ZESZYTY NAUKOWE POLITECHNIKI ŚLĄSKIEJ 1996
Seria: ELEKTRONIKA z.5 Nr kol. 1341

Jacek **KONOPACKI**

Instytut Elektroniki
Politechniki Śląskiej

CYFROWA ANALIZA I SYNTEZA SYGNAŁU MOWY

Streszczenie. Artykuł zawiera krótki przegląd podstawowych metod cyfrowej analizy i syntezy sygnału mowy. Dokładniej omówiono analizę formantową, kodowanie liniowo predykcyjne i syntezę w dziedzinie czasu. Podano przykłady opatrzone praktycznymi uwagami pomocnymi przy syntezie sygnału mowy.

DIGITAL SPEECH ANALYSIS AND SYNTHESIS

Summary. A brief review of fundamental speech analysis and synthesis algorithms for digital processing is presented in the paper. Formant analysis, linear predictive coding and time-domain synthesis are described more precisely. Some examples of speech synthesis and useful remarks for its preparations are also included.

ANALYSE ET SYNTHÉSE NUMERIQUE DE SIGNAL VOCAL

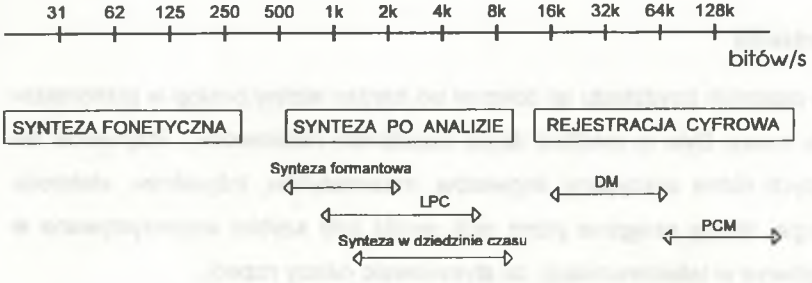
Résumé. La brève revue des algorithmes fundamentaux pour analyser et synthéser le signal vocal est présenté dans cet article. Analyse de formants, prédiction linéaire et synthése en domaine temporel sont decrits le plus précisément. Quelques exemples de synthése de la voix et les remarques utiles pour leurs préparations sont aussi contenus.

1. Wprowadzenie

W ciągu ostatnich trzydziestu lat dokonał się bardzo istotny postęp w przetwarzaniu sygnału mowy. Było to możliwe dzięki współpracy naukowców i inżynierów reprezentujących różne dyscypliny: lingwistów, matematyków, inżynierów, elektroników. Z drugiej strony osiągane przez nich wyniki były szybko wykorzystywane w praktyce, głównie w telekomunikacji, co stymulowało dalszy rozwój.

Do połowy lat sześćdziesiątych przetwarzanie sygnału mowy odbywało się prawie wyłącznie na drodze analogowej. Zastosowanie maszyn cyfrowych stworzyło nowe możliwości i wówczas powstały algorytmy, które nie miały odpowiedników w przetwarzaniu analogowym. Prawdziwa rewolucja dokonała się jednak z chwilą opracowania układów wielkiej skali integracji (VLSI). Wykonane w postaci układów scalonych tanie syntezatory mowy mogły znaleźć zastosowanie w sprzęcie powszechnego użytku (mówiące kalkulatory, sprzęt gospodarstwa domowego). Jednocześnie w specjalistycznym sprzęcie można było zastosować bardziej wyszukane algorytmy (np. kompresji i kodowania) działające w czasie rzeczywistym.

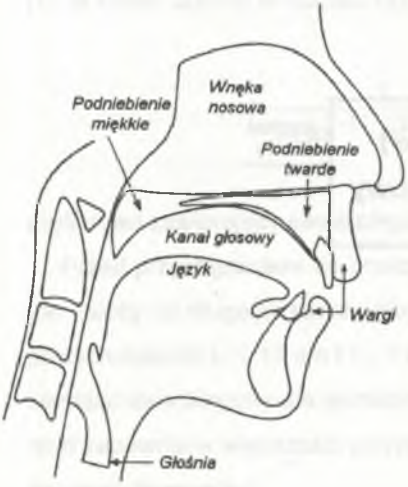
Z przetwarzaniem sygnału mowy związane są zagadnienia analizy i syntezy mowy, przy czym w konkretnym zastosowaniu może wystąpić każda z tych czynności osobno lub obie jednocześnie. Podczas identyfikacji i weryfikacji mówiącego mamy do czynienia z analizą mowy. Tylko synteza występuje w maszynach czytających automatycznie. Natomiast typowym przykładem, gdzie mogą wystąpić obydwie czynności, jest przesyłanie sygnału mowy. Wykonanie przed kodowaniem analizy sygnału mowy, w czasie której wnika się w jego strukturę, pozwala uzyskać większe współczynniki kompresji w stosunku do tradycyjnych metod kodowania sygnałów akustycznych, takich jąk: modulacja impulsowo-kodowa (PCM), modulacja delta (DM), różnicowa modulacja impulsowo-kodowa (DPCM) itp. Na rys. 1. [5] przedstawiono zależność liczby bitów potrzebnych do zakodowania jednej sekundy sygnału mowy od użytej metody syntezy. Jak widać, największe współczynniki kompresji zapewnia synteza fonetyczna. Jej zastosowanie wymaga poznania podstawowych dźwięków mowy ludzkiej (tzw. fonemów) oraz reguł ich łączenia, intonacji itp.



Rys. 1. Liczba bitów na 1 sek sygnału mowy dla różnych metod syntezy
Fig. 1. The data rates associated with various speech synthesis methods

Niniejszy artykuł poświęcony jest cyfrowej analizie i syntezie mowy. W rozdziale drugim opisano model wytwarzania mowy oraz podano podstawowe metody wyznaczania parametrów tego modelu. Kolejne dwa rozdziały poświęcone są syntezie sygnału mowy prowadzonej w dziedzinie częstotliwości i czasu. Artykuł zawiera przykłady działania opisanych algorytmów zastosowanych do rzeczywistego sygnału mowy. Pomimo że praca nie ma charakteru oryginalnego, zamieszczono w niej praktyczne uwagi pomocnicze przy cyfrowej syntezie formatowej i syntezie z wykorzystaniem kodowania liniowo-predykcyjnego (z ang. linear predictive coding - LPC).

