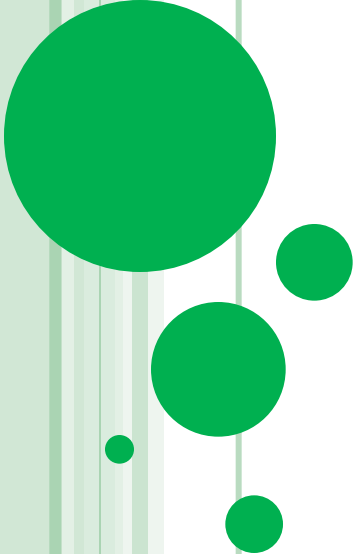


Akustyka mowy

Synteza mowy



mgr inż. Kuba Łopatka
Katedra Systemów Multimedialnych
klopatka@sound.eti.pg.gda.pl
pok. 628, tel. (348) 63-32

PLAN WYKŁADU

- Pojęcie i ogólny schemat działania syntezy mowy
- Analiza językowa i fonetyczna
- Podejścia do syntezy sygnału mowy
- Synteza formantowa
- Synteza artykulacyjna
- Synteza konkatenacyjna
- Modelowanie prozodii
- Zastosowania i przykłady syntezy mowy

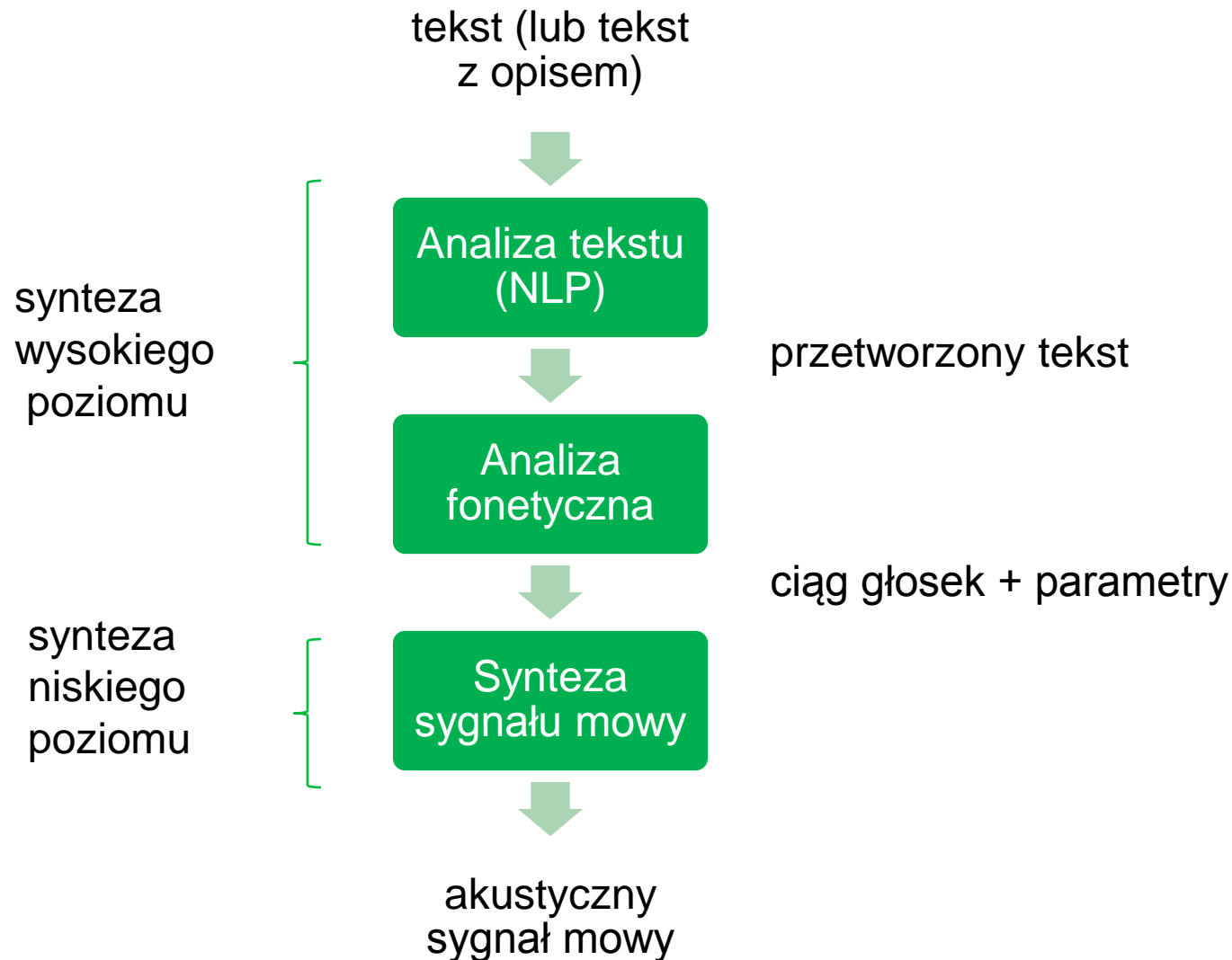
SYNTEZA MOWY

Synteza mowy – (ang. TTS - *Text-To-Speech*) – zamiana tekstu w formie pisanej na sygnał akustyczny, którego brzmienie naśladuje brzmienie ludzkiej mowy.

Podstawowe cele syntezy to:

- zrozumiałość treści wypowiedzi,
- naturalność brzmienia.

SCHEMAT DZIAŁANIA SYSTEMU TTS



ANALIZA JĘZYKOWA

Pierwszy etap przetwarzania – analiza tekstu. W analizie wykorzystywane są metody z dziedziny przetwarzania języka naturalnego (ang. *Natural Language Processing – NLP*). Zadania wchodzące w skład analizy tekstu wejściowego:

- normalizacja tekstu,
- analiza morfologiczna,
- analiza syntaktyczna,
- analiza semantyczna,
- analiza prozodyczna.

ANALIZA JĘZYKOWA



ANALIZA FONETYCZNA

Zamiana wypowiedzi dostępnej w formie tekstowej na ciąg fonemów.

- uwzględnienie zjawisk fonetycznych obowiązujących w języku (np. utrata dźwięczności, wygłos)
- wyjątki fonetyczne (np. marznąć) i słowa obce → słownik

Należy przyjąć standard opisu głosek (np. alfabet SAMPA, IPA, AS).

KOLEJNE ETAPY PRZETWARZANIA

Zosia dała Stefanowi 5,50 zł.

normalizacja:

zosia dała stefanowi pięć złotych pięćdziesiąt groszy

analiza morfologiczna:

zo·sia da·ła ste·fa·no·wi pięć zło·tych pięć·dzie·siąt gro·szy

analiza prozodyczna:

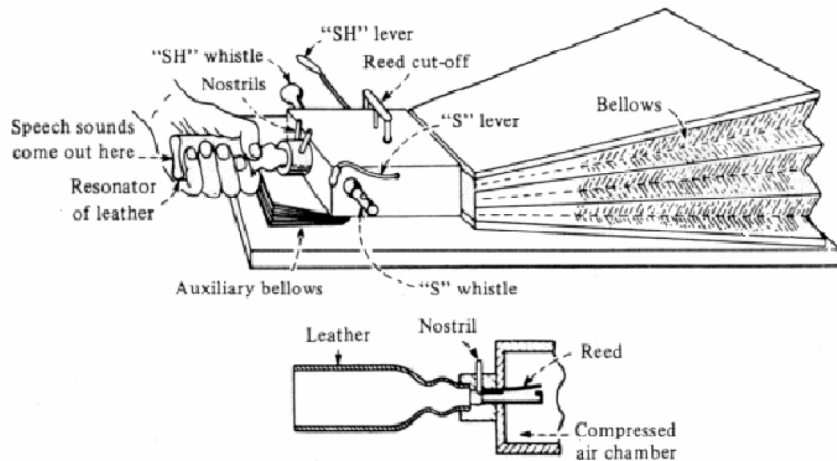
zo sia da ła ste fa no wi
pięć zł o tych pięć dzie siąt gro szy

analiza fonetyczna:

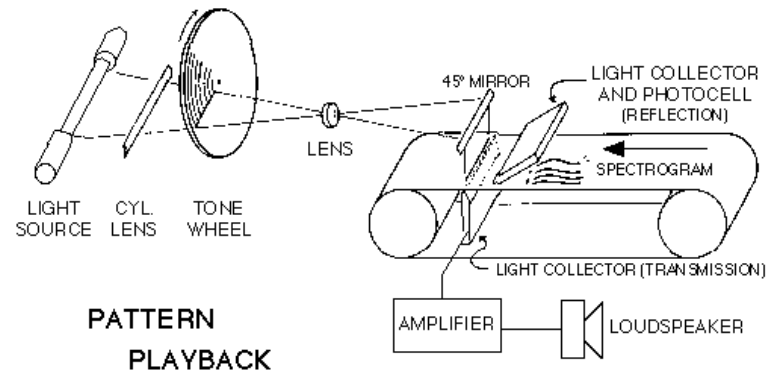
zośadałastefanowipjęźzłotyhpjeńżeśódgrošy

HISTORIA

Pierwsze syntetyzery – mechaniczne (von Kempelen 1791)



Pattern Playback – 1950 – maszyna „czytająca” spektrogram



Pierwszy syntetyzer formantowy – 1964 r.

