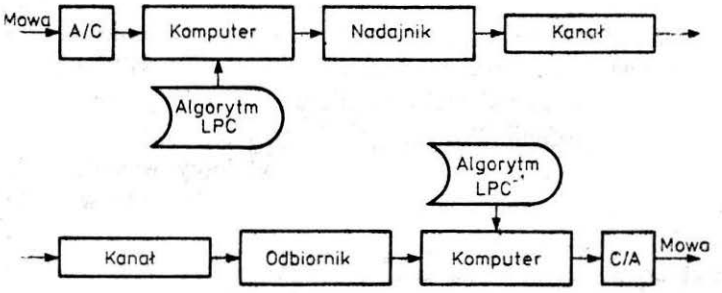



6-11. Struktura wokodera formantowego. U góry nadajnik, u dołu odbiornik. W nadajniku wydziela się trajektorie zmian czasowych formantów, a w odbiorniku, sterując odpowiednio generatorami i sumując ich sygnały, odtwarza się sygnał mowy

wadzono szczegółów procesu wydzielania częstości formantowych ani konstrukcji układu sterowanego parametrycznie przez informacje o częstotliwościach formantowych, a służącego do generacji sygnału wyjściowego w odbiorniku. Problem wydzielania formantów był bowiem dyskutowany w p. 4.4, zatem mimo dużej pomysłowości, jaką odznaczali się niektórzy twórcy wokoderów w budowie układów wydzielających formanty na drodze analogowej lub cyfrowej — trudno by było dodać tutaj coś naprawdę istotnie nowego. Jedyne, co można stwierdzić, to fakt większej, niż w przypadku systemów rozpoznawania mowy, tolerancyjności wokoderów na ewentualne niedokładności i błędy przy lokalizacji formantów. W szczególności problem ciągłości czasowej formantu, silnie podkreślany w dyskusji ukierunkowanej na rozpoznawanie mowy, staje się w wokoderze mniej ważny z tego powodu, że jeśli nawet część analizująca (nadajnik) wokodera wygeneruje (na skutek błędu) „skoki" w wartościach formantu, to zostaną one „wygładzone" w odbiorniku na skutek bezwładności przestrajanych parametrycznie układów syntezy mowy. Z drugiej strony dyskutowany przy syntezie mowy na podstawie wartości formantów (p. 2.3) problem stanów przejściowych między głoskami i trudności związanych z ich generacją — w wokoderach także nie istnieje. Przejścia od jednego do drugiego fonemu są bowiem śledzone bezpośrednio w nadajniku i odtwarzane w odbiorniku wiernie według oryginału — co zapewnia na ogół nie tylko dużą zrozumiałość mowy, ale także — w pewnym zakresie — odtwarza jej indywidualne cechy związane z głosem konkretnego mówcy. Innymi słowy, zbudować wokoder formantowy jest łatwiej niż osobno system analizy formantów i osobno system syntezy mowy na podstawie częstotliwości formantowych — co jednak nie znaczy, że konstrukcja wokodera według schematu z rys. 6-11 jest łatwa. Próby wykorzystania częstotliwości formantowych do kompresji sygnału mowy w kanale telekomunikacyjnym są jednak stale podejmowane, gdyż stopień kompresji (stosunek objętości sygnału po kompresji do objętości sygnału oryginalnego) jest w przypadku użycia wokodera formantowego bardzo duży, kilkakrotnie większy niż wyniki osiągane przy wokoderach kanałowych. W praktyce jakość sygnału odtwarzanego w wokoderach formantowych bywa również lepsza niż w wokoderach kanałowych, zatem gdyby nie wspomniane trudności z wydzielaniem formantów w nadajniku i ich wykorzystywaniem w odbiorniku, można by było zupełnie zarzucić technikę wokoderów pasmowych i interesować się jedynie wokoderami formantowymi — zanim oczywiście nie nadejdzie era wokoderów fonemowych, wzmiankowanych w poprzednim podrozdziale.

Do kompresji sygnału mowy w kanale telekomunikacyjnym można używać (lub próbować używać) wszystkich znanych parametrów mowy, istotne znaczenie ma jednak tylko jeszcze jedna koncepcja. W podrozdziale 4.5 omówiono technikę liniowej predykcji, wskazując, że jest to technika nowoczesna, związana z użyciem metod cyfrowych, a przy tym bardzo skutecznie opisująca sygnał mowy w kategoriach pewnych jego parametrów — możliwych do interpretacji jako współczynniki transmitancji toru głosowego w danym stadium procesu artykulacji. Można więc przyjąć, że wydobywa-

nie, przesyłanie i wykorzystywanie w odbiorniku parametrów funkcji liniowej predykcji stanowi kolejną możliwość konstrukcji wokodera (rys. 6-12). W istocie, koncepcja taka jest realizowana i daje bardzo dobre rezultaty. Z informacji, jakie napływają — bardzo skąpo zresztą — z laboratoriów najbardziej renomowanych firm produkujących sprzęt telekomunikacyjny, wynika, że istnieje już kilka udanych modeli wokodera liniowo-predykcyjnego i jest to aktualnie najbardziej obiecujący kierunek badań. Zastosowanie do transmisji sygnału mowy parametrów liniowej predykcji daje możli-




**6-12. Struktura wokodera opartego na zasadzie predykcji liniowej. W nadajniku komputer oblicza parametry liniowej predykcji sygnału. Po przesłaniu przez kanał, współczynniki predykcji liniowej służą do obliczeniowego (znowu potrzebny jest komputer) odtwarzania przebiegu sygnału mowy**

wość uzyskania bardzo dobrej zrozumiałości sygnału mowy odbieranej przy ekstremalnie małej zajętości przenoszącego transmisję kanału. Doniesienia firmy Siemens, cytowane na końcu książki, podają jako typowy dla wokodera liniowo-predykcyjnego strumień transmitowanej informacji rzędu $2 \cdot 10^3$ bit/s. Porównując to ze znanymi oszacowaniami objętości informacyjnej pełnego sygnału mowy (por. p. 4.6) dochodzimy do wniosku, że technika liniowej predykcji pozwala osiągać kilkusetkrotne „zagęszczenie" ilości przekazywanej informacji, co wyrażone w nieco innej formie pozwala sądzić, że stosując wokoder działający na omawianej zasadzie możemy tą samą siecią połączeń przekazywać kilkaset razy więcej rozmów telefonicznych. W istocie jest to wynik imponujący. Oczywiście efekty metod liniowej predykcji nie są osiągane za darmo — w stosunku do wszystkich wcześniej omawianych metod, technika liniowej predykcji stawia najwyższe wymagania aparaturze nadajnika i odbiornika sygnału. Obliczenia parametrów liniowej predykcji wymagają mocy obliczeniowej sporego komputera, jeśli mają być wyznaczane na bieżąco (w czasie rzeczywistym), co jest oczywiście wymogiem koniecznym dla wokodera. Podobnie wymagające są algorytmy odtwarzania sygnału na podstawie parametrów predykcyjnych — jest to zresztą obszar, w którym powstaje ostatnio najwięcej prac badawczych. Jeśli jednak utrzyma się dotychczasowy trend w mikroelektronice, jeśli koszt wykonywania obliczeń będzie sukcesywnie malał, a koszty przesyłania informacji — mimo zastosowania światłowodów — pozostaną duże, wówczas wokodery liniowo-predykcyjne mogą liczyć na upowszechnienie. Chyba że rozwiną się techniki rozpoznawania mowy — co już kilkakrotnie sugerowano.

Podsumowując trzeba stwierdzić, że w problemie kompresji sygnału mowy przy jego przesyłaniu kanałem telekomunikacyjnym nie powiedziano jeszcze ostatniego słowa. Powstają wciąż nowe opracowania, a żadne z nich nie zostało jeszcze powszechnie zastosowane w praktyce.