2. Model sygnału mowy

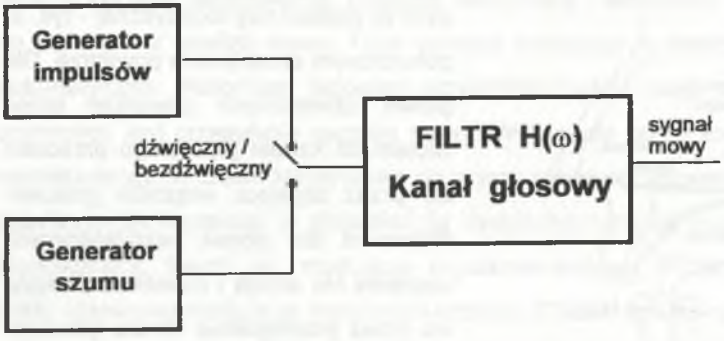


Rys. 2. Przekrój narządu mowy
Fig. 2. Section of the speech organs

Mowa wytwarzana jest w kanale głosowym (w postaci rury akustycznej - rys. 2) pobudzonym przepływem powietrza. Dla głosek dźwięcznych powietrze przed wlotem do kanału głosowego przeciska się przez drgające wiązadła głosowe. Natomiast dla głosek bezdźwięcznych wiązadła nie drgają i powietrze przepływa przez przewężenie kanału głosowego w sposób turbulentny. Kanał głosowy nie ma stałego kształtu, zmienia się on w zależności od typu głoski. Dodatkowo dla głosek nosowych otwiera się wnęka nosowa zmieniając parametry traktu głosowego.

Przedstawiony układ wytwarzania mowy można zamodelować w dwojaki sposób: za pomocą syntezatora widmowo-parametrycznego lub syntezatora konfiguracyjnego [16]. Pierwszy z syntezatorów stanowi filtr liniowy o parametrach wolnozmiennych, który aproksymuje charakterystykę częstotliwościową kanału głosowego. Drugi symuluje bezpośrednio zjawiska występujące w kanale głosowym, a więc wyraża ciśnienie powietrza jako funkcję czasu i położenia w rurze akustycznej o zmiennym przekroju. Przeprowadzone w dalszej części tego artykułu rozważania

dotyczą jedynie syntezatora widmowo-parametrycznego, którego uproszczony schemat przedstawia rysunek 3. Układ składa się z filtru, o charakterystyce częstotliwościowej H(ω), pobudzanego ze źródła szumowego dla głosek bezdźwięcznych lub ze źródła kwaziokresowych impulsów dla głosek dźwięcznych. Z charakteru mowy wynika, że w krótkich odcinkach czasu (10 do 20 ms) parametry kanału głosowego nie zmieniają się. W takim razie można przyjąć, że również charakterystyka filtru H(ω) jest w tym czasie stała. Proces syntezy mowy musi być poprzedzony jej analizą, która polega na: wyznaczeniu współczynników filtru, wybraniu odpowiedniego źródła oraz określeniu częstotliwości tego źródła (tzw. tonu krtaniowego) dla głosek dźwięcznych. Czynności te należy przeprowadzić w kolejnych odcinkach czasu (segmentach czasowych).



Rys. 3. Schemat syntezatora widmowo-parametrycznego
Fig. 3. Basic speech sythesis model

Charakterystyka częstotliwościowa kanału głosowego jest typu rezonansowego. Jeśli przyjąć, że kanał głosowy jest rurą akustyczną o długości 17 cm otwartą z jednej strony (usta) i zamkniętą z drugiej (głośnia), to trzy pierwsze rezonanse występują dla częstotliwości 500, 1500 i 2500 Hz [5]. W rzeczywistości charakterystyka kanału głosowego jest bardziej złożona i częstotliwości rezonansowe (zwane formantami) nie leżą tak regularnie. Nadal jednak można podać przedziały, w których znajdują się kolejne formanty [21]. Stała jest także ich liczba i wynosi 4 do 5 w zakresie do 5 kHz. Ponadto położenie formantów na osi częstotliwości jest powiązane z cechami osobniczymi i może być użyte do rozpoznania mowy [3, 4, 24].

Występowanie formantów w charakterystyce kanału głosowego sprawia, że w synezatorze widmowo-parametrycznym można filtr H(ω) przedstawić w postaci kaskadowo połączonych filtrów rezonansowych drugiego rzędu. Liczba tych filtrów zależy od liczby formantów. Opisany układ nazwano syntezatorem formantowym. W czasie analizy sygnału mowy dla każdego filtru formantowego należy wyznaczyć jego częstotliwość rezonansową i szerokość pasma. Pierwsze syntezatory formantowe wykorzystywały analogowe filtry rezonansowe [7]. Obecnie syntezatory widmowo-parametryczne realizuje się na drodze cyfrowej.

Z reguły dla filtru H(z) zakłada się, że posiada on same bieguny. W praktyce dla wielu dźwięków (szczególnie głosek nosowych) transmitancja H(z) powinna zawierać zera. Ponieważ zera te leżą na płaszczyźnie Z wewnątrz okręgu jednostkowego [1], to każdy czynnik w liczniku H(z) o postaci $(1-az^{-1})$ można aproksymować przez:

$$\frac{1}{1+az^{-1}+a^2z^{-2}+\ldots} \qquad dla \quad |a|<1 \tag{1}$$

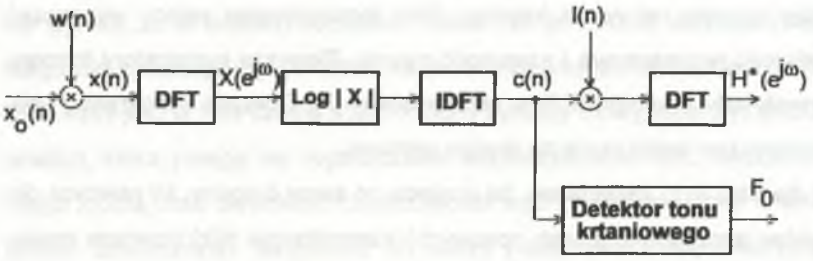
czyli układ zawierający same bieguny.