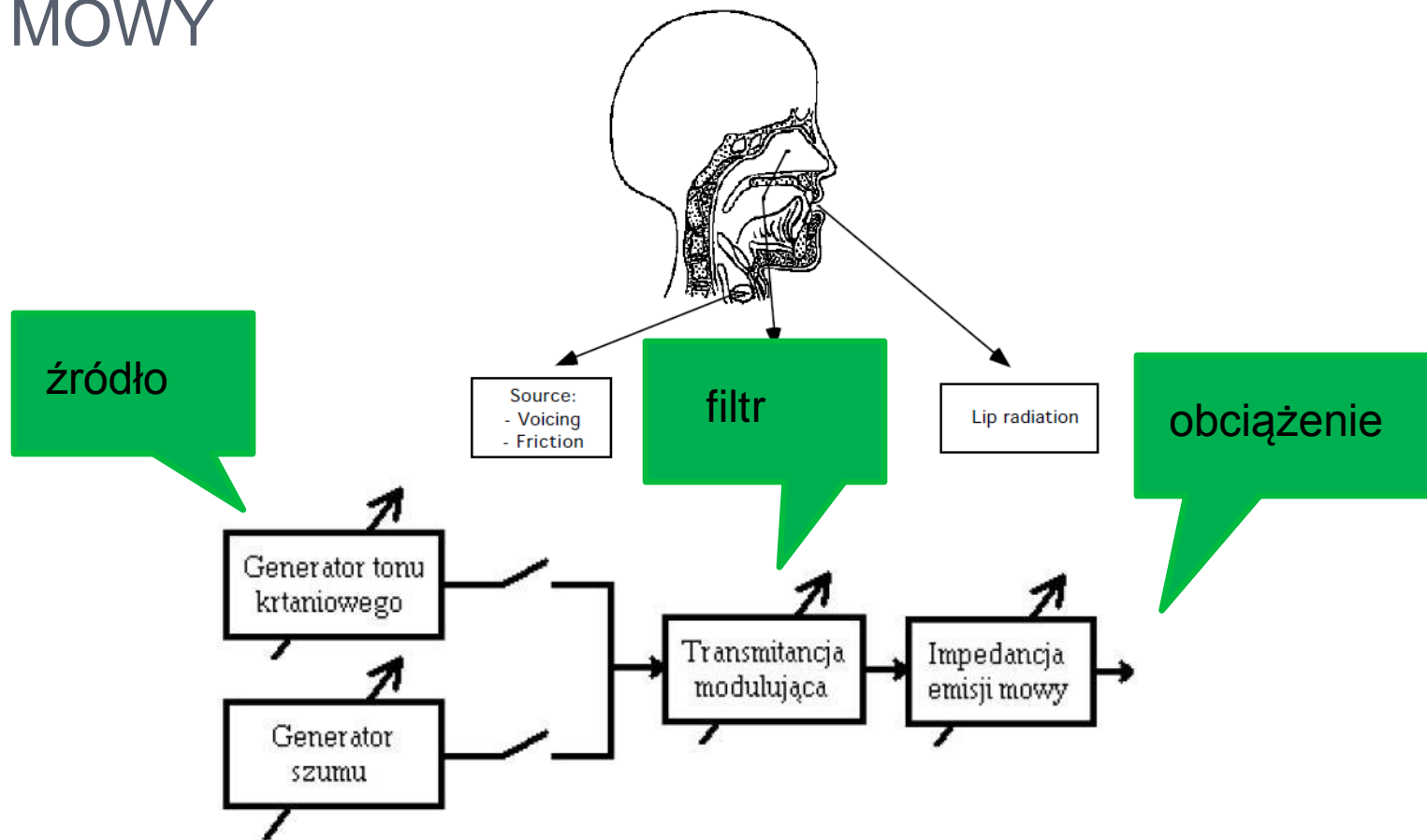
Później – synteza artykulacyjna i konkatencyjna

SYNTEZA SYGNAŁU MOWY

Można wyróżnić 3 podstawowe podejścia do syntezy sygnału mowy:

- Odwzorowanie widma sygnału mowy – metoda formantowa, synteza LPC;
- Fizyczne odwzorowanie mechanizmów wytwarzania mowy – metoda artykulacyjna;
- Wykorzystanie nagranych próbek sygnału mowy – metoda konkatencyjna.

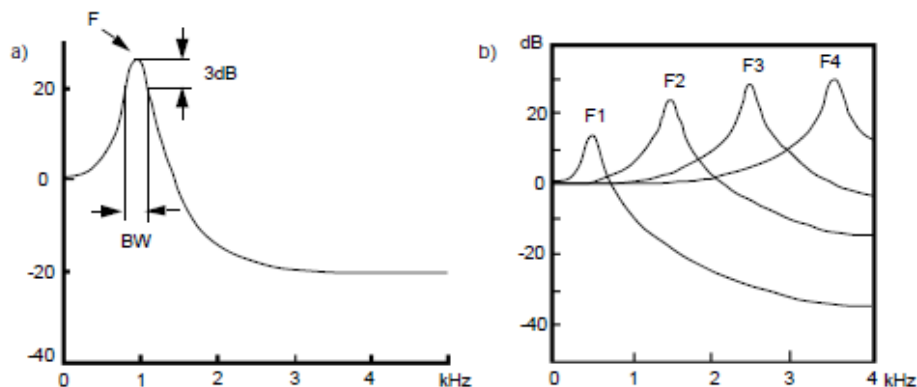
SCHEMAT ZASTĘPCZY WYTWARZANIA MOWY



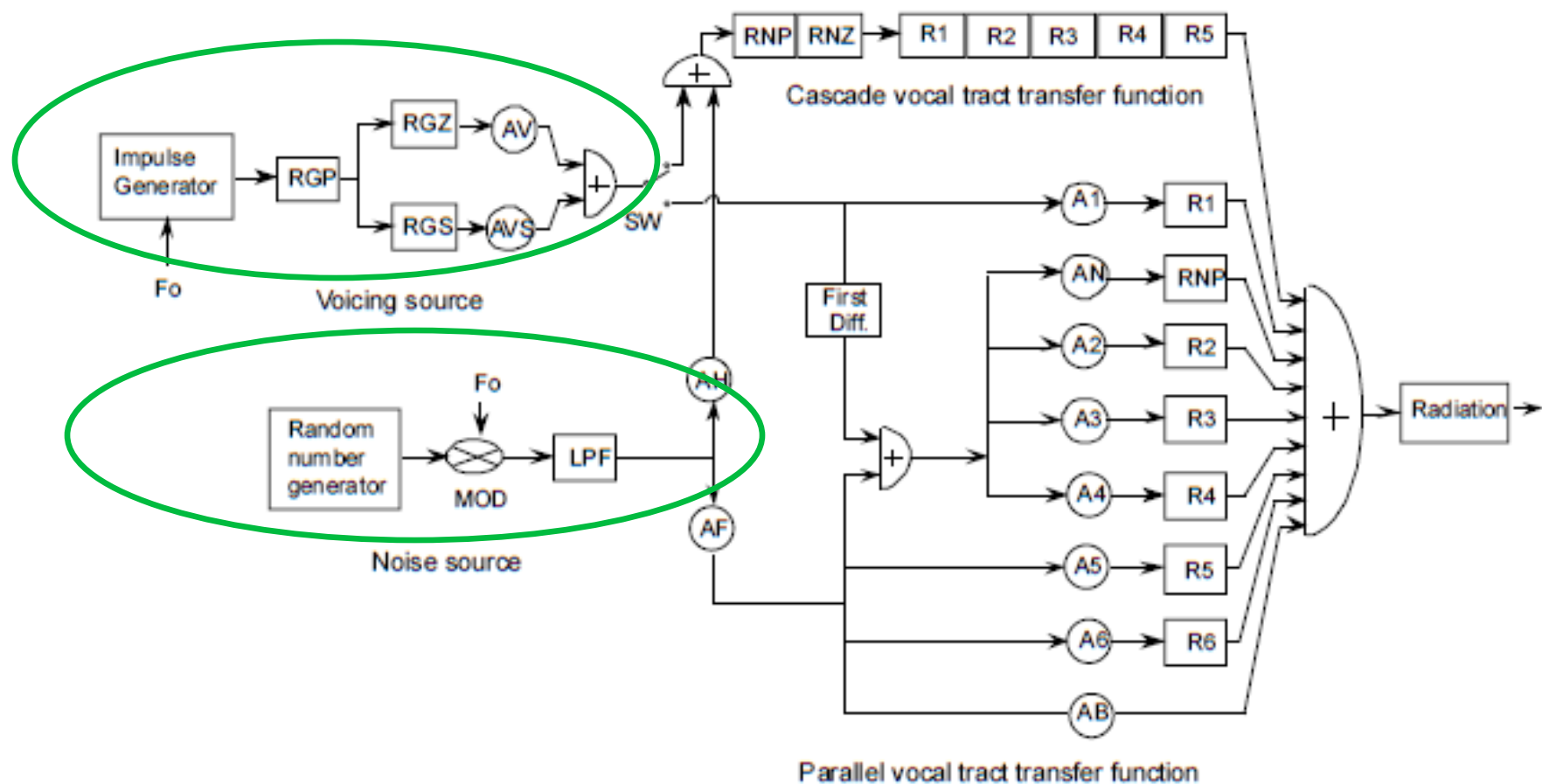
Najpopularniejszą realizację pierwszego podejścia do syntezy jest zastosowanie **modelu źródło-filtr** w celu odwzorowania charakteru widmowego sygnału mowy.