## 6.3. Wybrane problemy kryptofonii

W poprzednich rozdziałach książki skupiono uwagę głównie na zagadnieniach maksymalnie zrozumiałego przekazywania mowy. Obecnie, na zakończenie, przedstawione zostanie w telegraficznym skrócie kilka wybranych zagadnień z zakresu metod maksymalnie niezrozumiałego przekazywania mocy, czyli kryptofonii. Potrzeby przekazywania mowy w formie niezrozumiałej dla przypadkowego, postronnego odbiorcy wyłaniają się bardzo często nie tylko w zagadnieniach wojskowych, ale także w życiu gospodarczym, komunikacji pomiędzy firmami, a nawet osobami prywatnymi. Upowszechnienie użytkowania telefonii i (zwłaszcza) radiotelefonii spowodowało wzrost zainteresowania metodami utajniania mowy, gdyż inaczej te najdogodniejsze środki komunikacji międzyludzkiej stają się mało przydatne ze względu na niemożliwość wykorzystywania ich do przekazywania wiadomości w jakimkolwiek sensie i stopniu poufnej. Przechwytywanie i podsłuch rozmów telefonicznych stało się podstawowym źródłem pozyskiwania informacji nie tylko przez wojsko, służby specjalne i policję, ale przez wywiad gospodarczy, konkurencyjne firmy, czy wręcz przestępców poszukujących materiału do szantażu. W tej sytuacji zapotrzebowanie na urządzenia, dokonujące celowej i odwracalnej deformacji sygnału mowy przed jego przesłaniem w kanale telekomunikacyjnym, stale rośnie. O ileż wygodniej i swobodniej można rozmawiać, jeśli wiadomo, że sygnał zabezpieczony jest przed podsłuchem ze strony przypadkowego „hobbisty". Mówimy tu o zabezpieczeniu przed podsłuchem postronnej osoby nie dysponującej rozbudowanym laboratorium akustycznym i środkami odtwarzania mowy, gdyż w praktyce każda metoda utajniania mowy może zostać przy odpowiednim nakładzie pracy rozszyfrowana, a przesłany sygnał — odtworzony w jego oryginalnej, nie utajnionej postaci. Metody utajniania mają więc ten sam sens, jak zamki na drzwiach: mają zniechęcić przypadkowego złodzieja (w tym przypadku — podsłuchiwacza amatora) i opóźnić ewentualną akcję prawdziwego fachowca. Trzeba bowiem być świadomym, że przy dzisiejszych, bardzo efektywnych metodach analizy sygnału, angażowanych komputerach i wiedzy na temat kryptografii, kryptofonii i metodach deszyfracji kodów — żaden system utajniania nie jest stuprocentowo pewny. Przeciwnie, można być pewnym, że każdy szyfr, kod, technika utajniania, maskowania i zniekształcania zostanie prędzej czy później rozszyfrowana, a jedynym czynnikiem, na jaki można mieć wpływ stosując rozliczne zabezpieczenia — to czas, jaki będzie potrzebny zespołowi łamiącemu szyfr. Dąży się więc do tego, aby czas ważności i aktualności przesyłanej wiadomości był mniejszy od czasu niezbędnego do jej odszyfrowania przez osobę

nieupoważnioną (nie znającą klucza, według którego dokonano maskowania sygnału).

Metody utajniania sygnału mowy podzielić można ze względu na postać, w jakiej występuje sygnał, na cyfrowe i analogowe. Efektywniejsze i bogatsze w możliwości są metody cyfrowe, gdyż w przypadku kiedy sygnał jest w postaci serii dyskretnych kodów (PCM, Delta lub dowolnych innych), wówczas do jego utajnienia można by użyć wszelkich, bardzo rozbudowanych, łatwo dostępnych i doskonale wszechstronnie poznanych metod kodowania, szyfrowania i maskowania danych alfanumerycznych — gdyż czym w końcu różni się ciąg symboli kodowych przenoszących sygnał mowy od ciągu symboli kodowych, przekazujących tekst pisany? Technika komputerowa, mikroprocesory, elektroniczne maszyny szyfrujące i deszyfrujące — wszystko to może być użyte do maskowania treści zawartych w sygnale mowy, a „dawkowanie trudności" szyfru może być tu szczegółowo i precyzyjnie odmierzane. Innymi słowy, upowszechnienie cyfrowej transmisji sygnału, wprowadzenie cyfrowej telefonii a także upowszechnienie komputerów w telekomunikacji będzie sprzyjać i ułatwiać szyfrowanie i utajnianie sygnału mowy. Techniki takiego szyfrowania są przedmiotem dyskusji w specjalistycznej literaturze i wykraczają daleko poza problematykę, którą można wiązać z hasłem analizy sygnału mowy.

Specyficzne i bardzo ciekawe problemy wynikają natomiast przy próbach maskowania sygnału mowy traktowanego jako sygnał analogowy. Maskowanie polega w tym przypadku na celowym i odwracalnym niszczeniu struktury czasowej lub/i częstotliwościowej sygnału, w ten sposób, aby generalnie upodobnić przesyłany sygnał do szumu białego. Zasada działania urządzeń kryptofonicznych (tak zwanych skramblerów) polega więc na przykład na następujących zabiegach (stosowanych oddzielnie lub łącznie w różnych kombinacjach):

— zmianie struktury widma: przestawienie pasma, odwrócenie widma, przesyłanie poszczególnych pasm oddzielnie i montowanie ich w odbiorniku,

— zmianie struktury czasowej sygnału: przestawianie kolejności fragmentów czasowych sygnału, okresowo zmienna inwersja fazy sygnału, zmiana proporcji czasowych (iloczasów) poszczególnych głosek,

— zmianie struktury amplitudowej sygnału: spłaszczenie dynamiki sygnału, mowy, wypełnianie przerw szumem, nieregularne, zmienne w czasie wzmocnienie sygnału, modulacja obwiedni czasowo-amplitudowej sygnału.

Efektywne maskowanie i utajnianie sygnału mowy napotyka duże trudności ze względu na nadmiarowość sygnału mowy, a także z powodu nad wyraz efektywnego rozpoznawania, nawet bardzo zniekształconego sygnału mowy. Okazuje się, że nawet zmieniając położenie na osi czasu elementów mowy uzyskuje się w wielu przypadkach sygnał, który jedynie przy pierwszym czytaniu robi wrażenie całkowicie losowego, niezrozumiałego bełkotu. Możliwość wychwycenia słuchem i zinterpretowania takich elementów sygnału, jak częstość tonu krtaniowego, częstości formantów i ich zmiany, rytm wypowiedzi, zachowany dzięki dużym różnicom amplitudy elementów

samogłoskowych i spółgłoskowych wpływają na możliwość bezpośredniego odgadnięcia treści zamaskowanej wypowiedzi — szczególnie w tych przypadkach, kiedy zbiór możliwych komunikatów jest znany lub może być odtworzony. Jeśli nawet odsłuchowe rozpoznanie przemieszanego czasowo sygnału ze skramblera jest niemożliwe, wówczas stosunkowo prosta analiza sygnału pozwala odczytać regułę maskowania i odtworzyć sygnał w wersji oryginalnej.

Lepsze wyniki dają na ogół metody widmowe. Tu już nawet najprostszy z możliwych zabieg inwersji widma (odwrócenie widma sygnału w ten sposób, aby obszar dużych częstotliwości przypadał na obszar — w zamaskowanym sygnale — częstości małych i na odwrót) bardzo skutecznie utrudnia zrozumienie wypowiedzi. Tymczasem realizacja takiej inwersji jest technicznie niesłychanie prosta: wystarczy dokonać modulacji sygnału (amplitudowej) i brać pod uwagę odpowiednio przesuniętą wstęgę boczną.

Jeszcze skuteczniejsze są metody wokoderowe. Mowa podzielona zostaje na pasma, a następnie sygnał jest odtwarzany przy zmienionej numeracji pasm. Zrozumiałość tak spreparowanej mowy jest niewielka, a trudności dla potencjalnego nie uprawnionego odbiorcy są bardzo duże. Trzeba bowiem odgadnąć, jakie były oryginalne pasma, jak je pomieszano i jakie są reguły rozkładu szerokości i częstości środkowych pasm, które w dodatku z reguły zmienia się co jakiś czas podczas trwania transmisji. Oczywiście dla zachowania efektywności porozumiewania się nadawcy z upoważnionym (właściwym) odbiorcą sygnału mowy, ten ostatni musi dysponować informacją na temat sposobu zaszyfrowania sygnału i sprawną aparaturą na bieżąco deszyfrującą sygnał. Naturalnie pojawia się przy tym problem odpowiedniego zabezpieczenia zarówno klucza (informacji o metodzie szyfrowania), jak i deszyfrującej aparatury.

Podsumowując ten krótki podrozdział trzeba powiedzieć, że wiedza na temat sygnału mowy może służyć zarówno jej sprawnemu i maksymalnie zrozumiałemu przekazywaniu, jak i może być użyta do uczynienia transmisji mowy całkowicie niezrozumiałą. Obszerniejsze omówienie problematyki utajniania i szyfrowania mowy można znaleźć w specjalistycznych publikacjach, zebranych w wykazie literatury na końcu książki. Celem przedstawionego podrozdziału było jedynie zasygnalizowanie problemu, wskazanie na możliwości i zachęcenie do ewentualnych prac i studiów w tej dziedzinie, gdyż — co trzeba raz jeszcze podkreślić — powszechność telefonii i radiotelefonii rozmównej spowoduje już wkrótce wzrost zainteresowania możliwościami zabezpieczenia rozmowy przed podsłuchem. Tym samym problematyka, uprawiana dotychczas w ośrodkach wojskowych i siłą rzeczy mało znana szerszemu ogółowi, będzie mogła znaleźć się w zakresie zainteresowania niemal wszystkich laboratoriów zajmujących się problematyką sygnału mowy.