Przed przystąpieniem do analizy należy jeszcze określić rząd M filtru H(z). Rząd ten zależy od długości kanału głosowego L i częstotliwości próbkowania F_p. Dla typowych danych L = 17 cm i F_p = 10 kHz otrzymamy w przybliżeniu M = 10 [1]. Dokładając dwa bieguny dla aproksymacji ewentualnego zera uzyskamy M = 12. Taki rząd zapewnia w większości przypadków prawidłową identyfikację kanału głosowego do pięciu formantów.

2.1. Wyznaczenie współczynników filtru H(z)

Do obliczenia współczynników filtru modelującego kanał głosowy stosuje się dziś najchętniej dwie metody - filtracji hamomorficznej [2, 16, 17] i predykcji liniowej [2, 5, 11]. Zgodnie z przedstawionym modelem sygnał mowy jest splotem funkcji pobudzenia i odpowiedzi impulsowej kanału głosowego. Operację odwrotną, czyli rozplot, można uzyskać za pomocą filtracji homomorficznej. Na rysunku 4 przedstawiono schemat blokowy przetwarzania sygnału mowy x(n) pozwalający uzyskać nie tylko

obwiednię widma kanału głosowego, ale również rodzaj pobudzenia [21]. Na wstępie sygnał x(n) jest mnożony przez funkcję okna Hamminga w(n). Analizę kończy estymacja formantów na podstawie obwiedni widma kanału głosowego $H^*(e^{j\omega})$.



Rys. 4. Schemat homomorficznej analizy mowy
Fig. 4. Block diagram of the system for estimating formant frequencies and pitch period

Bardziej dogodna i częściej stosowana jest metoda predykcji liniowej, gdyż pozwala bezpośrednio obliczyć współczynniki filtru H(z). Jeśli transmitacja H(z) posiada same bieguny, to:

$$H(z) = \frac{G}{A(z)}; \quad A(z) = 1 + \sum_{k=1}^{M} a_k z^{-1} \quad (M - rząd\ \ filtru) \tag{2}$$

Odpowiedź impulsowa tego filtru jest równa:

$$s(n) = G\delta(n) - \sum_{k=1}^{M} a_k \ s(n-k) \tag{3}$$

Dla n > 0 powyższy wzór upraszcza się do postaci:

$$s(n) = -\sum_{k=1}^{M} a_k \ s(n-k) \tag{4}$$

czyli s(n) jest kombinacją liniową poprzednich wartości. Jeśli modelowany sygnał jest rzeczywiście odpowiedzią impulsową poszukiwanego filtru, to ostatnie równanie jest spełnione dokładnie. W przeciwnym wypadku otrzymamy estymatę tej odpowiedzi, którą można oznaczyć jako s*(n):

$$s^*(n) = -\sum_{k=1}^{M} a_k \; s(n-k) \qquad n \rangle 0 \tag{5}$$

Minimalizując błąd średniokwadratowy E

$$E = \sum_n e^2(n); \qquad e(n) = s(n) - s^*(n) \tag{6}$$

otrzymuje się równanie:

$$\sum_{k=1}^{M} a_k \sum_n s(n-k)s(n-i) = -\sum_n s(n)s(n-i) \qquad 1 \le i \le M \tag{7}$$

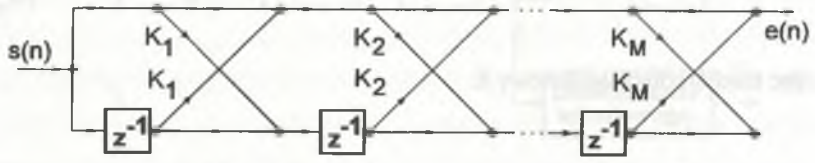
z którego można obliczyć współczynniki a_k. W zależności od sposobu obliczania sumy ze wskaźnikiem n we wzorze (7) otrzymuje się dwie metody - autokowariancji lub autokorelacji. Częściej stosuje się metodę autokorelacji i wówczas wzór (7) ma postać:

$$\sum_{k=1}^{M} a_k R(i-k) = -R(i) \qquad 1 \le i \le M \tag{8}$$

gdzie R(i-k) jest macierzą autokorelacji typu Toeplitza. Rekurencyjny algorytm rozwiązania równania (8) opracował Levinson, a następnie zmodyfikował go Durbin [5,11]. Po wyznaczeniu współczynników predykcji a_k wzmocnienie G oblicza się z zależności [11]:

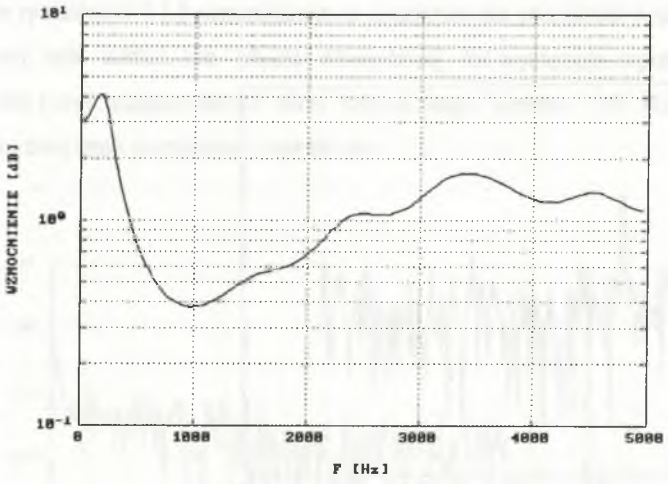
$$G^2 = R(0) + \sum_{k=1}^{M} a_k \; R(k) \tag{9}$$

Predykcja liniowa realizowana w strukturze bezpośredniej (wzór (5)) jest bardzo popularna w obliczeniach komputerowych. W komercyjnych scalonych syntezatorach chętniej stosuje się strukturę kratową [5] (rys. 5), która wynika wprost z algorytmu Levinsona-Durbina. Jej zaleta polega na tym, że dla stabilnego filtru współczynniki K_i (rys. 5) są zawsze mniejsze od jeden, a więc strukturę tę łatwo zrealizować w arytmetyce stałoprzecinkowej.