METODY SYNTEZY

Synteza formantowa – modelowanie traktu głosowego jako połączenie rezonatorów – filtrów elektrycznych (LC) lub cyfrowych. Łączna charakterystyka częstotliwościowa układu filtrów ma być zbliżona do charakterystyki aparatu mowy człowieka. Podejście to ma w założeniu odwzorować formantowy charakter sygnału mowy.

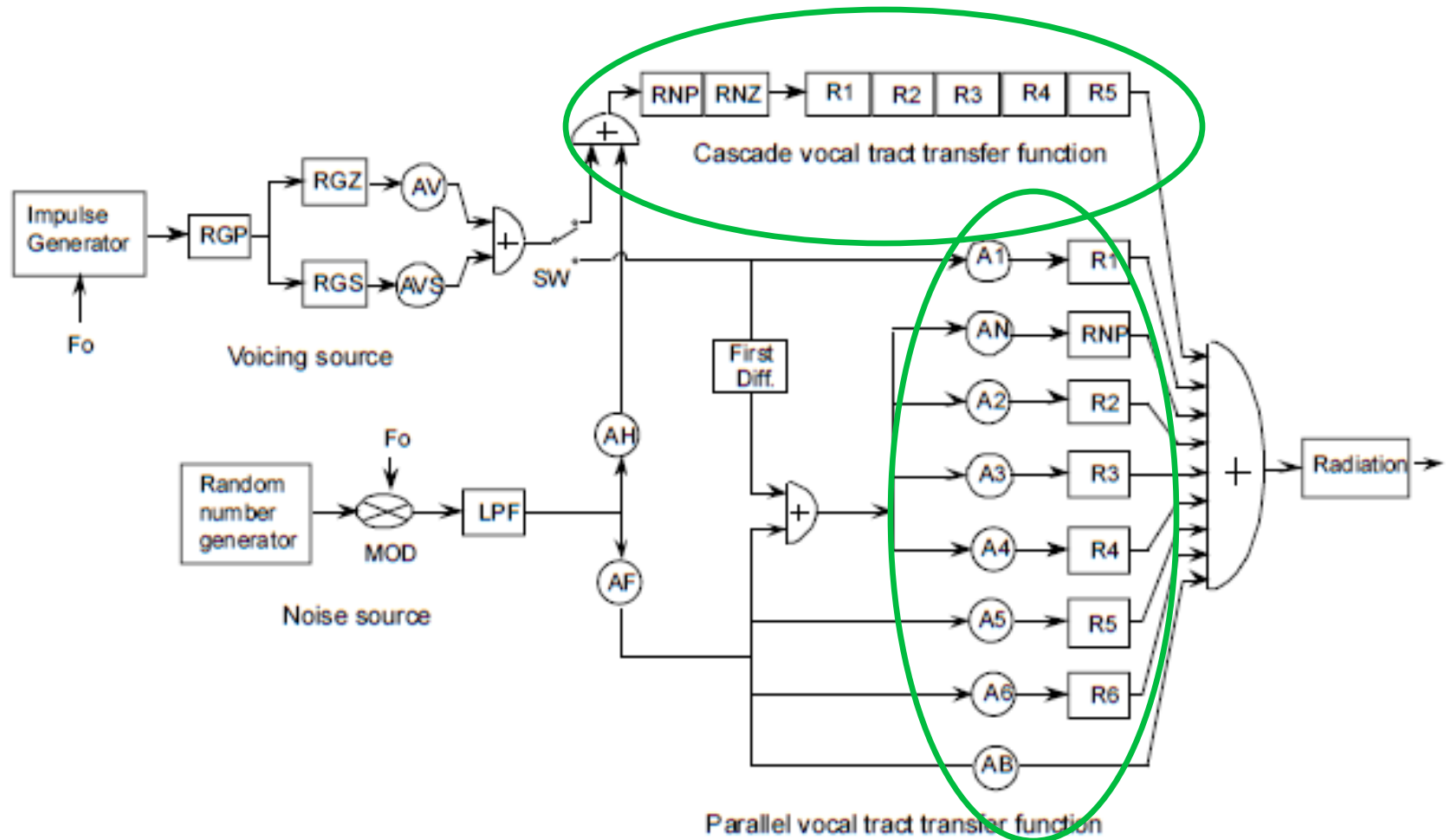


PRZYKŁAD SYNTETYZERA FORMANTOWEGO



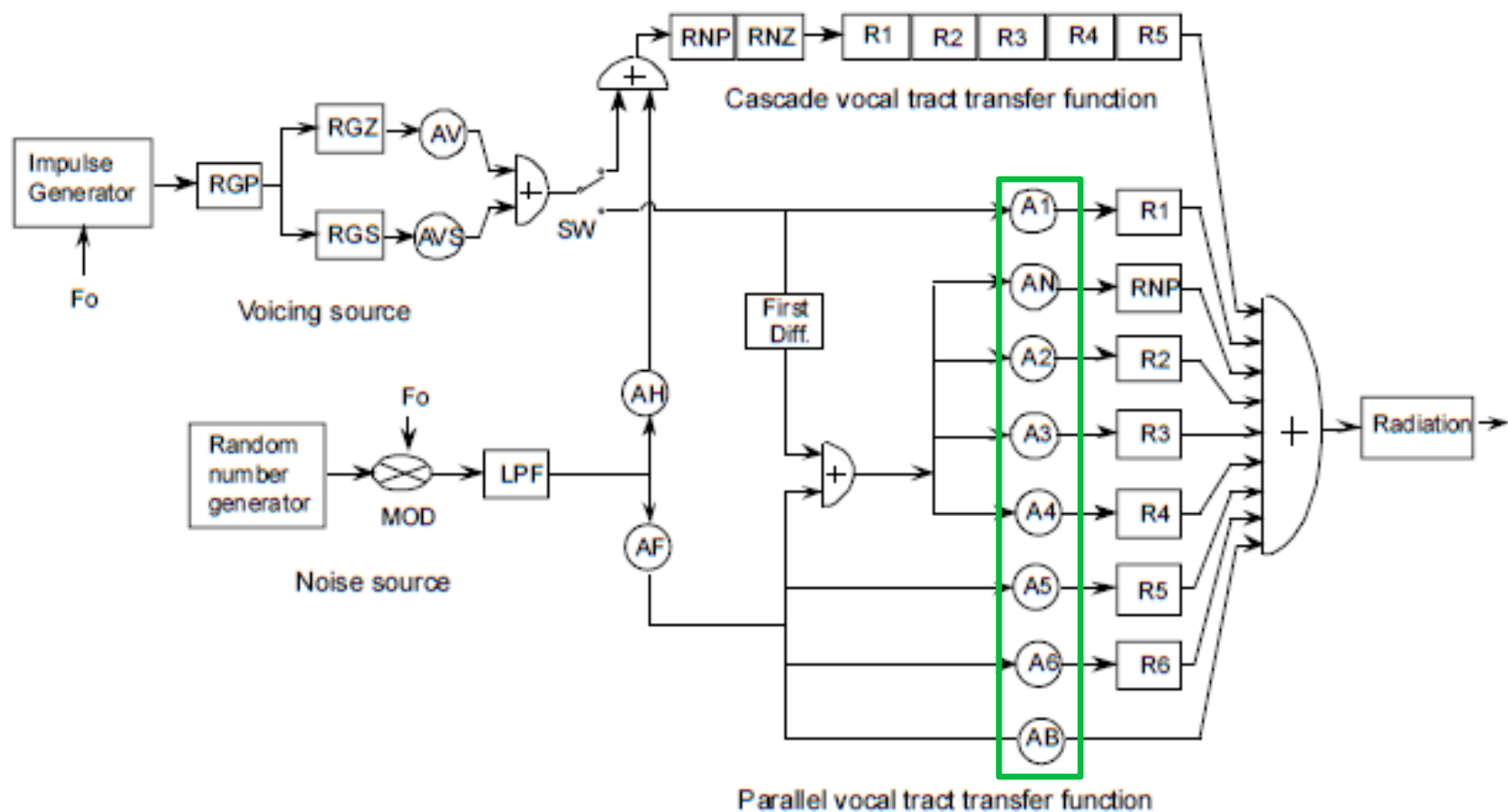
Dwa rodzaje pobudzenia: tonalne (dla głosek dźwięcznych) i szumowe (dla bezdźwięcznych i trących)