# Zakończenie

Prezentowana książka nie wyczerpała wszystkich zagadnień wiążących się z problemem sygnału mowy. Ale też — co trzeba podkreślić — nie pojawiła się ona na „bibliograficznej pustyni". Zagadnienia sygnału mowy, jego analizy, rozpoznawania, przesyłania, syntezy i wykorzystania były i są tematem wielu prac. Książkę tę napisano z myślą o uzupełnieniu istniejącego obrazu, o dodaniu informacji tam, gdzie jest ich dostępnych niewiele, a wstrzymaniu się od powtarzania zagadnień powszechnie znanych, dobrze opracowanych i wielokrotnie opisanych. Dlatego wiele razy odwoływano się w tekście książki do literatury, której wykaz zamieszczono na dalszych stronach, dlatego dobierano materiał poszczególnych rozdziałów i wybierano sposób jego prezentacji mając na uwadze istniejące i przygotowywane prace innych autorów, dlatego wreszcie dokonano obszernych studiów literaturowych w celu znalezienia takiej formuły książki, która nie dublując innych pozycji może dostarczyć sumę niezbędnych podstawowych informacji i może wnieść niektóre nowe wiadomości — odmienne od przedstawianych w pozostałych pracach.

Na zakończenie wypada jednak wskazać te pozycje literatury, które w największym stopniu zaważyły na koncepcji książki. Zaczynając od pracy dra Czesława Basztury (*Źródła, sygnały i obrazy akustyczne. Przetwarzanie, analiza, rozpoznawanie*), która zapewne ukaże się niemal równocześnie

z tą książką, a której istnienie „zwalniało" Autora od konieczności pisania o sprawach dobrze rozpracowanych tamże (sygnał mowy w aspektach telekomunikacyjnych, technika liniowej predykcji, analiza cepstralna i wiele innych — potraktowanych tu skrótowo zagadnieniach). Przy okazji Autor składa drowi Baszturze podziękowania za umożliwienie zapoznania się z powstającą książką, dzięki czemu możliwe było uniknięcie powtórzeń. Wiele informacji na temat zagadnień nie rozwiniętych w tej książce znaleźć można w pracy zbiorowej, której redaktorem był prof. Janusz Kacprowski: *Akustyka mowy i diagnostyka akustyczna*, Warszawa 1980, wyniki zaś najnowszych badań nad sygnałem mowy polskiej zbierane są okresowo w wydawanych przez PWN, pod redakcją prof. Wiktora Jassema, pracach zbiorowych zatytułowanych *Speech analysis and synthesis*. Ostatni, piąty tom tego wydawnictwa ukazał się w roku 1980. Z książek nieco starszych wymienić trzeba podstawowe dla wszystkich zajmujących się mową dzieło prof. Wiktora Jassema *Podstawy fonetyki akustycznej* wydane przez PWN w 1973 roku. Istnienie tej książki zwalniało Autora — w jego mniemaniu — od konieczności obszerniejszego dyskutowania fonetycznych aspektów mowy, na przykład stosunku elementów mowy żywej (na przykład fonemów) do stosowanego ortograficznego zapisu. Kolejna i ostatnia już wymieniana tu książka, to wydana bardzo dawno, bo aż w 1966 roku książka M. A. Sapożkowa *Sygnał mowy w telekomunikacji i cybernetyce*. Książka stanowiła w swoim czasie prawdziwą encyklopedię wiedzy o sygnale mowy, układach jego formowania, zapisywania, transmisji, ograniczania w objętości informacyjnej i rozpoznawania. Wprawdzie technika poszła ogromnie naprzód i dla współczesnego elektronika schematy gęsto upakowane lampami są jaskrawym anachronizmem, jednak sygnał mowy nie zmienił się od tamtych czasów, a nasza wiedza o nim — wbrew pozorom — nie wzbogaciła się aż tak bardzo. Z tego względu zagadnienia obszernie dyskutowane w tej książce: struktura sygnału mowy, jego elementy, zasady oceny jego jakości itd., mogły być tu potraktowane skrótowo.

Mimo wspomnianych skrótów książka jest nadspodziewanie obszerna. Bardzo wiele można bowiem napisać i powiedzieć na temat tak prostego i elementarnego na pozór obiektu — sygnału mowy. A przecież dla każdego człowieka są to sprawy oczywiste — wystarczy powiedzieć, usłyszeć, zrozumieć, zatelefonować... Dopiero kiedy zamiast nas wytwarzają mowę lub mają ją rozpoznawać komputery — uświadamiamy sobie złożoność tego procesu i nikłość naszej wiedzy w stosunku do rozmiarów problemu. W ten sposób — nie po raz pierwszy i nie po raz ostatni — te mądre maszyny pomagają nam lepiej zrozumieć i poznać nas samych. Zrozumieć i zadumać się nad doskonałością tworów Natury, które przed tysiącleciami wytworzyły system artykulacji dostosowany do wcześniej perfekcyjnie stworzonego słuchu i jeszcze wcześniej uformowanego mózgu, dzięki czemu ludzie uzyskali najważniejsze narzędzie rodzącej się cywilizacji — **sygnał mowy**.

## Sygnał mowy a tekst pisany

Opisując w tekście książki — szczególnie w tabelach, na rysunkach oraz przy prezentacji przykładów — określone zjawiska zachodzące w sygnale mowy, napotykano na trudności związane z różnicami, jakie zachodzą pomiędzy pisownią a wymową tych samych wyrazów, fraz i zdań. Wymowa podlega własnym prawom, wynikającym z anatomii i fizjologii narządów mowy, a także z uwarunkowań natury kulturowej, tradycji i regionalnych obyczajów. Natomiast pisownia jest skodyfikowana przez ortografię i w znacznym stopniu odbiega od rzeczywistej wymowy. Rozbieżności te są wielokierunkowe: często ta sama litera używana jest do zapisu zupełnie różnych brzmieniowo głosek (na przykład w wyrazie *babka* pierwsze i drugie *b* oznacza odmienny dźwięk), innym zaś razem tę samą głoskę rejestruje się pisząc — zależnie od tradycji — odmienne litery (by wspomnieć tylko o dwoistości *u* oraz *ó* w języku polskim). Dla zapisu jednej głoski można używać więcej niż jednej litery (*rz*, *sz*, *cz*, *dz*, *dzi* — by wymienić tylko niektóre typowe dla naszego języka dźwięki), natomiast często także używa się jednej litery dla zapisu dwu kolejnych fonemów — przykładowo *ą* odpowiada w wymowie sekwencji fonemów *om*, a głoski odpowiadającej literze *c* w ogóle nie ma — gdyż wymawiane jest zawsze *ts*. Podobnych przykładów można mnożyć bez liku, a ich wspólnym mianownikiem jest postawiona na wstępie teza: pomiędzy językiem mówionym i językiem pisanym jest trudna do przebycia przepaść, wyjątkowo dobrze znana tym wszystkim, którzy w trudzie opanowują wymowę nieznanego języka na podstawie drukowanych podręczników.

Chcąc więc opisywać — tak jak w tej książce — sygnał mowy jako głosową formę języka, trzeba koniecznie posłużyć się jednoznacznym i powszechnie przyjętym systemem notacji, rejestrującym brzmienie poszczególnych wyrazów i głosek w sposób niezależny od ich tradycyjnej pisowni. System taki jest znany i używany dla notacji zjawisk dźwiękowych we wszystkich językach świata. Wykorzystuje specjalne symbole międzynarodowej transkrypcji fonematycznej, które — dokładnie stosowane i precyzyjnie określane — pozwalają odwzorowywać zjawiska zachodzące podczas mówienia w sposób równie wierny i szczegółowy, jak rejestracja na taśmie magnetofonowej. Symbole transkrypcji szczegółowej są niewygodne w użyciu, gdyż niemal wszystkie odbiegają od typowych czcionek używanych w drukarniach, co sprawia kłopoty poligraficzne, a w dodatku dla wiernego odwzorowania zjawisk zachodzących podczas mówienia opatrywane są licznymi dodatkowymi symbolami, sygnalizującymi między innymi stopień otwarcia lub przymknięcia ust, położenie języka lub artykulację nosową. Zainteresowanych szczegółową transkrypcją (nie tylko zresztą głosek polskich) odesłać więc należy do książki profesora Wiktora Jassema *Podstawy fonetyki akustycznej*, w której wszystkie te subtelności obszernie wyjaśniono. Dla potrzeb tej książki przyjęto transkrypcję uproszczoną, łatwiejszą w za-

pisie i prostszą w stosowaniu. Jej zasady przedstawione będą dalej wraz z krótką charakterystyką głosek języka polskiego.