Rys. 5. Struktura kratowa predyktora (w postaci grafu)
Fig. 5. Lattice predictor

Metoda predykcji liniowej ma swoje zastosowanie również wtedy, gdy wymagane jest wyznaczenie częstotliwości formantowych. Mając dane współczynniki a_k można rozwiązać równanie zespolone A(z) = 0. Pierwiastki tego równania to bieguny transmitancji H(z), z których (jeśli są zespolone) oblicza się częstotliwości formantowe. Wielokrotne rozwiązanie równania zespolonego w kolejnych segmentach jest czasochłonne i w większości zastosowań nie do przyjęcia. Toteż opracowano cały szereg efektywniejszych algorytmów. Najprostsze polegają na „przeglądaniu" charakterystyki widmowej filtru H(z) punkt po punkcie i wybraniu lokalnych maksimów [12]. Zdarza się jednak, że dwa formanty leżą w niewielkiej odległości od siebie i w widmie występuje jedno z maksimów zamiast dwóch (rys. 6). W takim wypadku można obliczyć charakterystykę częstotliwościową filtru H(z) na okręgu o promieniu mniejszym od jeden, co spowoduje „wyostrzenie" maksimów i nawet blisko leżące formanty będą rozróżnialne [14]. Inne podejście polega na wykorzystaniu algorytmu split Levinsona [6], który - jak pokazano w [25] - pozwala precyzyjniej wyznaczać częstotliwości formantowe. Kolejne rozwiązania stosowane w automatycznych systemach rozpoznawania mówcy zaproponowano w [9, 22].



Rys. 6. Widmo segmentu, w którym brak maksimum dla drugiego formantu
Fig. 6. Spectrum of the segment in which is no maximum for second formant

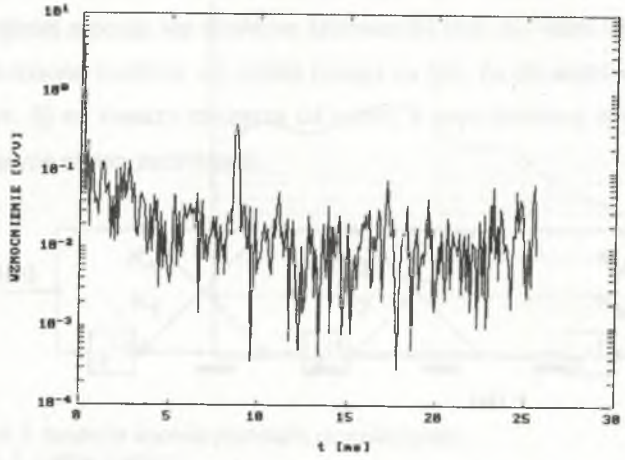
2.2. Estymacja częstotliwości tonu krtaniowego

Wyznaczenie częstotliwości tonu krtaniowego jest nieodłączną częścią systemów przetwarzania sygnału mowy. Częstotliwość tę wykorzystuje się we wszystkich syntezatorach widmowo-parametrycznych, a także w układach rozpoznawania mówcy. Algorytm estymacji tonu krtaniowego można podzielić na trzy grupy [18]:

1. wykorzystujące czasowe własności sygnału mowy,
2. wykorzystujące własności tego sygnału w dziedzinie częstotliwości,
3. wykorzystujące jednocześnie czasowe i częstotliwościowe własności sygnału.

Algorytmy pierwszej grupy operują bezpośrednio na próbkach sygnału mowy. Częstotliwość tonu krtaniowego wyznaczana jest przez detekcję wartości szczytowych, zliczania punktów przejść przez zero lub obliczenie funkcji autokorelacji [8,20].

W drugiej grupie znajdują się algorytmy, które wykorzystują fakt, że w widmie sygnału okresowego występują harmoniczne dla częstotliwości podstawowej i jej wielokrotności. Dokonując odpowiedniego przekształcenia widma, np. obliczenia cepstrum [15,21], można ustalić poszukiwaną częstotliwość.



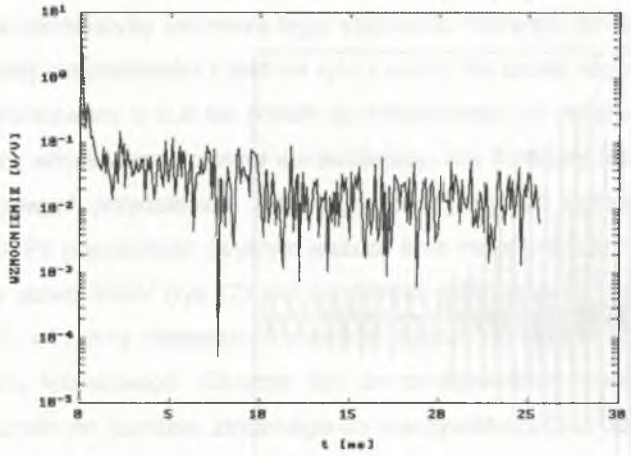
Rys. 7. Cepstrum dla głoski /o/
Fig. 7. Cepstrum for vowel /o/

Algorytmy hybrydowe (trzeciej grupy) wykorzystują techniki częstotliwościowe do wygładzenia sygnału czasowego, by następnie na podstawie np. funkcji autokorelacji obliczyć częstotliwość tonu krtaniowego [13, 23].

W większości algorytmów, gdy nie można wyznaczyć częstotliwości i tonu krtaniowego, określa się, czy jest to sygnał bezdźwięczny, czy brak sygnału (cisza). Informacja ta jest szczególnie ważna w systemach pracujących aytomatycznie.

Spośród wielu algorytmów bardzo trudno wybrać najlepszy. Rabiner wraz z współpracownikami [18] dokonał kompleksowego porównania siedmiu różnych metod. Analizie poddano pojedyncze słowa i całe zdania wypowiedziane przez różne osoby (mężczyzn, kobiety, dzieci) zarejestrowane za pomocą różnych mikrofonów. Okazało się, że każdy z algorytmów miał swoje mocne i słabe strony. Jak pokazały testy, metoda obliczania cepstrum jest bardzo dobra do wyznaczania częstotliwości tonu krtaniowego. Gorzej wypada na tle innych, gdy ją zastosować do określenia, czy dany segment mowy jest dźwięczny, czy bezdźwięczny.

Na rysunkach 7 i 8 przedstawiono cepstrum dla głosek /o/ oraz /s/ (fragmenty po 50 ms). Jak widać, dla głoski dźwięcznej /o/ występuje wyraźne maksimum dla 8,7 ms (czyli częstotliwość tonu krtaniowego wynosi 115 Hz). Dla głoski bezdźwięcznej brak wyraźnych maksimów.