PRZYKŁAD SYNTETYZERA FORMANTOWEGO



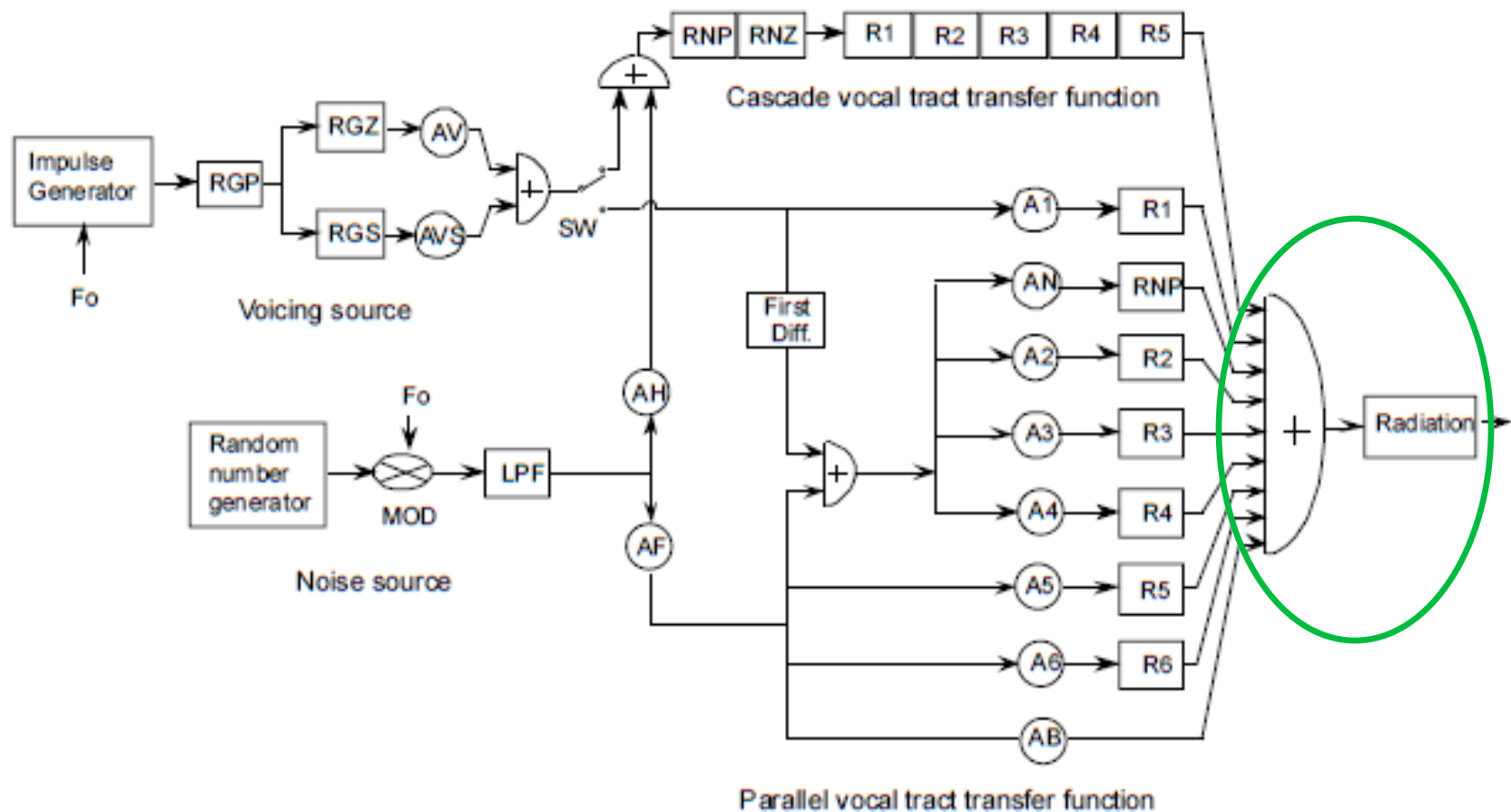
Układ filtrów (rezonatorów) połączonych równolegle bądź kaskadowo.

PRZYKŁAD SYNTETYZERA FORMANTOWEGO



Parametrami są wzmacnienia, częstotliwości środkowe i szerokości pasm filtrów modelujących formanty.

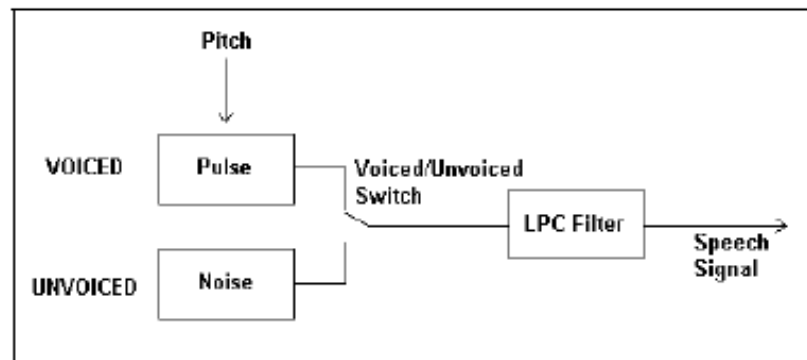
PRZYKŁAD SYNTETYZERA FORMANTOWEGO



Połączenie filtrów tworzy łączną charakterystykę traktu głosowego, obciążoną dodatkowo impedancją emisji mowy.

METODY SYNTEZY

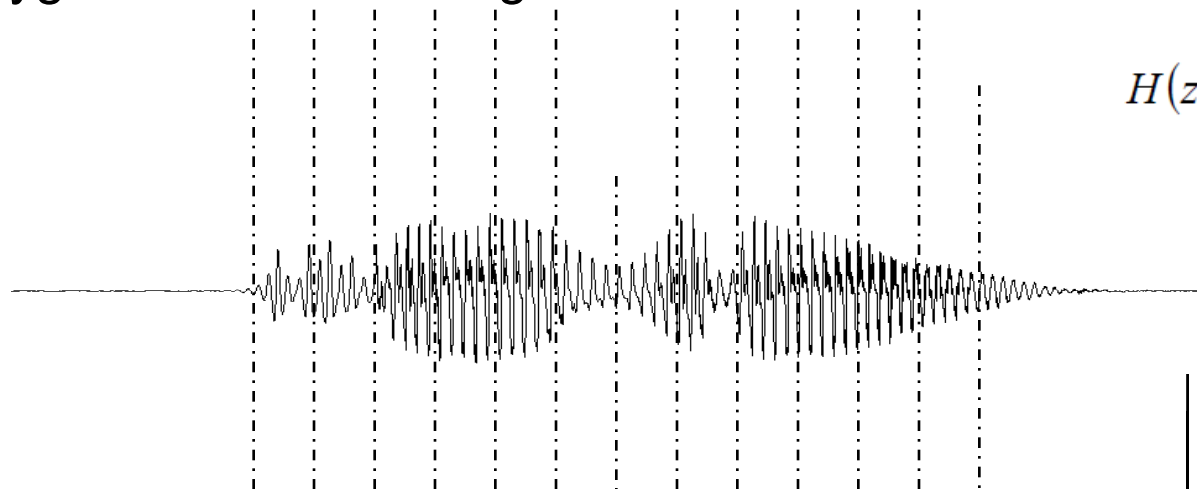
- Synteza LPC – wykorzystuje liniowe kodowanie predykcyjne (ang. LPC – *linear predictive coding*) do odwzorowania charakterystyki przenoszenia traktu głosowego. Metoda LPC pozwala na rozbitcie sygnału mowy na pobudzenie i transmitancję traktu głosowego, modelowaną przez filtr biegunowy (*all-pole filter*).



METODY SYNTEZY

○ LPC – przypomnienie

Liniowe kodowanie predykcyjne (ang. *Linear Predictive Coding* – LPC) – technika analizy sygnału mowy polegająca na przedstawieniu sygnału mowy jako odpowiedzi filtru typu biegunowego (*all-pole filter*) na sygnał tonu krztaniowego.



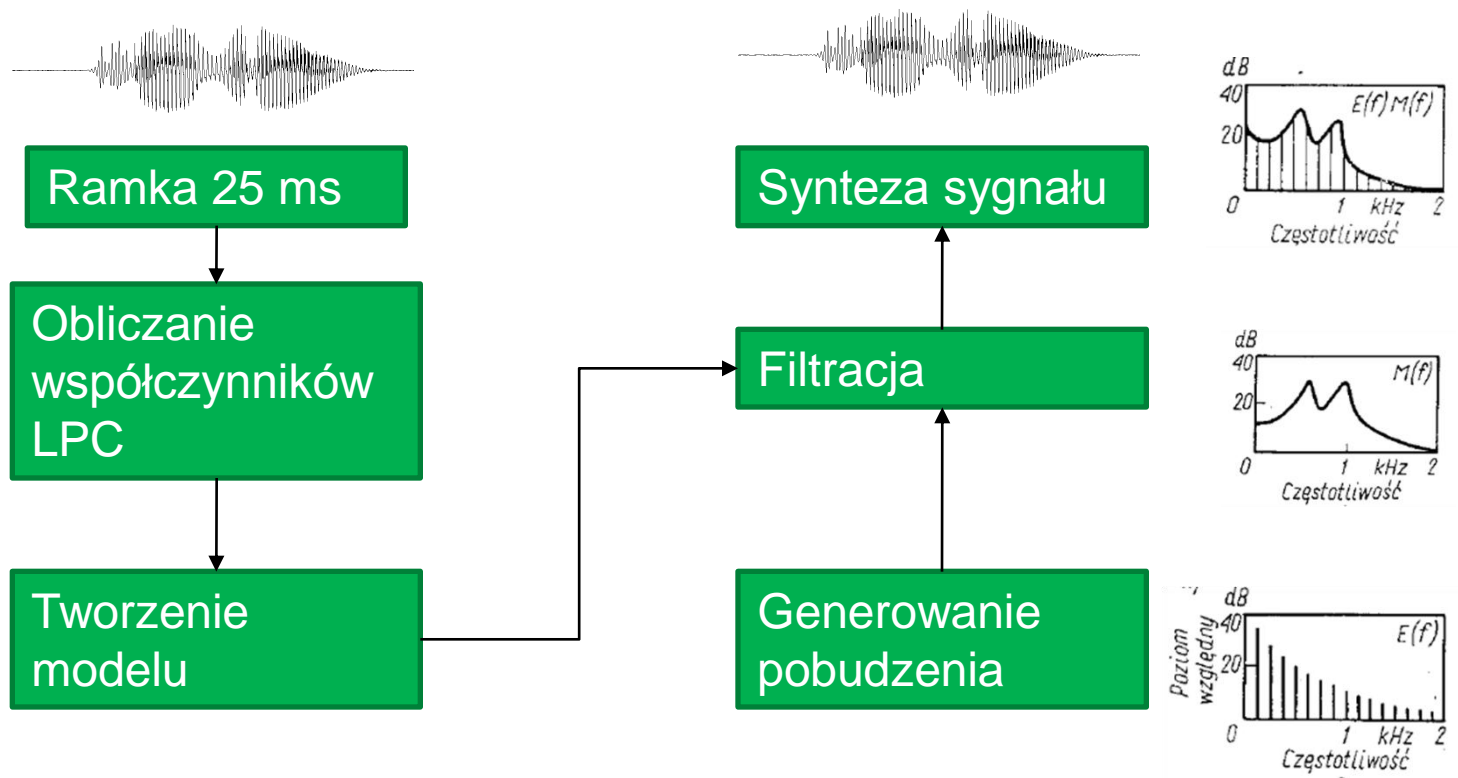
$$H(z) = G \cdot \frac{1}{1 - \sum_{k=1}^p a_k \cdot z^{-k}}$$