## Samogłoski

Powstają podczas swobodnego przepływu powietrza wzdłuż linii środkowej języka i są wszystkie bez wyjątku dźwięczne. W transkrypcji szczegółowej wyróżnia się (dla różnych języków) łącznie kilkadziesiąt samogłosek, natomiast dla języka polskiego celowe jest wyróżnienie sześciu samogłosek: **i** (lis), **ɨ** (pył), **e** (szewc), **a** (rak), **o** (rok), **u** (mur). Warto zwrócić uwagę, że głoska zapisywana ortograficznie jako Y w transkrypcji ma zapis **ɨ** (i przekreślone). Wynika to z faktu, że w międzynarodowej transkrypcji znak **y** zarezerwowano dla głosek brzmiących jak w niemieckim słowie *süß* lub francuskim *lutte*, **Y** zaś to dźwięk występujący w niemieckim küssen. Należy też odnotować różnicę w podanej liście głosek, opartej na akustycznej analizie zjawisk zachodzących podczas artykulacji mowy, w stosunku do „szkolnej" listy polskich samogłosek. Uwzględnia się w niej nosowe **ę** i **ą**. Głoski te pominięto w podanej liście, ponieważ nie istnieją. Zjawisko zapisywane ortograficznie jako ę lub ą jest zawsze dwugłoską złożoną z e lub o (odpowiednio) i którejś z głosek nosowych (zależnie od kontekstu) — najczęściej jest to ŋ (patrz dalej). Wiadomość ta jest zapewne dla wielu Czytelników sprzeczna z ich subiektywnymi odczuciami, ale analiza spektrogramów sygnału mowy nie pozostawia w tej sprawie cienia wątpliwości.

## Spółgłoski zwarte

Powstają podczas chwilowego całkowitego zatrzymania przepływu powietrza z płuc, po którym następuje plozja — wybuchowy wypływ powietrza połączony z charakterystycznym dźwiękiem. Zależnie od tego, czy podczas zwarcia i plozji struny głosowe drgają, czy nie — mamy do czynienia z odmianą dźwięczną lub bezdźwięczną danej głoski. Zależnie od miejsca zwarcia wyróżnić można głoski zwarte wargowe — **p** (pas) i **b** (bas), zębowe — **t** (tom) i **d** (dom), podniebienne — **c** (kino) i **ɟ** (ginąć) i tylnojęzykowe — **k** (kura) i **g** (góra). Przy dokładnej transkrypcji wyróżnia się jeszcze odmiany głosek t i d: dziąsłową i cerebralną, zaznaczane oddzielnymi symbolami ze względu na ich odmienność akustyczną. Warto także zwrócić uwagę na rozróżnienie c i k oraz **ɟ** i **g** — nie występuje w piśmie, ale konieczne przy analizie sygnału mowy.

## Spółgłoski trące

Artykulacja tych głosek polega na wywołaniu turbulencji powietrza wypływającego z płuc w miejscu celowo utworzonego przewężenia w narządach mowy. Głoski te, podobnie jak wcześniej omówione, mają odmiany dźwięczne i bezdźwięczne, a ich klasyfikacja jest oparta na miejscu utworzenia szczeliny, przy czym nie wszystkie możliwe lokalizacje są wykorzystywane

w języku polskim — na przykład brak w nim spółgłosek trących dwuwargowych (jak na przykład w japońskim wyrazie **fudżi**), zębowych (słynne angielskie **the**) lub środkowojęzykowych (jak w niemieckim **ich**). Ponadto transkrypcja dokładna wyróżnia tu wiele subtelnych różnic, możliwych do pominięcia przy prezentacji transkrypcji przybliżonej, stosowanej w tej książce. Oto w języku polskim występują spółgłoski trące zębowo-wargowe — **f** (frak) i **v** (wrak), zazębowe — **s** (kosa) i **z** (koza), zadziąsłowe — **ʃ** (szary) i **ʒ** (żar), dziąsłowo-środkowojęzykowe — **ɕ** (siano) i **ʑ** (ziarno) oraz tylnojęzykowe — **x** (niech). Ta ostatnia głoska praktycznie nie ma w języku polskim dźwięcznego odpowiednika, chociaż przy szczególnie starannej wymowie frazy **niech będzie** można zauważyć występowanie udźwięcznionego ch (co fonetycznie zapisuje się symbolem **ɣ**).

## Spółgłoski zwarto-trące

W głoskach tych występuje z reguły para elementów — głoska zwarta i segment odpowiadający głosce trącej, ale o krótszym czasie trwania. Oba elementy mają to samo miejsce artykulacji i łączą się w charakterystyczną całość, co przesądza o traktowaniu ich jako odrębnych głosek. Głoski te, nie występujące w ogóle w wielu językach (na przykład we francuskim) najobficiej występują w języku polskim, który ma ich aż 6. Ich podział wynika z miejsca artykulacji, zatem wyróżnia się: zazębowe-**t͡s** (praca) i **d͡z** (sadza), dziąsłowe — **t͡ʃ** (czytać) i **d͡ʒ** (drożdże), dziąsłowo-środkowojęzykowe — **t͡ɕ** (ciało) i **d͡ʑ** (działo).

## Spółgłoski nosowe

Przy artykulacji głosek nosowych opuszczony języczek podniebienia miękkiego udostępnia dla emisji głosu jamę nosową, podczas gdy jama ustna, zamknięta w punkcie zależnym od rodzaju artykułowanej głoski, stanowi „bocznik akustyczny". Klasyfikacja głosek nosowych zależna jest od punktu zamknięcia jamy ustnej, w związku z czym wyróżnia się głoski: dwuwargowe — **m** (matka), dziąsłowe — **n** (nora), środkowojęzykowe — **ɲ** (koń) i tylnojęzykowe — **ŋ** (bank). Spółgłoski nosowe praktycznie nie występują w formie bezdźwięcznej, natomiast uczestniczą w formowaniu innych głosek nazalizowanych — na przykład **ę**. Przy artykulacji tej ostatniej głoski występuje zawsze **e**, po którym — zależnie od kontekstu — może występować głoska **n** (tęcza), **ɲ** (miękki) lub **ŋ** (ręka).

## Spółgłoski boczne

Przy artykulacji głosek bocznych powietrze uchodzi obok języka — po jednej stronie lub obustronnie. W języku polskim jest w zasadzie jedna głoska omawianego typu, mianowicie **l** (lody). Głoska **ɫ** (łydka) zanika na rzecz głoski płynnej **w** (dłoń) i dostrzegana jest jedynie w bardzo starannej wymowie scenicznej.

## Samogłoski niesylabiczne

Głoski te pod względem artykulacyjnym przypominają samogłoski, natomiast ich funkcje są identyczne ze spółgłoskami. W języku polskim są dwie takie głoski: płynne **j** (**j**odła) i zaokrąglone **w** (**ł**awka). Warto zwrócić uwagę na transkrypcję głoski ortograficznie zapisywanej jako *ł*. Międzynarodowa transkrypcja przypisuje jej symbol *w* zgodnie z wymową angielską, natomiast symbol *l* w transkrypcji oznacza inny dźwięk („kresowe" Ł).

## Spółgłoska drżąca

Wyjątkowo nieregularny obraz ma głoska **r** (**r**yba). Podczas jej artykulacji język uderza o podniebienie, tworząc wyjątkowo nieregularny zespół elementów akustycznych: szumy sąsiadują tu z odcinkami periodycznego przebiegu sygnału, impulsy obok formantów oraz okresy przerw.

# Literatura

Wykaz literatury podzielono na części, zgodnie ze strukturą rozdziałów i podrozdziałów książki. Ma to tę zaletę, że Czytelnik odesłany do wykazu z konkretnej partii tekstu może łatwo zidentyfikować, o jakie pozycje literatury chodzi. Ma to jednak również wadę, polegającą na braku uporządkowania podawanych pozycji bibliograficznych, a także dlatego, że niektóre, szczególnie bogate w treść pozycje literatury mogą być wykazane kilkakrotnie, w różnych kontekstach i w różnych zestawieniach. W sumie wydaje się jednak, że zalety proponowanego układu przeważają nad jego wadami.

Generalne uwagi na temat wytwarzania mowy znaleźć można w pracach:

Anderson S. R.: *Wprowadzenie do fonologii.* Ossolineum, Wrocław—Warszawa—Kraków—Gdańsk—Łódź 1982.
Dłuska M.: *Fonetyka polska.* PWN, Warszawa—Kraków 1983.
Fant G.: *Acoustic theory of speech production.* S' — Gravenhage, Mouton and Co., 1960.
Flanagan J. L.: *Speech analysis, synthesis and perception.* Springer Verlag, Berlin, Heidelber, New York 1970.
Kacprowski J.: *Akustyczne modelowanie organu mowy.* Prace IPPT PAN, Warszawa 1982.
Sapożkow M. A.: *Sygnał mowy w telekomunikacji i cybernetyce.* WNT, Warszawa 1966.