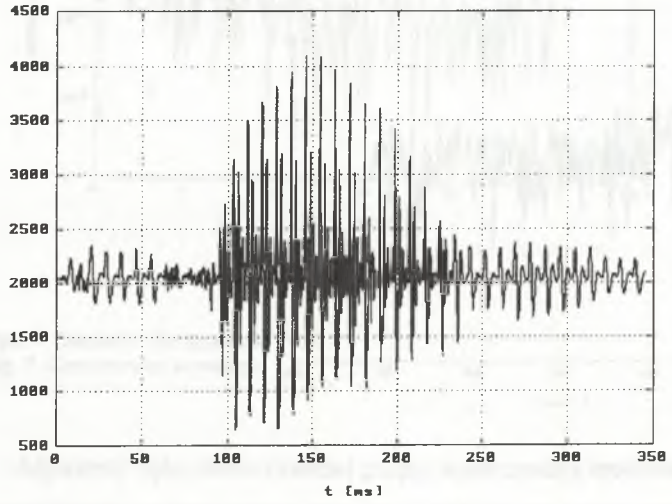


Rys. 8. Cepstrum dla głoski /s/
Fig. 8. Cepstrum for sound /s/

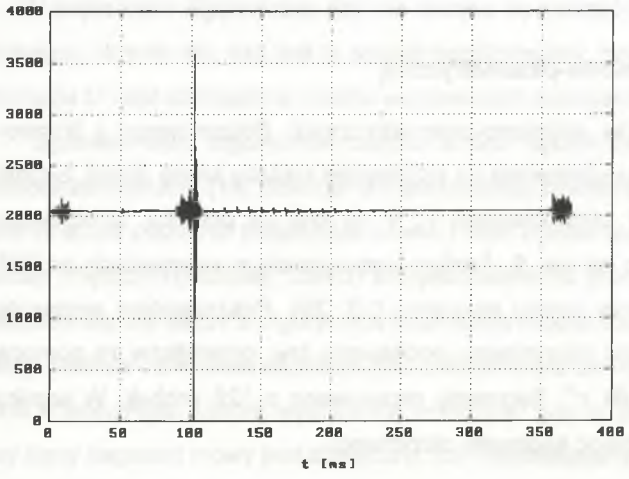
3. Przykłady syntezy widmowo-parametrycznej

Działanie syntezatorów widmowo-parametrycznych (formatowego i liniowo-predykcyjnego) zostanie przedstawione na przykładzie syntezy słowa /dom/. Sygnał oryginalny spróbkowany z częstotliwością F_p = 10 kHz za pomocą 12-bitowego przetwornika AC pokazano na rys. 9. Analizę tego sygnału w segmentach po 256 próbek przeprowadzono przy użyciu programu ILS [26]. Poszczególne segmenty mnożono przez funkcję okna Hamminga i poddawano tzw. preemfazie za pomocą filtru o transmitancji $1 - 0{,}98\ z^{-1}$. Segmenty przesuwano o 128 próbek. W wyniku działania programu dla każdego segmentu otrzymano:

* współczynniki K_i (predyktora kratowego - rys. 5), które przeliczono na współczynnik predykcji liniowej a_k,
* częstotliwości środkowe i szerokości pasm filtrów formantowych,
* energię sygnału oryginalnego, na podstawie której estymowano wzmocnienie G,
* częstotliwość tonu krtaniowego (dla głosek dźwięcznych).

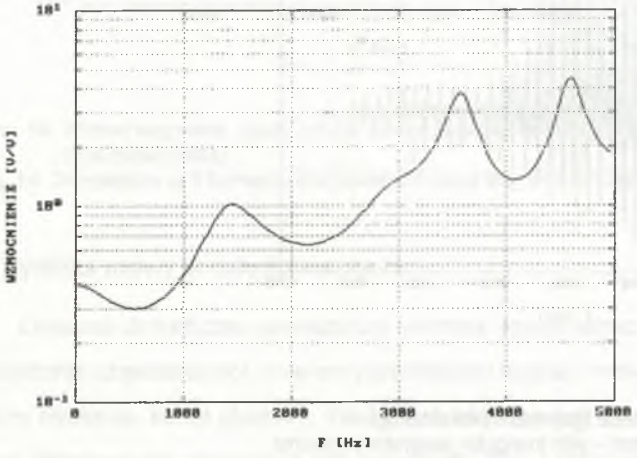


Rys. 9. Przebieg czasowy dla słowa /dom/
Fig. 9. Waveform for word /dom/ (digitised)

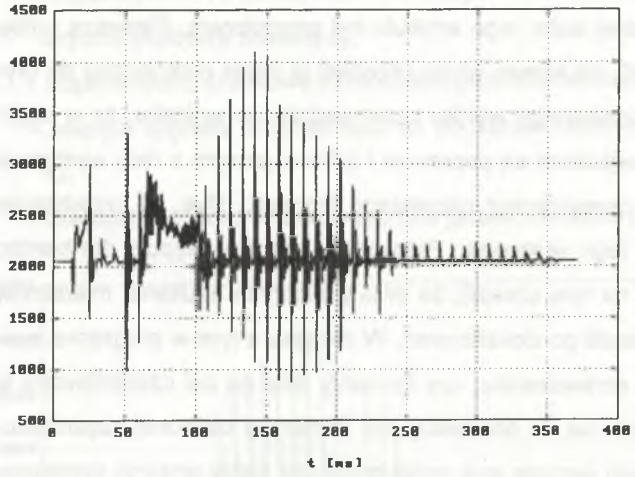


Rys. 10. Pierwsza próba syntezy słowa /dom/
Fig. 10. Synthesised word /dom/ - first attempt