Analiza
LPC

$\begin{bmatrix} a_1 \\ a_2 \\ \dots \\ a_p \end{bmatrix}$

METODY SYNTEZY

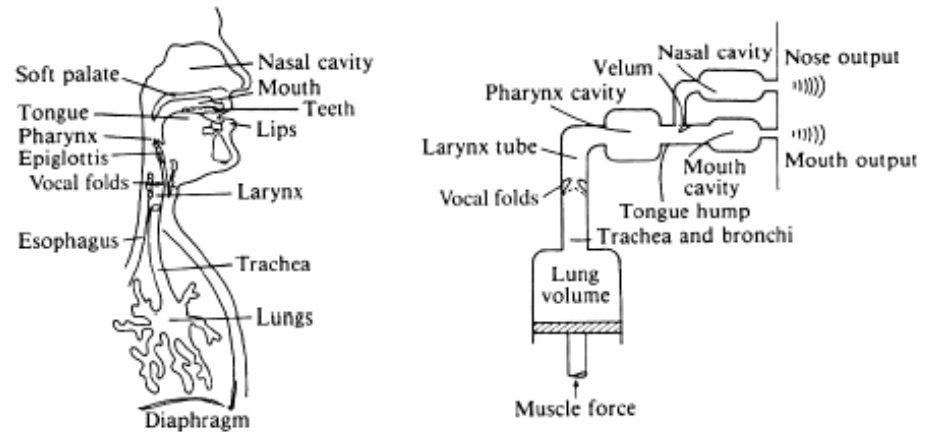
Synteza LPC – schemat działania



METODY SYNTEZY

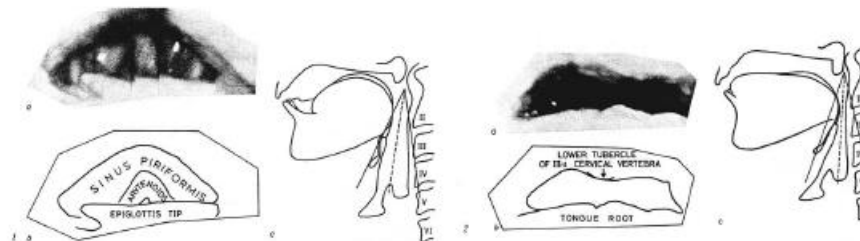
Synteza artykulacyjna – polega na fizycznym odwzorowaniu mechanizmu generowania dźwięków mowy. Wykorzystując modelowanie matematyczne, uwzględnia się zjawiska zachodzące podczas przenoszenia dźwięku przez trakt głosowy. Charakter generowanego sygnału zmienia się w zależności od parametrów, takich jak wymiary i ustawienia poszczególnych organów mowy. Metoda jest w założeniu wierniejsza od formantowej, ale dalece bardziej skomplikowana.

METODY SYNTEZY



Synteza artykulacyjna

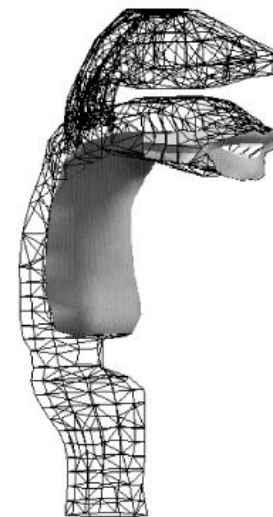
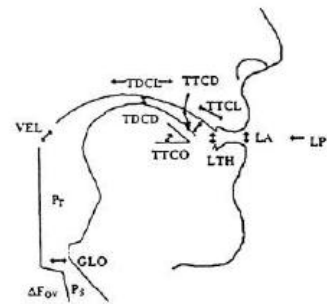
- o modelowanie geometrii traktu głosowego
- o pozyskanie parametrów na drodze analizy przekroju traktu głosowego, rezonansu magnetycznego itp.



METODY SYNTEZY

Przykładowe parametry dwuwymiarowego modelu artykulacyjnego:

| | |
|-----------------|---|
| LP | lip protrusion |
| LA | lip aperture |
| TDCL | tongue dorsum constrict location |
| TDCD | tongue dorsum constrict degree |
| LTH | lower tooth height |
| TTCL | tongue tip constrict location |
| TTCD | tongue tip constrict degree |
| TTCO | tongue tip constrict orientation |
| VEL | velic aperture |
| GLO | glottal aperture |
| Ps | subglottal pressure |
| Pt | transglottal pressure |
| ΔF_{0v} | change in virtual fundamental frequency |



Trójwymiarowy model traktu głosowego →

METODY SYNTEZY

Synteza konkatenacyjna – łączenie (konkatenacja) wypowiedzi z nagranych fragmentów głosu lektora (segmentów) zawierających słowa, sylaby lub złączenia głosek. Jest to obecnie najczęściej spotykana metoda syntezy, dająca wysoką zrozumiałość i naturalność brzmienia. Dla poprawnego działania konkatenacyjnego systemu TTS konieczne jest zebranie bazy segmentów obejmujących cały system fonetyczny języka.

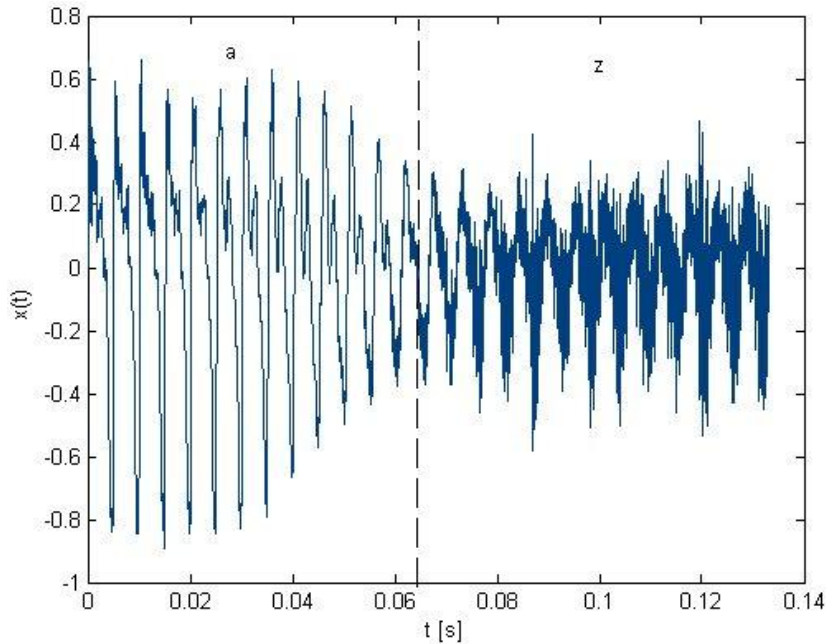
WYBÓR SEGMENTÓW

Segmenty możliwe do wykorzystania w syntetyzerze konkatencyjnym:

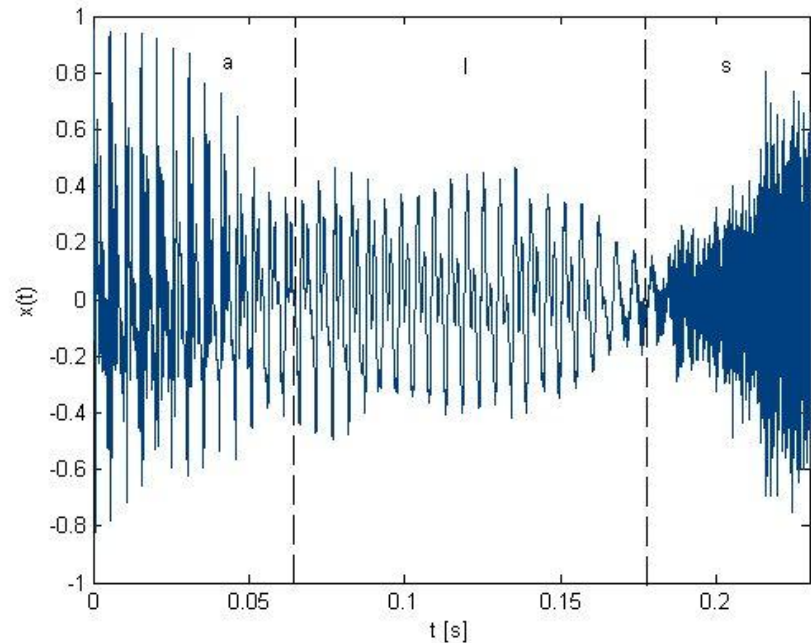
- fonem (głoska),
- difon,
- trifon,
- sekwencja fonemów,
- półsylaba
- sylaba,
- wyraz,
- zdanie.

dłuższe segmenty → lepsza jakość → obszerniejsza baza

PRZYKŁADOWE SEGMENTY



difon – połączenie dwóch głosek
liczba difonów w j. polskim – $37^2=1369$



trifon – połączenie 3 głosek
liczba trifonów – $37^3=50653$

DIFONY

Brzmienie głoski jest bardzo mocno zależne od głosek poprzednich i następnych (koartykulacja).
Difony zawierają przejście między dwoma głoskami wraz ze stanami ustalonymi obu głosek.