Uzupełnienia rozważań z rozdziału 2.1 można znaleźć w pracach:

Barney H. I., Dunn H. K.: *Speech Synthesis w Manual of Phonetics.* (L. Kaiser — editor), Amsterdam 1957.
Kacprowski J.: *Teoretyczne podstawy syntezy samogłosek polskich w rezonansowych układach formantowych.* Rozprawy Elektrotechniczne, VIII, 1962, str. 127÷203.

Lawrence W.: *The synthesis of speech from signals which have a low information rate.* Communication theory (W. Jacson — editor), London 1953.
Wierzchowska B.: *Fonetyka i fonologia języka polskiego*, Ossolineum, Wrocław 1980.

Biologiczne aspekty procesu artykulacji mowy rozszerzyć można na podstawie podręczników:

Best Ch. H., Taylor N. B.: *Żywy organizm — zarys fizjologii człowieka.* PZWL, Warszawa 1966.
Bochenek A., Reicher M.: *Anatomia człowieka.* Tomy I÷VIII, PZWL, Warszawa 1966—1982.
Miętkiewski E.: *Zarys fizjologii lekarskiej.* PZWL, Warszawa 1977.
Traczyk W. Z.: *Fizjologia człowieka w zarysie.* PZWL, Warszawa 1982.
Sylwanowicz W.: *Anatomia człowieka.* PZWL, Warszawa 1978.

Opis funkcjonowania traktu głosowego znaleźć można w pracach:

Fant G. C. M.: *Speech sounds nad features.* The MIT Press, Cambridge, Mass., 1972.
Flanagan J. L.: *Source — system interactions in the vocal tract.* Ann. N. Y. Avada. Sci., vol 155, 1968, pp. 9÷15.
Flanagan J. L.: *Voices of man and machines.* J. Acoust. Soc. Amer., vol. 51, 1972, pp. 1375÷1387.
Oppenheim A. V.: *Sygnały cyfrowe — przetwarzanie i zastosowanie.* WNT, Warszawa 1982 (rozdział 3).

Opis różnych prób modelowania systemu artykulacyjnego oraz generatorów pobudzenia akustycznego znaleźć można w pracach:

Coker C. H.: *A model of articulatory dynamics and control.* Proc. of the IEEE, vol. 64, nr 4, 1976, pp. 452÷460.
Dennis J. B.: *Computer control of an analog vocal tract.* Proc. Stockholm Speech Comm. Seminar, R. I. T., Stockholm 1962.
Dunn H. K.: *The calculation of vovel resonances and an electrical vocal tract.* J. Acoust. Soc. Amer., vol. 22, 1950, pp. 740÷753.
Flanagan J. L.: *Some properties of the glottal sound source.* Journ. Speech and Hearing Research, vol. 1, 1968, pp. 99÷116.
Hecker M. H. L.: *Studies of nasal consonants with an articulatory speech synthezier.* Journ. Acoust. Soc. Amer., vol. 34, 1962, pp. 179÷188.
House A. S.: *Analog studies of nasal consonants.* Journ. Speech and Hearing Disordes vol. 22, 1957, pp. 190÷204.
House A. S., Stevens K. N.: *Analog studies of the nasalization of vovels.* Journ. Speech and Hearing Disorders, vol. 21, 1956, pp. 218÷232.
Ishizaka K., Flanagan J. L.: *Synthesis of voiced sounds from a two-mass modell of the vocal cords.* Bell Syst. Techn. J., vol. 51, 1972, pp. 1233÷1268.
Ishizaka K., French J. C., Flanagan J. L.: *Direct determination of vocal tract wall impedance.* Journ. Acoust. Soc. Amer., vol. 55, 1974.
Kacprowski J.: *Fizyczne modele źródła krtaniowego.* Archiwum Akustyki, vol. 12, l. 1977, str. 47÷70.
Kacprowski J.: *Model symulacyjny kanału głosowego z uwzględnieniem zjawiska nazalizacji.* Archiwum Akustyki, vol. 12, 4, 1977, str. 181÷302.
Kacprowski J.: *Teoretyczne podstawy syntezy samogłosek polskich w rezonansowych układach formantowych.* Rozprawy Elektrotechniczne, vol. 8, 1, 1962, str. 127÷203.
Rabiner L. R.: *Digital — formant synthesier for speech synthesis.* Journ. Acoust. Soc. Amer., vol. 43, 1968, pp. 822÷828.

J. Q. Steward: *An electrical analogue of the vocal organs.* Nature, vol. 110, 1922, pp. 311 ÷ ÷312.
Stevens K. N., Fant G. C. M.: *An electrical analog of vocal tract.* Journ. Acoust., Soc., Amer., vol. 25, 1953, pp. 734÷742.
Tarnoczy T.: *Opening Time and Opening Quotient of the Vocal Cords.*, Journ. Acoust. Soc. Amer., vol. 23, 1951, pp. 42÷44.

Informacje na temat metod i urządzeń do syntezy mowy można znaleźć w następujących materiałach firmowych i publikacjach naukowych:

Buric M. R.; Kohut J., Olive J. P.: *Speech synthesis.* The Bell System Technical Journal, vol. 60, nr 7, 1981, pp. 1621÷1631.
Clark J. E.: *A review of techniques for speech synthesis.* Journ. of Electr. and Electron. Eng. Australia, vol. 1, nr 1, 1981, pp. 21÷28.
Cooper F. S.: *Speech Synthetizers.* IV-th International Congress of Phonetic Sciences, 1962, pp. 3÷13.
Fant G., Martony J.: *Speech Synthesis.* Speech Transmision Laboratory Quarterly Progress and Status Report, Royal Institute of Technology, Stockholm, nr 2, 1962, pp. 18÷24.
Flanagan L. J., Coker C. H., Bird M. C.: *Digital computer simulation of a formant-vocoder speech synthesizer.* 15-th Ann. Meeting Audio Engr. Soc., Preprint 307, 1963.
Kacprowski J., Mikiel W.: *Realizacja procesu syntezy mowy za pomocą syntezatora SYNFOR II.* Praca IPPT PAN, nr 25, 1968.
Kacprowski J., Mikiel W.: *Simplified rules for parametric synthesis of nasal and stop consonanats in C-V Syllables by means of the terminal analog speech synthesizer.* Acustica, 16, 1965, pp. 356÷364.
Kacprowski J.: *Theoretical Bases of the Synthesis of Polish Vovels in Resonance Circuits.* Speech Analysis and Synthesis (W. Jassem — editor) nr 1, 1968, pp. 219÷288.
Kiełczewski G.: *Digital synthesis of speech and its prosodie features by means of microphonemic method.* Sprawozdanie II UW, zeszyt 65, 1978.
Myślecki W.: *Reguły generacji pobudzenia krtaniowego w procesie syntezy fraz mowy polskiej.* Raport nr I 28 (PRE-080) 79, Politechnika Wrocławska, 1979.
Rabiner R. L.: *Digital — formant synthesizer for speech synthesis.* Jour. Acoust. Soc. Amer., vol 43, 1968, pp. 822÷828.
Rabiner R. L., Jackson L. B., Schafer W. R., Coker C. H.: *Digital hardware for speech synthesis.* IEEE Trans. Commun. Techn., vol. COM-19, 1971, pp. 1016÷1020.

Informacje na temat konkretnych konstrukcji syntezatorów można znaleźć:
— w materiałach firm produkujących odpowiedni sprzęt:

National semiconductor linear databook, 1982.
Votrax speech synthesiser. Data Sheet, 1980.
TMS 5200 Voice Synthesis Procesor. Data Manual 1980.

— w periodykach zajmujących się tą problematyką:

*MEA 800 voice synthesizer: principles and interfacing.* Technical Publication 101 Electronic Components and Materials, Philips, 1983.
Elektron, Sep. 1981.
Electronic Desig, June 1981.
Robotics Age, Nov/Dec. 1981.
Interface Age, June 1979.
Radioelektronik, Grudzień 1984.
Speech Technology.

Podstawową pozycją literatury w zakresie percepcji mowy jest nadal książka:

Flanagan J. L.: *Speech analysis, synthesis and perception.* Springer Verlag, Berlin, Heidelberg, New York 1970.