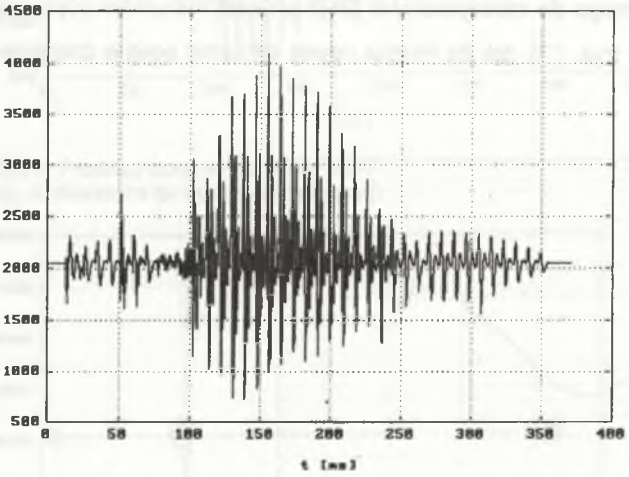
Syntezę formantową przeprowadzono korzystając z programu napisanego w ramach pracy dyplomowej [10], której autor tego artykułu był promotorem. Pierwszą próbę syntezy przedstawia rys. 10, na którym widać przebieg w ogóle niepodobny do oryginału. Wnikliwa analiza uzyskanego wyniku doprowadziła do wniosku, że w większości segmentów dane wejściowe są poprawne i tylko w jednym z nich występuje błąd gruby spowodowany niewykryciem pierwszego formantu. Rys. 11 przedstawia charakterystykę widmową tego segmentu. Pierwszy formant występuje dla bardzo małej częstotliwości i jest na tyle szeroki, że prosty algorytm szukania maksimów zastosowany w ILS nie potrafił go zlokalizować. W związku z tym w programie syntezy wbudowano korektor sprawdzający, czy formanty leżą na osi częstotliwości w typowych przedziałach. Jeśli nie, to dodawany jest formant o ustalonej częstotliwości. Po poprawkach (wykryto jeszcze inne mniej znaczące błędy analizy) syntetyczne słowo /dom/ (rys.12) jest już bardzo podobne do oryginału (z wyjątkiem głoski /d/), lecz przy odtwarzaniu wyraźnie słychać brzęczenie pochodzące od generatora tonu krtaniowego. Okazuje się, że zastosowanie pobudzenia o przebiegu piłokształtnym (bardziej zbliżonego do rzeczywistości [24]) zamiast impulsów poprawia zasadniczo jakość syntezy (rys. 13), tak że można nawet odróżnić pewne charakterystyczne cechy mówcy.



Rys. 11. Widmo segmentu, dla którego analizator nie wykrył pierwszego formantu
Fig. 11. Spectrum of the segment for which first formant has been omitted

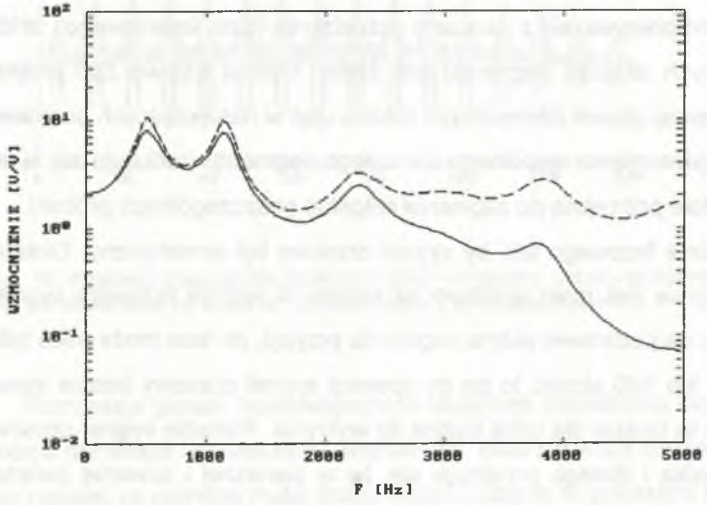


Rys. 12. Słowo /dom/ po syntezie (generator impulsów)
Fig. 12. Synthesised word /dom/ (after corrections) with pulse generator as source



Rys. 13. Słowo /dom/ po syntezie (generator piłokształtny)
Fig. 13. Synthesised word /dom/ - with triangular generator as source

Dla syntezy przeprowadzonej bezpośrednio przy użyciu współczynników predykcji nieprzyjemne zniekształcenie mowy pochodzące od generatora pobudzającego jest bardziej wyraźne. Zmiana kształtu sygnału pobudzającego eliminuje ten nieprzyjemny efekt, ale nie w takim stopniu jak poprzednio. Wydaje się, że przyczyna różnego brzmienia mowy dla tych syntezatorów leży w charakterystyce widmowej. Na rys. 14 pokazano te charakterystyki dla wybranego segmentu głoski /o/. Jak widać, dla syntezatora formantowego charakterystyka po ostatnim maksimum szybko opada tłumiąc tym samym wyższe harmoniczne pochodzące od pobudzenia.



Rys. 14. Widmo segmentu głoski /o/ dla syntezatora formantowego (linia ciągła) i syntezatora LPC (linia przerywana)
Fig. 14. Comparison of 5 formants filter spectrum (solid line) with LPC spectrum

4. Synteza mowy w dziedzinie czasu

Opisane dotychczas syntezatory widmowo-parametryczne dokonują syntezy w dziedzinie częstotliwości, bowiem parametrami sygnału mowy są współczynniki filtru, który modeluje kanał głosowy. Wadą tych układów jest skomplikowany proces syntezy. Wymaga on utworzenia odpowiedniego filtru i obliczenia odpowiedzi tego filtru na pobudzenie. Dodatkowe utrudnienie sprawia konieczność zapewnienia łagodnych zmian sygnału wyjściowego podczas przechodzenia do następnego segmentu.

Inne podejście, zwane syntezą w dziedzinie czasu, zapropnował Mozer [4, 5]. Metoda ta operuje bezpośrednio na próbkach sygnału mowy. Wykorzystując własności tego sygnału pozwala ona zapisać zawartą w nim informację na dużo mniejszej liczbie bitów (kompresja). Sama synteza polega na prostej dekompresji sygnału. Jak pokazano na rys. 1, synteza w dziedzinie czasu pozwala uzyskać współczynniki kompresji podobne do tych, jakie dają syntezatory widmowo-parametryczne.