Składanie wypowiedzi z difonów:

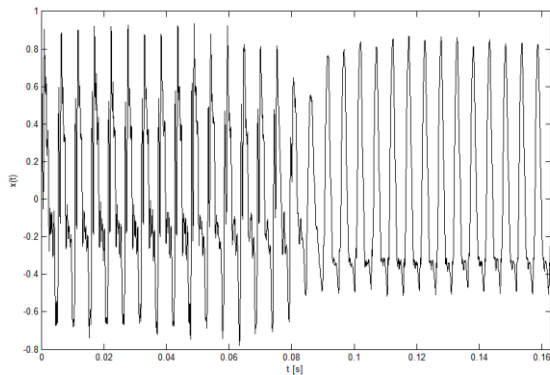
Zosia dała Stefanowi 5,50 zł.

#-z,z-o,o-ś,ś-a,a-d,d-a,a-ł,ł-a,a-s,s-t,t-e,e-f,f-a,a-n,n-
o,o-w,w-i,i-p,p-j,j-ę,ę-dź,dź-z,z-ł,ł-o,o-t,t-y,y-h,h-p,p-
j,j-e,e-ń,ń-dź,dź-e,e-ś,ś-ą,ą-d,d-g,g-r,r-o,o-sz,sz-y,y-
#

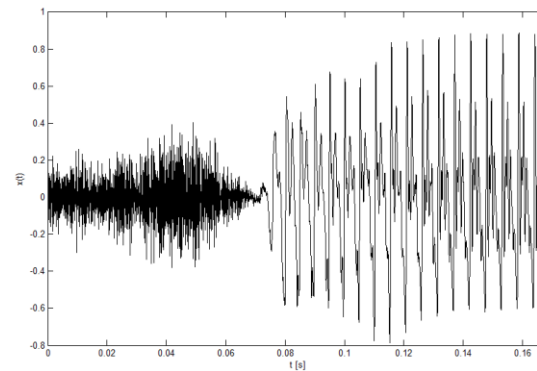
DIFONY

Przykłady difonów

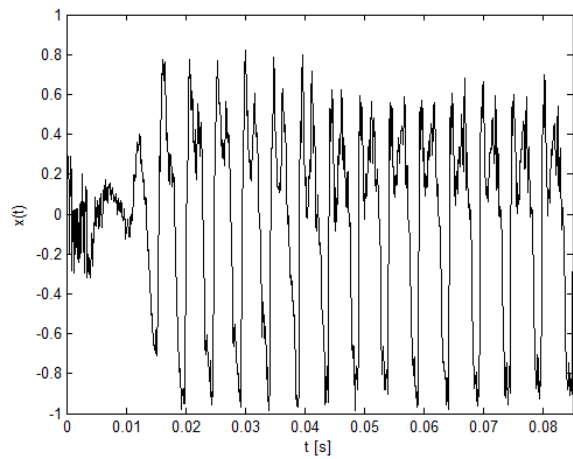
a-m



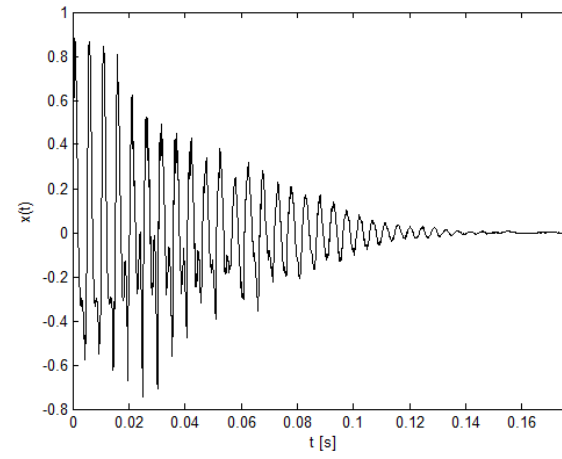
SZ-O



t-e



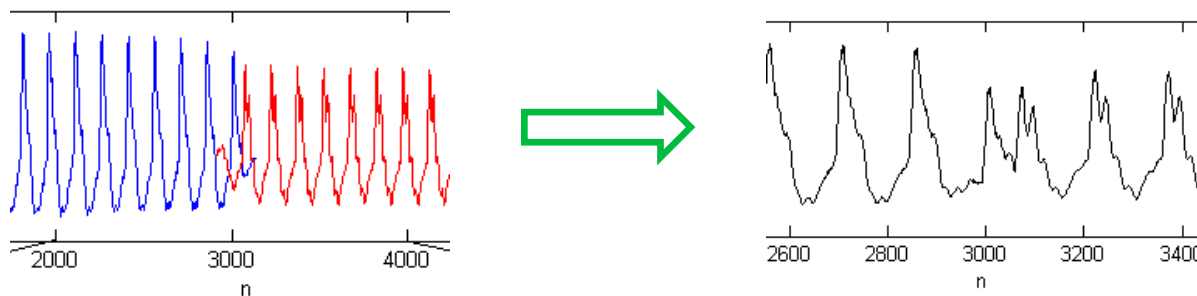
n-#



DIFONY

- Fazy początkowe i końcowe

dla optymalnego połączenia difonów fazy początkowe i końcowe difonów (dla dźwięcznych głosek) powinny być zgodne.



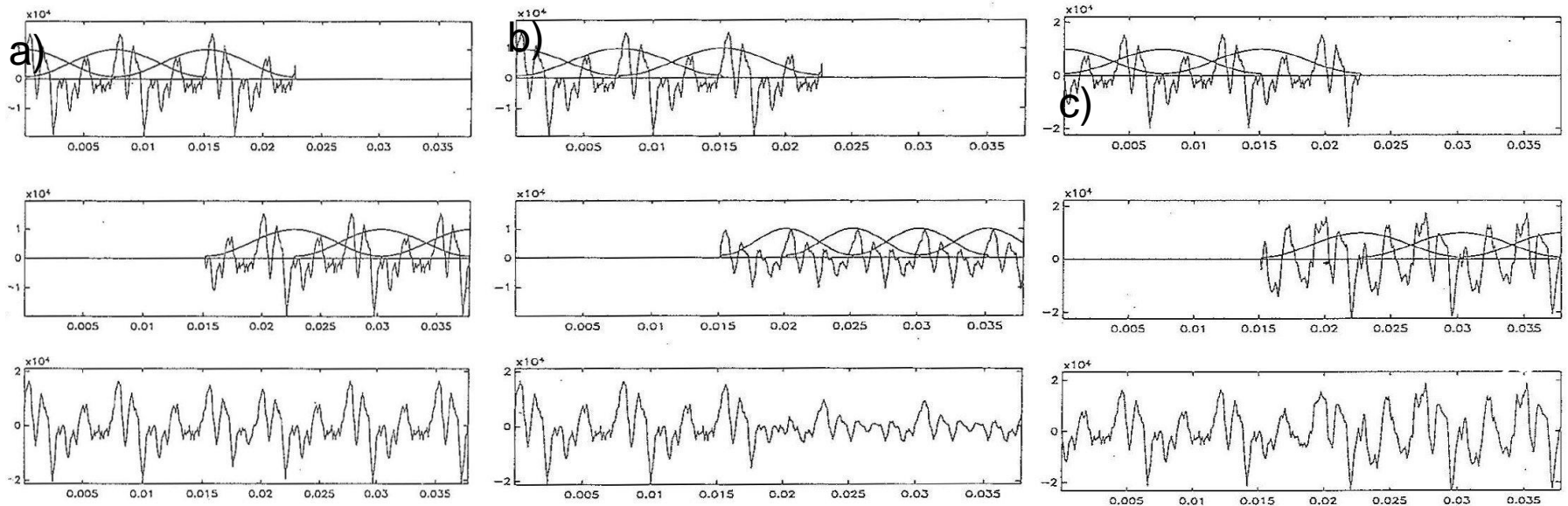
- Próbką przejścia

dla poprawnego połączenia difonów konieczna jest znajomość próbki, na którą przypada przejście między fonemami.