Inne pozycje na ten temat to np.:

Ainsworth W. A.: *Mechanisms of speech recognition.* Int. Ser. in Natural Philosophy, vol. 85. Pergamon Press, Oxford—New York—Toronto—Sidney—Paris—Frankfurt 1976.
Bekesy G.: *Experiments in hearing.* Mc Graw Hill, New York 1960.
Daris H.: *Advances in the Neurophysiology and Neuroanatomy of the Cochlea.* Journ. Acoust. Soc. Amer., vol. 34, nr 8, part 2, 1962, pp. 1377÷1385.
Pozin N. V.: *Modelirowanie nejronnych struktur.* Nauka, Moskwa 1970.
Tadeusiewicz R., Izworski A.: *Cybernetic modelling of the initial part of the man auditory system,* Postępy Cybernetyki, vol. 7, 1984, pp. 43÷57.
Zwicker E., Feldtkeller R.: *Das Ohr als Nachrichtempfäger.* Hirzel Vg, Stuttgart 1967.

Model systemu słuchowego człowieka został omówiony m.in. w pracach:

Engstrom H., Ades H. W., Hawkins J. E.: *Structure and functions of the sensory hairs of the inner ear.* Journ. Acoust. Soc. Amer., vol. 34, 1962 pp. 1356÷1364.
Floch A., Kimura R., Lundquist P. G.: *Morphological basis of directional sensivity of the outer hair cells in the organ of Corti.* Journ. Acoust. Soc. Amer., vol. 34, 1962, pp. 1351÷÷1356.
Hass G. F.: *Electric Network Effects in the Cochlea.* Journ. Acoust, Soc. Amer., vol. 53, nr 1, 1973, pp. 2÷6.
Majewski J., Tadeusiewicz R.: *Cyfrowy model generowania impulsów czynnościowych w neuronie.* Archiwum Automatyki i Telemechaniki, vol. 24, nr 4, 1979, str. 499÷510.
Mikrut Z., Tadeusiewicz R.: *Identyfikacja i modelowanie neuronu na maszynie cyfrowej.* Archiwum Automatyki i Telemechaniki, vol. 23, nr 3, 1978, str. 345÷357.
Tadeusiewicz R.: *A Computer model of the initial part of the humen ear.* Proceedings of MECO'79, Acta Press, Anaheim, Calgary, Zurich 1979.
Tadeusiewicz R., Majewski J.: *Model dynamiki komórki nerwowej.* Zeszyty Naukowe AGH, nr 58, 1978, str. 159÷166.
Tadeusiewicz R.: *Modelowanie systemu słuchowego człowieka.* Prace I Ogólnopolskiego Seminarium System — Modelowanie — Sterowanie, Zakopane 1974.
Tadeusiewicz R., Pachowicz P.: *Model komórki rzęsatej w narządzie Cortiego człowieka.* Elektrotechnika, vol. 1, nr 4, 1982, str. 143÷154.
Tadeusiewicz R.: *Wybrane zagadnienia modelowania złożonych systemów biologicznych.* Zeszyty Naukowe AGH, nr 58, Kraków 1978, str. 167÷175.

Na temat psychologicznych aspektów percepcji mowy można przeczytać w:

Carlson R., Granstrom B. (eds.): *Representation of Speech in the Peripheral Auditory System.* Elsevier, Amsterdam 1982.
Czystowicz L., Goluszina A., Lubinskaja V., Malinnikowa T., Żukowa M.: *Psychological Methods in Speech Perception Research.* Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsfoeschung, vol. 21, 1968, pp. 33÷39.
Czystowicz L.: *Riecz, artikulacia i vospriniatie,* AN SSSR, Leningrad, 1965.
Sapożkow M. A.: *Sygnał mowy w telekomunikacji i cybernetyce.* WNT, Warszawa 1966.
Fletcher H.: *Speech and Hearing in Communication.* New York 1953.
Liberman M. A.: *Some Results of Research on Speech Perception.* Journ. Acoust. Soc. Amer., vol. 29, 1957, pp. 117÷123.
Luchsinger R., Arnold G. F.: *Voice, Speech, Language.* Belmont, California 1965.
Malecki I.: *Akustyka radiowa i filmowa.* Warszawa 1955.

Miller G. A.: *Language and Communication.* New York 1951.
Mol H.: *Fundamentals of Phonetics.* Hague 1963.
Paget R.: *Human Speech.* London 1930.
Zalewski J.: *Badania wyrazistości fonemowej mowy polskiej w funkcji częstotliwości.* Zeszyty Naukowe Politechniki Wrocławskiej, nr 101, 1965, str. 81÷92.

Ogólne informacje na temat metod opisu sygnału mowy można znaleźć między innymi w pracach:

Flanagan J. L.: *Synthesis and recognition of speech: teaching computers to listen.* Bell Laboratories Record, 1981, pp. 123÷151.
Jassem W.: *Podstawy fonetyki akustycznej.* PWN, Warszawa 1973.
Oppenheim A. V. (ed): *Sygnały cyfrowe — przetwarzanie i zastosowanie.* WNT, Warszawa 1982.
Oppenheim A. V., Schafer R. W.: *Cyfrowe przetwarzanie sygnałów.* WKŁ, Warszawa 1979.
Sapożkow M. A.: *Sygnał mowy w telekomunikacji i cybernetyce.* WNT, Warszawa 1966.

Informacje na temat opisu sygnału mowy w dziedzinie czasu można znaleźć w pracach:

Barney H. L., Dunn K. H.: *Speech Analysis.* Manual of Phonetics (L. Kaiser — editor), Amsterdam 1957, pp. 180÷201.
Beranek L. L.: *Acoustic Measurements.* New Yok 1949.
Denes P.: *The Use of Computers for Research in Phonetics.* IV International Congress of Phonetic Sciences, 1962, pp. 149÷154.
Fant G.: *The Acoustics of Speech.* III International Congres on Acoustics, vol. 1, 1959 pp. 199÷201.
Gray G., Wise C. M.: *Bases of Speech.* New York 1959.
Jakobson R., Fant G., Halle M.: *Preliminaries to Speech Analysis.* MIT resp. nr 13, Cambridge. Mass., 1972.
Niederjohn R. J.: *A mathematical formulation and comparison of zero-crosing analysis techniques which have been applied to automatic speech recognition.* IEEE Trans. on Acoustics Speech and Signal Proc., vol. ASSP-23, nr 4, 1975.
Wierzchowska B.: *Opis fonetyczny języka polskiego.* Warszawa 1967.

Opis sygnału mowy w dziedzinie częstotliwości podano w pracach:

Bluestein L. I.: *A linear filtering approach to the computation of the discrete fourier transform.* IEEE Trans. Audio Electroacoustics, vol. AU-18, Dec. 1970, pp. 451÷455.
Brigham E. O.: *The Fast Fourier Transform.* Prentice Hall, New York 1974.
Cooley J. W., Tukey J. W.: *An Algorithm for the Machine Calculation of Complex Fourier Series.* Math. of. Comp., vol. 19, nr 90, 1965, pp. 297÷301.
Dziurnikowski A.: *Uśrednianie częstotliwości tonu krtaniowego przy korelacyjnej metodzie jej estymacji, wykorzystującej algorytm liniowej predykcji.* Arch. Akust. T 16, z 1, 1981, s. 53÷74.
Papoulis A.: *The Fourier Integral and its Applications.* Mc Graw — Hill, 1962.
Randall R. B.: *Application of B and K Equipment to frequency analysis.* Brüel and Kjaer, Nearum. 1977.
Silverman H. E., Dixon N. R.: *A parametrically controlled spectral analysis system for speech.* IEEE Trans. Acoust. Speech Signal Proc. vol. ASSP-22, Oct. 1974, pp. 362÷381.
Wahrmant C. G.: *Averaging Time of Measurements.* B and K Application Note nr 11÷138, 1977.
Zalewski J., Majewski W.: *Cross correlation of long-term spectra as a speaker identification technique.* Acustica, vol. 34, nr 1, 1975, pp. 20÷24.

Na temat czasowo-częstotliwościowej zmienności sygnału mowy napisano w pracach:

Dudley H.: *Remarking speech.* Journ. Acoust. Soc. Amer., vol. 11, 1939, pp. 169÷177.
Randall R. B.: *High Speed Narrow Band Analysis wirth Digital Output.* B and K Application Note, nr 12—192, 1975.
Schafer B. W., Rabiner L. R.: *Design and simulation of a speech analysis-synthesis system based on short-time fourier analysis.* IEEE Trans. Audio Electroacoustic, vol. AU-21, June 1973, pp. 165÷174.
Tadeusiewicz R.: *Bioniczna koncepcja analizy mowy.* Metody bezpośredniego wprowadzania informacji tekstowej i obrazowej w systemach informatycznych. Jabłonna 1973.
Tadeusiewicz R., Jaworowski J.: *ART 73 b — język do przetwarzania informacji akustycznej za pomocą maszyny cyfrowej.* Archiwum Akustyki, vol. 10, nr 1, 1975, str. 11÷24.
Tadeusiewicz R.: *KART-1, konwerter do bezpośredniego wprowadzania informacji akustycznej do maszyny cyfrowej.* Archiwum Akustyki, vol. 10. nr 3, 1975, str. 217÷224.