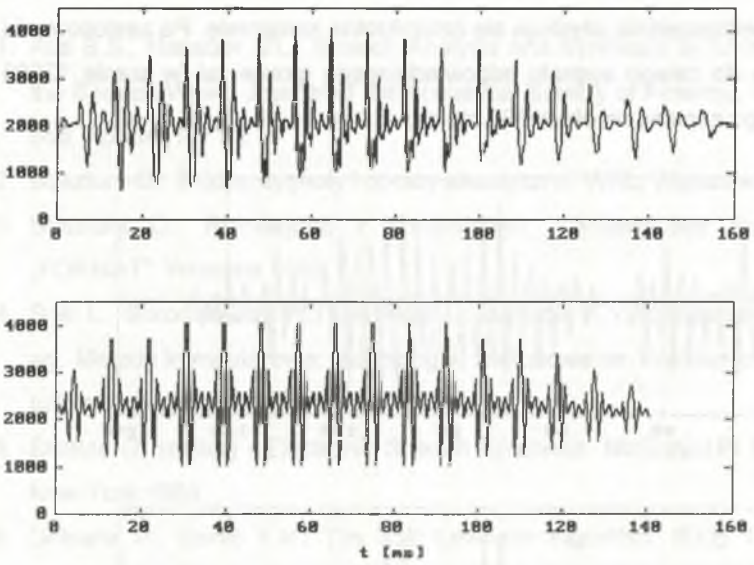
Kompresja sygnału mowy przebiega nieco inaczej dla głosek dźwięcznych i bezdźwięcznych. Wspólną cechą jest segmentacja i decymacja. Dla głosek dźwięcznych jeden segment pokrywa się z okresem pobudzenia (tonu krtaniowego), a dla głosu bezdźwięcznych długość segmentu jest stała i wynosi typowo 256 próbek. Kolejne etapy kompresji głosek dźwięcznych można ująć w następujących punktach:

a) Określenie wzmocnienia wspólnego dla całego segmentu (redukuje się w ten sposób liczbę bitów potrzebną do zapisania amplitud poszczególnych próbek).

b) Dobranie widma fazowego tak, by sygnał czasowy był symetryczny. Okazuje się, że ucho ludzkie jest mało wrażliwe na zmiany w widmie fazowym sygnału mowy. Jeśli więc na podstawie widma segmentu przyjąć, że faza może mieć tylko dwie wartości 0 lub 180 stopni, to po tej operacji sygnał czasowy będzie symetryczny i zmiana ta będzie dla ucha trudna do wykrycia. Ponadto sygnał czasowy dość szybko zanika i dlatego przyjmuje się, że w pierwszej i czwartej ćwiartce segmentu jest on równy zeru.

c) Powtórzenie kilku segmentów. Widmo amplitudowe sygnału mowy zmienia się wolno. Podczas syntezy można więc jeden segment powtórzyć kilkakrotnie, co zapewnia dalszą kompresję.

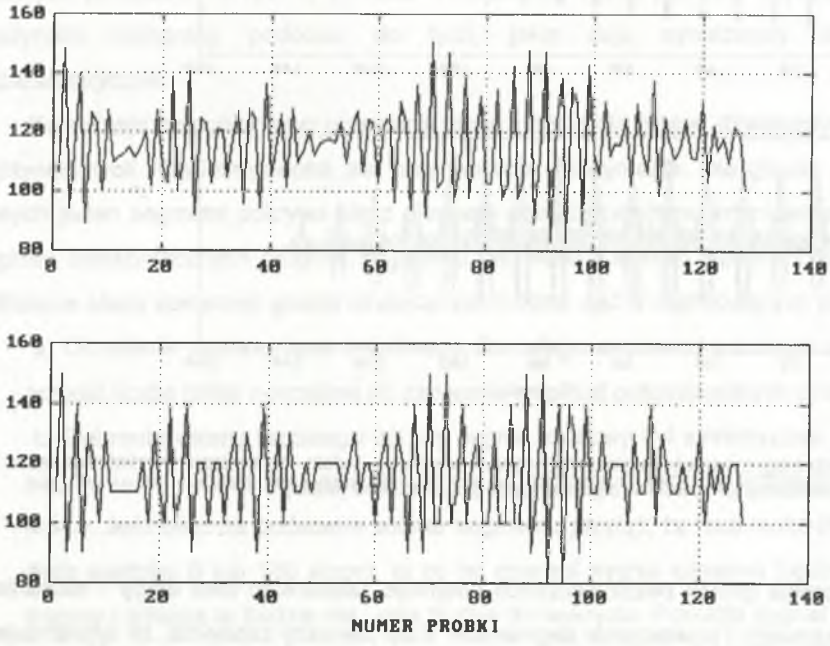
Na rys.15 (górny wykres) pokazano przebieg czasowy głoski /o/ ze słowa /dom/ (1500 próbek 12-bitowych), oraz przebieg otrzymany za pomocą opisanej syntezy w dziedzinie czasu dla tej samej głoski, którą wcześniej zakodowano na 170 bitach (dolny wykres). Przy odtwarzaniu uzyskany dźwięk jest zrozumiały, chociaż słychać lekkie brzęczenie. Jest to spowodowane prawdopodobnie tym, że przyjęto wszystkie segmenty o jednakowej długości ignorując niewielkie zmiany tonu krtaniowego. Zmiany te można uwzględniać dodając zerowe próbki na początku dłuższych segmentów.



Rys. 15. Przebiegi czasowe dla głoski /o/; u góry - oryginalny, u dołu - po syntezie w dziedzinie czasu
Fig. 15. Waveforms for vowel /o/; digitised (upper), synthesised (lower)

Kompresja głosek bezdźwięcznych obejmuje zasadniczo dwa etapy - dobranie widma fazowego i powtarzanie segmentów. Etap pierwszy zapewnia, że sygnał daje się zapisać za pomocą małej liczby bitów. Zostanie to pokazane na 128 próbkach 8-bitowych wybranych z głoski /s/. Przyjmując, że widmo fazowe tego segmentu jest ograniczone do dwóch wartości (0 i 180 stopni), otrzymamy przebieg jak na rys. 16 (górny wykres). Jak widać, amplitudy próbek gromadzą się wokół ośmiu poziomów, czyli da się zapisać na trzech bitach plus jedno wzmocnienie dla całego segmentu (rys. 16, dolny wykres). Należy zaznaczyć, że inna operacja na widmie może zapewnić lepszą kompresję, gdyż tym razem nie jest wymagana symetria sygnału. Jeśli chodzi o etap drugi, to nie wolno realizować go identycznie jak dla głosek dźwięcznych, ponieważ wprowadzi się do sygnału składową okresową. Zakładając, że widmo sygnału odtworzonego wstecz jest identyczne jak widmo sygnału odtworzonego w przód oraz że widmo części segmentu jest średnio takie samo jak całego segmentu, można 128 próbek powtórzyć według następującego algorytmu: od 1 do 128; od 128 do 1 (czyli ten sam sygnał odtworzony wstecz); od 65 do 128 i od 1 do

64; oraz od 64 do 1 i od 128 do 65. W ten sposób unika się wprowadzenia składowej okresowej i jednocześnie uzyskuje się czterokrotną kompresję. Po zastosowaniu opisanej metody do całego sygnału odpowiadającego głosce /s/ (w sumie 12300 bitów) udało się go zapisać na około 300 bitach.