DIFONY - NIEDOPASOWANIE

Po połączeniu difonów możliwe jest niedopasowanie:

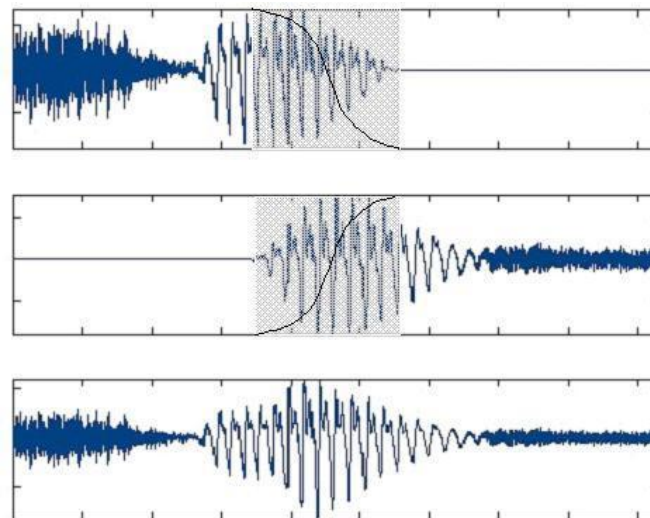
- a) fazy (różne fazy)
- b) tonu podstawowego (różna wysokość)
- c) obwiedni widmowej (różne brzmienia głosek)



KONKATENACJA

Metody konkatencji difonów:

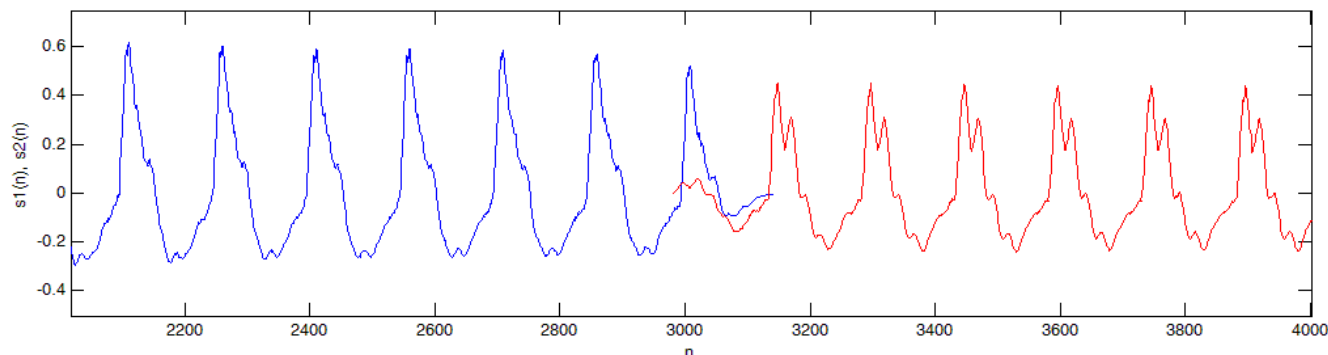
- o przemiksowanie (*cross-fade*)
- zapewnia przejście gładkie pod względem barwy (naturalna interpolacja)
- zmiana zakładki powoduje zmianę tempa wypowiedzi
- możliwe problemy z niedopasowaniem fazy
- przy różnych okresach podstawowych sąsiednich difonów występuje dwugłos



KONKATENACJA

Metody konkatencji difonów:

- PSOLA (*Pitch-Synchronous OverLap and Add*) – połączenie zgodnie z okresem podstawowym. Kolejny difon jest dołączany w miejscu, gdzie rozpoczynałby się kolejny okres podstawowy sygnału. Zapewnia ciągłość tonu podstawowego i w przypadku zgodności faz początkowych i końcowych – również ciągłość fazy.



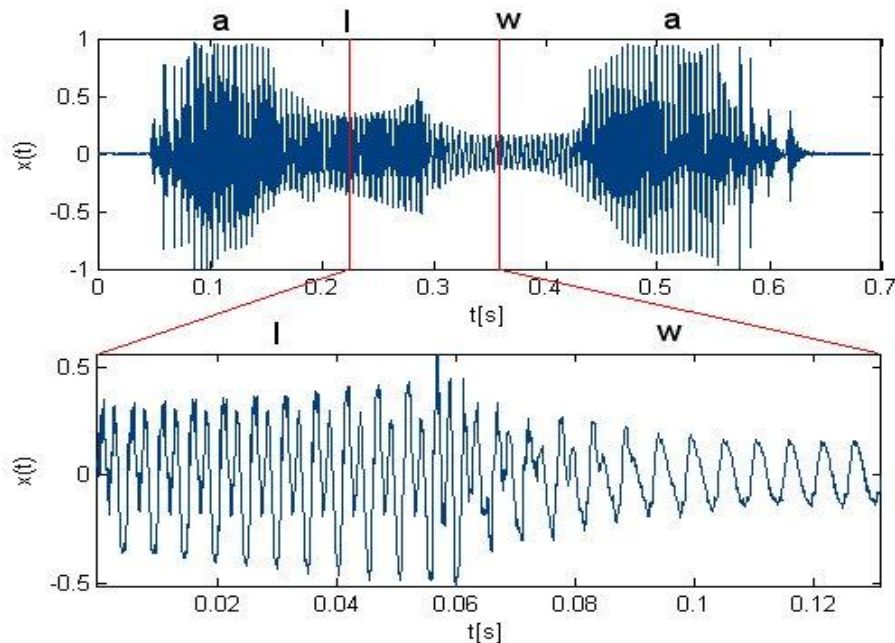
KONKATENACJA

Metody konkatencji difonów:

- MBROLA (*Multi-Band Resynthesis OverLap and Add*) – stosowanie dodatkowego przetwarzania na bazie segmentów (offline) w celu uzyskania lepszego dopasowania łączonych difonów. Dzięki przetwarzaniu zapewnione jest dopasowanie tonu podstawowego, fazy i obwiedni widmowej difonów.

DIFONY – NAGRANIE I EKSTRAKCJA

Nagranie difonów – konieczny materiał językowy zawierający wszystkie połączenia głosek. Możliwe wykorzystanie logatomów – jednostek pozbawionych znaczenia. Należy zwrócić uwagę na równomierną barwę głosu i wysokość tonu.



DIFONY - NAGRANIE

Materiał językowy:

- same difony – zbyt trudne do wymówienia
- wyrazy zawierające difony – niebezpieczeństwo akcentowania ale łatwiejsza wymowa
- logatomy – wyrazy pozbawione znaczenia, ułatwia „automatyczne” czytanie przez lektora.

| difon | logatom |
|-------|---------|
| #-j | |
| j-e | jej |
| e-d%x | |
| d%x-e | dzędź |
| e-m | mem |
| m-y | mym |
| y-d | dyd |
| d-o | dod |
| o-m | mom |
| m-u | mum |

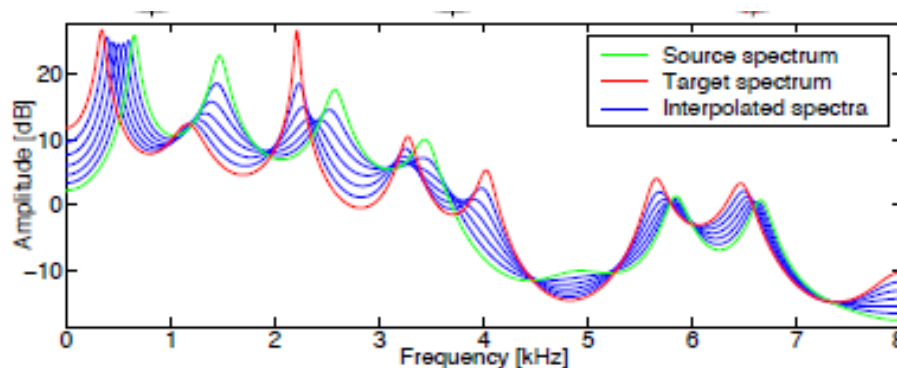
WYGŁADZANIE WIDMOWE DIFONÓW

Sąsiednie difony mają różne brzmienie → niedopasowanie obwiedni widma.

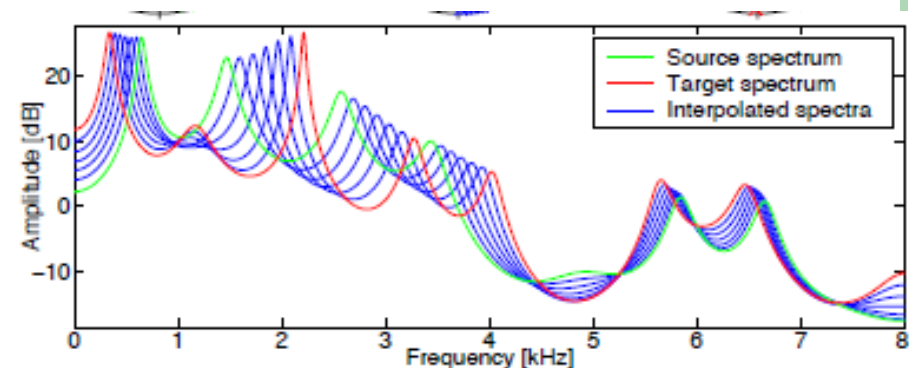
Celem wygładzania widmowego jest znalezienie pośrednich obwiedni widm, interpolujących widma dwóch sąsiadujących difonów.