Informacje na temat parametrycznego opisu sygnału mowy można znaleźć w pracach:

Izworski A., Tadeusiewicz R.: *Metody komputerowej ekstrakcji parametrów dystynktywnych z ciagłego sygnału mowy polskiej.* Archiwum Akustyki, vol. 18, nr 3, 1983, str. 253÷÷274.
Jassem W.: *Założenia ogólnego modelu rozpoznawania mowy.* Prace IPPT PAN nr 68/ /1977.
Lotko B., Tadeusiewicz R.: *Rytmiczno-częstotliwościowa metoda transformacji sygnału mowy dla celów akustycznego rozpoznawania ograniczonego słownika wyrazów.* Elektrotechnika, vol. 2, nr 4, 1983, str. 357÷365.
Majewski W., Hollien H.: *Formant frequency regions of Polish vowels.* Journ. Acoust. Soc. Amer., vol. 42, nr 5, 1967, pp. 1031÷1042.
Oppenheim A. V., Schafer R. W.: *Homomorphic analysis of speech.* IEEE Trans. Audio Electroacoustics, vol. AU-16, 1968, pp. 221÷226.
Schafer R. W., Rabiner L. R.: *System for automatic analysis of voiced speech.* Journ. Acoust. Soc. Amer., vol. 47, part 2, 1970, pp. 634÷648.
Wilusz T.: *Komputerowo wspomagane projektowanie sieci neuropodobnych dla przetwarzania sygnałów.* Rozprawa doktorska, AGH, Kraków 1983.
Wilusz T., Wołoszyn J.: *Model wydobywania pewnego rodzaju informacji zawartej w widmie sygnału.* Zeszyty Naukowe AE w Krakowie 1978.

Technikę liniowej predykcji opisano w pracach:

Atal B. S., Hanauer S. L.: *Speech analysis and synthesis by linear prediction of the speech wave.* Journ. Acoust. Soc. Amer., vol. 50, 1971, pp. 637÷655.
Atal B. S., Schroeder M. R.: *Adaptive predictive coding of speech signals.* Bell System Techn. Journ., vol. 49, no. 6, Oct. 1970, pp. 1973÷1986.
Durbin J.: *The fitting of time series models.* Rev., Inst. Intern. Statist., vol. 28, nr 3, 1969, pp. 233÷243.
Jurkiewicz J.: *Metoda predykcji liniowej o zmiennych parametrach do analizy mowy.* Raport nr I-28/P-009/84 Politechnika Wrocławska 1984.
Makhoul J.: *Linear prediction — a tutorial review.* Proc. IEEE, vol. 63, Apr. 1975, pp. 561÷580.
Makhoul J.: *Spectral analysis of speech by linear prediction.* IEEE Trans. Audio Electroacoustics, vol. AU-21, June 1973, pp. 140÷148.

Markel J. D., Gray A. H.: *Linear Prediction of Speech.* Springer Verlag, New York, 1976.
Shichor E., Silverman H.: *An improved LPC algorithm for voiced speech synthesis.* IEEE Trans. on Acoust. Speech and Signal. Proc. vol. ASSP-32, nr 1, 1984, pp. 180÷183.

Na temat opisu sygnału mowy z punktu widzenia teorii informacji można przeczytać w pracach:

Berry J.: *Some statistical aspects of conversational speech.* Communication theory (W. Jackson — editor), London 1953, pp. 392÷401.
Brachmański S. P., Majewski W.: *Gęstość prawdopodobieństwa wartości chwilowych sygnału mowy polskiej.* Arch. Akust. T 14, z 3, 1979, s. 207÷214.
Glenn W., Mitchcock B.: *With a speech pattern classifier computer listens to its mater's voice.* Electronics, May, 1971, pp. 84÷89.
Jassem W.: *Podstawy fonetyki akustycznej.* PWN, Warszawa 1973.
Lee Y. W.: *Statistical theory of communications.* J. Wiley, New York 1960.
Majewski J.: *Podstawy teorii informacji.* Skrypt AGH nr 955, Kraków 1984.
Roberts A. H.: *A statistical linguistic analysis of american english.* The Hague, 1965.
Steffen-Batogowa M.: *Częstość występowania głosek polskich.* Biul. PTJ, XVI, str. 145÷÷164.
Tadeusiewicz R.: *Metody oceny redundancji sygnału mowy i oszacowania redundancji dla mowy polskiej.* III Sympozjum MPN WEAIE AGH, Kraków 1977, str. 5÷8.
Young J. F.: *Information theory.* Butterworth, London 1971.

Informacje o sygnale mowy w automatyce można znaleźć w:

Lea W. A.: *Trends in speech recognition.* Prentice Hall, Englewood Cliffs 1980.
Reddy D. R.: *Speech recognition.* Academic Press, New York 1975.
Tadeusiewicz R.: *Głosowa komunikacja człowieka — operatora ze sterującą procesem maszyną cyfrową.* Prace II Krajowej Konferencji ,,Zastosowanie komputerów w przemyśle", Szczecin 1981, str. 232÷241.

O roli sygnału mowy w systemach sterowania napisano w:

Emeljanova V. N. (ed): *Analiz i sintiez reci v sistemach upravlenia,* Izd. AN SSSR, Moskva 1981.
Jassem W.: *Fonetyczno-akustyczne założenia automatycznego rozpoznawania fonemów.* Prace IPPT PAN, nr 14/1970.
Holmgren J. E.; *Toward Bell System application of automatic speech recognition.* The Bell System Technical Journal, vol. 62, nr 6, 1983, pp. 1865÷1880.
Kacprowski J.: *Teoretyczne podstawy metody automatycznego rozpoznawania samogłosek.* Archiwum Akustyki, vol. 2, nr 1, str 255÷266.
Newell A. (et al.): *Speech Underistanding Systems*; Final Raport of a Study Group, North — Holland/American Elservier, Amsterdam 1973.
Tadeusiewicz R.: *Głosowa łączność człowieka z maszyną cyfrową.* Zeszyty Naukowe AGH, Automatyka, nr 22, 1978.

Możliwości automatycznego rozpoznawania mowy dyskutowano w pracach:

Ackroyd M. H.: *Isolated word recognition using the weighted Levenshtein distance.* IEEE Trans. on Acoustic, Speech and Signal Processing, vol. ASSP-28, no 2, 1980, pp. 243÷÷244.
Atal B. S.: *Automatic speaker recognition based on pitch contoures,* JASA, vol. 52, nr 6, 1972.

Ciemiel G. I.: *Opoznavanie recevych signalov*. Nauka, Moskva 1971.
Górecki H., Skowiniak A., Tadeusiewicz R.: *Analiza możliwości zastosowania układów uczących się do rozpoznawania obrazów dźwiękowych*. Teoria sterowania — część V (H. Górecki — red.), Kraków 1975, str. 143÷156.
Stevens K. N.: *Towards a Model for Speech Recognition*. Journ. Acoust. Soc. Amer., vol. 32, 1960, pp. 47÷55.
Tadeusiewicz R.: *Układ do sterowania maszyn rozkazami wydawanymi słownie*. Zeszyty Naukowe AGH, Automatyka nr 13, 1976, str. 231÷238.
Włodarczyk H.: *Klasyfikacja głosek izolowanych niezależnie od sposobu artykulacji*, Archiwum Akustyki, vol. 5, 1970, str. 69÷84.

O sposobach wprowadzania sygnału mowy do systemów rozpoznawania napisano w pracach:

Atal B. S.: *Effectiveness of linear prediction charakteristics of the speech wave for automatic speaker identification and verification*. Journ. Acoust. Soc. Amer. vol. 55, 1974, pp. 1304÷÷1312.
Hanne R. J.: *Formant Analysis*. Univ. of Michigan Communication Sciences Laboratory, Report nr 12, 1965.
Tadeusiewicz R.: *KART1 — konwerter do bezpośredniego wprowadzania sygnałów akustycznych do maszyny cyfrowej*. Archiwum Akustyki, vol. 10, nr 3, 1975, str. 217÷224.
Tadeusiewicz R.: *Układ do bezpośredniego wprowadzania informacji akustycznych do maszyny cyfrowej*. Metody bezpośredniego wprowadzania i wyprowadzania informacj tekstowej i obrazowej w systemach informacyjnych. Jabłonna 1973.