Rys. 16. Fragment przebiegu dla głoski /s/; u góry - po dopasowaniu widma, u dołu - po kwantyzacji do 3 bitów
Fig. 16. Waveforms for sound /s/; 3-bit level matching (upper), and quantized

5. Podsumowanie

W artykule przedstawiono algorytmy analizy i syntezy mowy w dziedzinie czasu i częstotliwości. W różnych odmianach są one obecnie powszechnie stosowane w wielu systemach przetwarzania sygnału mowy. Od mniej więcej połowy lat osiemdziesiątych większość nowych prac dotyczy rozpoznawania mowy rozumianego jako: dekodowanie wypowiedzi, interpretacja wypowiedzi lub rozpoznawanie mówcy. Do tego celu próbuje się wykorzystać również bardzo popularne obecnie sieci neuronowe.

LITERATURA

1. Atal B.S., Hanauer S.L.: Speech Analysis and Synthesis by Linear Prediction of the Speech Wave. Journal of the Acoustical Society of America, vol. 50, pp. 637-655, August 1971*).
2. Basztura C.: Źródła, sygnały i obrazy akustyczne. WKŁ, Warszawa 1988.
3. Basztura C.: Rozmawiać z komputerem. Wydawnictwo Prac Naukowych „FORMAT", Wrocław 1992.
4. Bolc L., Borodziewicz W., Cytowski J., Jaszczak K.: Przetwarzanie sygnału mowy. Metody komputerowe, technologia, zastosowanie. Wydawnictwo Uniwersytetu Warszawskiego, Warszawa 1989.
5. Bristow G. (editor) - Electronic Speech Synthesis. McGraw-Hill Book Company, New York 1984.
6. Delsarte P., Genin Y.V.: The Split Levinson Algorithm. IEEE Trans. on ASSP vol.34 no.3, pp.470-478, June 1986.
7. Flanagan J.L.: Automatic Extraction of formant Frequencies from Continuous Speech. Journal of the Acoustical Society of America, January 1956*).
8. Gold B., Rabiner L.: Parallel Processing Techniques for Estimating Pitch Periods of Speech in the Time Domain. Journal of the Acoustical Society of America, vol.46, pp.442-448, August 1969*).
9. Hanson H.M. et al.: A system for finding Speech Formants and Modulations via Energy Separation. IEEE Trans. on Speech and Audio Processing, vol.2 no.3, pp.436-443, July 1994.
10. Kompel S.: Syntezator mowy współpracujący z IBM PC. Praca dyplomowa, Instytut Elektroniki Pol.Śl., Gliwice 1992.
11. Makhoul J.: Linear Prediction: A Tutorial Review. Proc. of the IEEE, vol.63, no.4, pp.561-580, April 1975.
12. Markel J.D.: Digital Inverse Filtering - A New Tool for Formant Trajectory Estimation. IEEE Trans. on Audio Electr., vol.20, pp.129-137, June 1972*).
13. Markel J.D.: The SIFT Algorithm for Fundamental Frequency Estimation. IEEE Trans. on Audio Electr., vol.20, pp.367-377, December 1972*).
14. McCandless S.S.: An Algorithm for Automatic Formant Extraction Using Linear Prediction Spectra. IEEE Trans. on ASSP vol.22, pp.135-141, April 1974*).

15. Noll A.M.: Cepstrum Pitch Determination. Journal of the Acoustical Society of America, vol.41, pp.293-309, February 1967*).
16. Oppenheim A.V.: Sygnały cyfrowe. Przetwarzanie i zastosowania. WNT, Warszawa 1982.
17. Oppenheim A.V., Schafer R.W.: Cyfrowe przetwarzanie sygnałów. WKŁ, Warszawa 1979.
18. Rabiner L.R. et al.: A Comparative Performance Study of Several Pitch Detection Algorithm. IEEE Trans. on ASSP vol.24, pp.399-417, October 1976*).
19. Rabiner L.R., Gold B: Theory and Application of Digital Signal Processing. Prentice-Hall, Inc. New Jersey 1975.
20. Ross M.J. et al.: Average Magnitude Difference Function Pitch Extractor. IEEE Trans. on ASSP vol.22, pp.353-362, October 1974*).
21. Schafer R.W., Rabiner L.R.: System for Automatic Formant Analysis of Voiced Speech. Journal of the Acoustical Society of America, vol.47, pp.634-648, February 1970*).
22. Snell R.C., Milinazzo F.: Formant Location from LPC Analysis Data. IEEE Trans. on Speech and Audio Processing, vol.1, no.2, pp.129-134, April 1993.
23. Soandhi M.M.: New Methods of Pitch Extraction. IEEE Trans. on Audio Electr., vol.16, pp.262-266, June 1968*).
24. Tadeusiewicz R.: Sygnał mowy. WKŁ, Warszawa 1988.
25. Willems L.F.: Rubust Formant Analysis for Speech Synthesis Applications. Manuscript no.616, 1988.
26. ILS-IEEE Interactive Laboratory System for IBM Personal Computers, Signal Technology, Inc.

*)Artykuły te wydrukowano także w „Speech Analysis" edited by Schafer R.W., Markel J.D., IEEE Press, New York 1979.

Recenzent: Prof.dr hab.inż. Ryszard Tadeusiewicz

Wpłynęło do Redakcji 15.10.1994 r.

Abstract

The paper describes one of the most popular for digital speech signal processing called synthesis by analysis. This method includes frequency domain synthesis (formant analysis/synthesis, linear predictive coding) and time domain synthesis. The schemes of the first one are based on human speech modelling as timevarying filter excited by noise source or pulse source. Data compression is achieved through storing the parameters of excitation and filter in place of the original waveform. By contrast, in time-domain synthesis, a compressed representations of waveform as a function of time are stored. The useful remarks for application of these methods and speech sythesis examples are included.