Informacje na temat wydzielania parametrów sygnału mowy można znaleźć w pracach:

Daxer W., Zwicker E.: *On line isolated word recognition using a microprocesor system*. Speech Communication, 1, 1982, pp. 21÷27.
Izworski A., Tadeusiewicz R.: *Metody komputerowej ekstrakcji parametrów dystynktywnych z ciągłego sygnału mowy polskiej*. Archiwum Akustyki, vol. 18, nr 3, 1983, str. 253÷÷274.
Tadeusiewicz R., Książkiewicz-Jóźwiak B., Wszołek W.: *Widmowe kryteria oceny stopnia deformacji sygnału mowy w protetyce stomatologicznej*. VII Konferencja Biocybernetyki i Inżynierii Biomedycznej, Gdańsk 1985, str. 39÷42.
Lotko B., Tadeusiewicz R.: *Analiza rytmiczno-częstotliwościowa jako podstawa rozpoznawania ograniczonego słownika wyrazów*. VI Krajowa Konferencja Biocybernetyki i Inżynierii Biomedycznej, Warszawa 1983, str. 412÷415.

Podstawowe informacje na temat metod rozpoznawania obrazów znaleźć można między innymi w opracowaniach:

Duda R. C., Hart P. E.: *Pattern Classification and Scene Analysis*. John Wiley nad Sons, New York 1973.
Kulikowski J. L.: *Cybernetyczne układy rozpoznające*. PWN, Warszawa 1972.
Lachenbruch P. A.: *Discriminant Analysis*. Hafner Press, London 1975.
Nilsson N.: *Maszyny uczące się*. PWN, Warszawa 1965.
Patrick E. A.: *Fundamentals of Pattern Recognition*. Prontice Hall Inc., Englewood Cliffs, N. J. 1972.
Sklansky J., Wassel G. N.: *Pattern classifiers and trainable machines*. Springer Verlag, New York, Heidelber, London 1981.
Tadeusiewicz R.: *Ocena przydatności wybranych metryk w minimalnoodległościowych metodach rozpoznawania mowy*. Archiwum Akustyki, vol. 18, nr 3, 1983, str. 275÷284.

Tadeusiewicz R.: *Rozpoznawanie obrazów — zarys teorii.* Skrypt Uniwersytetu Jagiellońskiego w Krakowie, 1985.

Na temat elementów systemu rozpoznającego napisano w pracach:

Bolc L., Cichy M., Różańska L.: *Przetwarzanie języka naturalnego.* WNT, Warszawa 1982.

Chenng R. S., Eisenstein B. A.: *Feature selection via dynamic programming for text independent speaker identification.* IEEE Trans. on Acoustics, Speech and Signal Processing, vol. ASSP-26, nr 5, 1978, pp. 397÷403.

Ciemiel G. I., Sorokin A. (red): *Problemy postroenia system ponimania reci.* Nauka, Moskva 1980.

Grocholewski S.: *Programowanie dynamiczne w automatycznym rozpoznawaniu mowy.* Elektrotechnika, tom 4, zeszyt 1, 1985, str. 109÷126.

Itakura F.: *Minimum prediction residual principle applied to speech recognition.* IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-23, Feb. 1975, pp. 67÷72

Jelinek J.: *Continuous speech recognition by statistical methods.* Proc IEEE, vol. 64, 1976, nr 4, pp. 532÷556.

Kuhn M. H., Tomaschewski H. H.: *Improvements in isolated word recognition.* IEEE Trans. Acoust. Speech Signal Processing, vol. ASSP — 31, Feb. 1983, pp. 157÷167.

Basztura Cz., Majewski W.: *Wpływ wybranych parametrów kanału telefonicznego na identyfikację głosu.* Arch. Akust., T 16, z 4, 1981, s. 404÷416.

Mazoń S., Tadeusiewicz R.: *Próba opisu pewnej klasy systemów naturalnej konwersacji człowieka z maszyną cyfrową.* Archiwum Automatyki i Telemechaniki, vol. 24, nr 2, 1979, str. 293÷299.

Mayers C., Rabiner L. R., Rosenberg A. E.: *Performance tradeoffs in dynamic time wraping algorithms for isolated word recognition.* IEEE Trans. Acoust. Speech Signal Processing, vol. ASSP-28, Dec. 1980, pp. 622÷633.

Rabiner L. R., Levinson S. E., Rosenberg A. E., Wilpon J. G.: *Speaker-independent recognition of isolated words using clustering techniques.* IEEE Trans., Acoust. Speech, Signal Processing, vol. ASSP-21, Aug. 1979, pp. 336÷349.

Rabiner L. R., Rosenberg A. E., Levison S. E.: *Considerations in dynamic time warping algorithms for discrete word recognition.* IEEE Trans. Acoust. Speech, Signal Processing. vol. ASSP — 26, Dec. 1978, pp. 575÷582.

Sakoe H., Chiba S.: *Dynamic programming algorithm optimization for spoken word recognition.* IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-26, Feb. 1978, pp. 43÷49.

Slutsker G. S.: *Nieliniejnyj mietod analiza riecziewych signałow.* Trudy NIIR, nr 2, 1968, str. 76÷82.

Tadeusiewicz R.: *Głosowa łączność człowieka z maszyną cyfrową.* Zeszyty Naukowe AGH, Automatyka, nr 22, 1978.

Velichko V. M., Zagoruyko N. G.: *Automatic recognition of 200 words.* Int. Journ. Man-Machine Studies, vol. 2, June 1970, pp. 223÷234.

White G., Neely R.: *Speech recognition experiments with linear prediction, bandpas filtering and dynamic programing.* IEEE Trans. Acoust., Speech, Signal Processing. vol. ASSP-24, Apr. 1976, pp. 183÷188.

Winograd T.: *Understanding Natural Language.* Edinburgh 1972.

Problemy związane z sygnałem mowy w telekomunikacji omówiono w pracach:

Alles H. G., Codon J. H., Fisher W. C., McDonald H. S.: *Digital signal processing in telephone switching.* Proc. Intern. Conf. on Comm., Minneapolis 1974.

Flanagan J. L.: *A Resonance — vocoder of baseband system for speech transsmision.* III International Congress on Acoustics, vol. I, 1959, pp. 211÷214.
Lea W. A.: *Trends in speech recognition.* Prentice Hall Inc., Englewood Cliffs, New Jersey 1980.
Lucky R. W.: *Techniques for adaptive equalization of digital communication system.* Bell System Techn. Journ., vol. 45, nr 2, Feb. 1966, pp. 255÷286.
Oppenheim A. V. (editor): *Sygnały cyfrowe — przetwarzanie i zastosowania.* WNT, Warszawa 1982.
Sapożkow M. A.: *Sygnał mowy w telekomunikacji i cybernetyce.* WNT, Warszawa 1966.
Steenehen H. J. M.; Hontgast T.: *A physical method for measuring speech transmission quality.* J. Acoust. Soc. Am., vol. 67, nr 1, 1980 pp. 318÷326.

O metodach kompresji sygnału mowy napisano w pracach:

Bially T., Anderson W.: *A digital channel vocoder.* IEEE Trans. Commun. Techn., vol. COM-18, nr 4, Aug. 1970, pp. 435÷442.
Fant G., Risberg A.: *Evaluation of speech compression systems.* Speech Transmission Laboratory Quarterly Progress and State Report, Royal Institute of Technology, Stockholm 1963, nr 2, pp. 15÷21.
Flanagan L. J., Golden M. R.: *Phase vocoder.* Bell System Techn. Journ., vol. 45, 1966, pp. 1493÷1509.
Gold B., Rader C. M.: *The channel vocoder.* IEEE Trans. Audio Electrocoustics, vol. AU-15, Dec. 1967, pp. 148÷160.
Golden E.: *Vocoder filter design: practical considerations.* Journ. Acoust. Soc. Amer., vol. 43, Apr. 1968, pp. 803÷810.
Marcou P., Daguet J.: *New methods of speech transmission.* Information Theory (C. Cherry — editor), London 1955.
Rabiner L. R.: *Digital — formant synthesizer for speech-synthezis studies.* Journ. Acoust. Soc. Amer., vol. 43, 1968, pp. 822÷828.
Schroeder M. R.: *Vocoders: analysis and synthesis of speech.* Proc. IEEE, vol. 54, May 1966, pp. 720÷734.

Informacje o problemach kryptofonii można znaleźć w pracach:

French R. C.: *Speech scrambling and synchronization.* Philips Res. Rep. nr 9, 1973.
Peciak J.: *Maskowanie i badanie jakości metod maskowania sygnałów mowy.* Zeszyty Naukowe WSMW, nr 86 A, Gdynia 1985.
Peciak J.: *O utajnianiu mowy bez tajemnic.* MON, Warszawa 1980.
Udalov S.: *Microprocessor-based techniques for analog voice privacy.* Conf. Rec. Int. Commun., June 1980, pp. 1641÷1645.