Ważne jest zachowanie częstotliwości formantowych

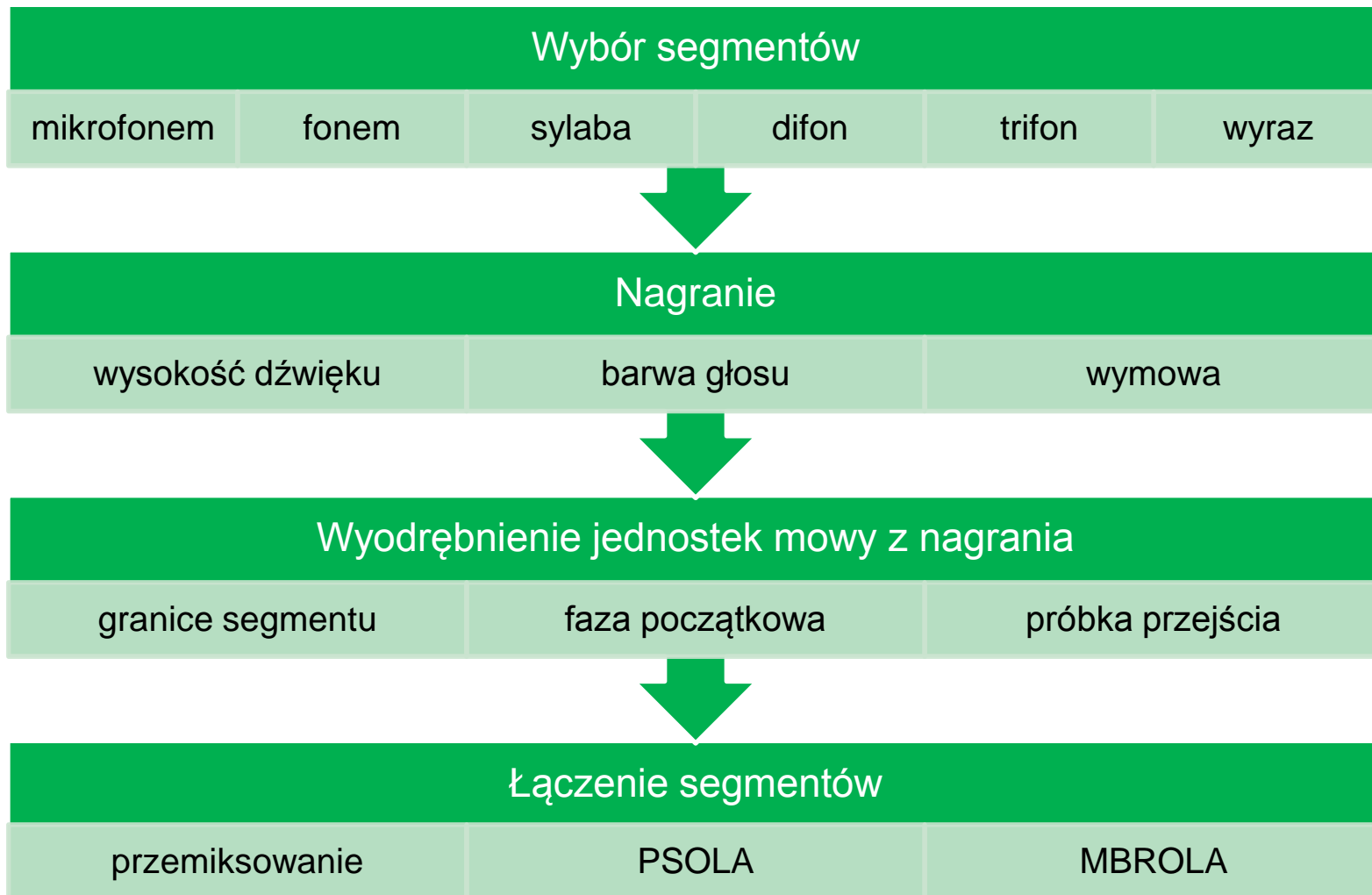
X Źle interpolowana obwiednia widma



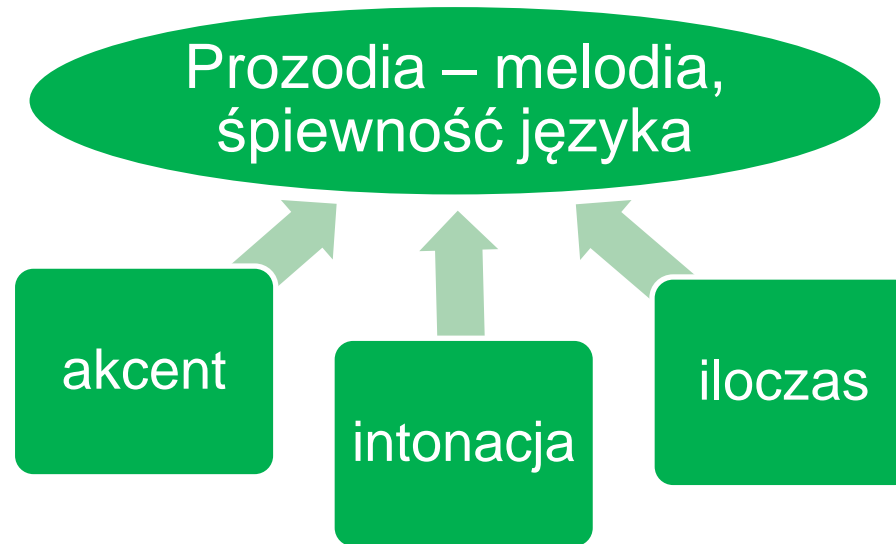
✓ Dobrze interpolowana obwiednia widma



SYNTEZA KONKATENACYJNA „W PIGUŁCE”



KSZTAŁTOWANIE PROZODII



Odwzorowanie prozodii jest konieczne dla naturalnego brzmienia syntetyzowanego sygnału. Bez jej kształtowania synteza brzmi jak „głos robota”.

KSZTAŁTOWANIE PROZODII

W syntezie konkatencyjnej kształtowanie prozodii wypowiedzi możliwe jest dzięki zastosowaniu odpowiednich algorytmów przetwarzania sygnału:

- zmiany częstotliwości podstawowej f_0 (*pitch shifting*)
- zmiany czasu trwania (*time stretching*)
- przetwarzanie dynamiki

Akcent

- Zmiana f_0 - podwyższenie lub obniżenie tonu
- Zmiana amplitudy - zwiększona intensywność
- Zmiana czasu trwania - wydłużenie samogłoski

Intonacja

- Zmiana f_0 - np. obniżenie tonu na końcu zdań oznajmujących i podniesienie na końcu zdań pytających

Iloczas

- Zmiana czasu trwania – przyspieszenie lub zwolnienie tempa wypowiedzi, wydłużenie akcentowanych samogłosek

KSZTAŁTOWANIE PROZODII

Algorytmy – zmiana czasu trwania i wysokości tonu jednostek mowy:

- resampling
- TD-PSOLA – Time-Domain *Pitch-Synchronous OverLap and Add*
- FD-PSOLA – *Frequency-Domain Pitch-Synchronous OverLap and Add*
- wokoder fazowy – *phase vocoder*

KSZTAŁTOWANIE PROZODII

TD-PSOLA

Powtarzanie ramek sygnału

→ wydłużenie.

Omijanie ramek sygnału

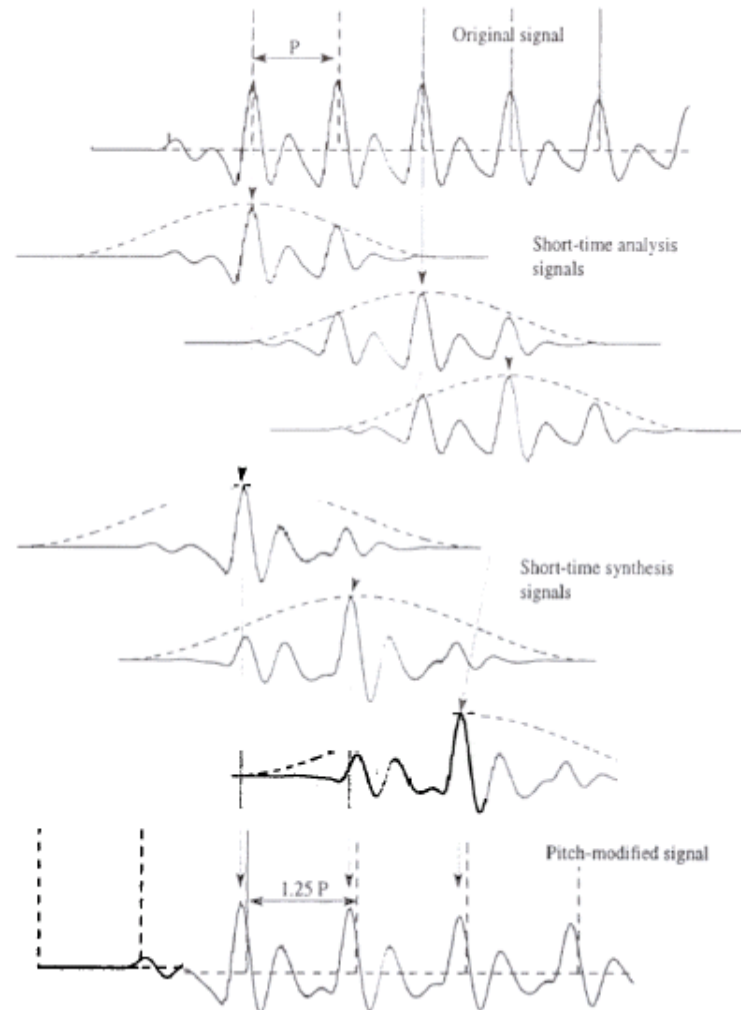
→ skrócenie.

Ramki rozmieszczone rzadziej

→ obniżenie f_0

Ramki rozmieszczone częściej →

podwyższenie f_0



KSZTAŁTOWANIE PROZODII

FD-PSOLA

- Synchroniczne pobieranie ramek sygnału
- Obliczanie transformaty Fouriera kolejnych ramek.
- Wyznaczanie obwiedni widmowej w celu rozłożenia widma sygnału na charakterystykę traktu głosowego i widmo okresowego pobudzenia.
- Modyfikacja widmowa pobudzenia w celu modyfikacji częstotliwości podstawowej.
- Wymnożenie zmodyfikowanego widma pobudzenia przez wcześniej wyznaczoną obwiednię widmową.
- Obliczenie odwrotnej transformaty Fouriera i resynteza sygnału, z ewentualnym powtarzaniem lub eliminacją ramek, jeśli zachodzi potrzeba modyfikacji czasowej.

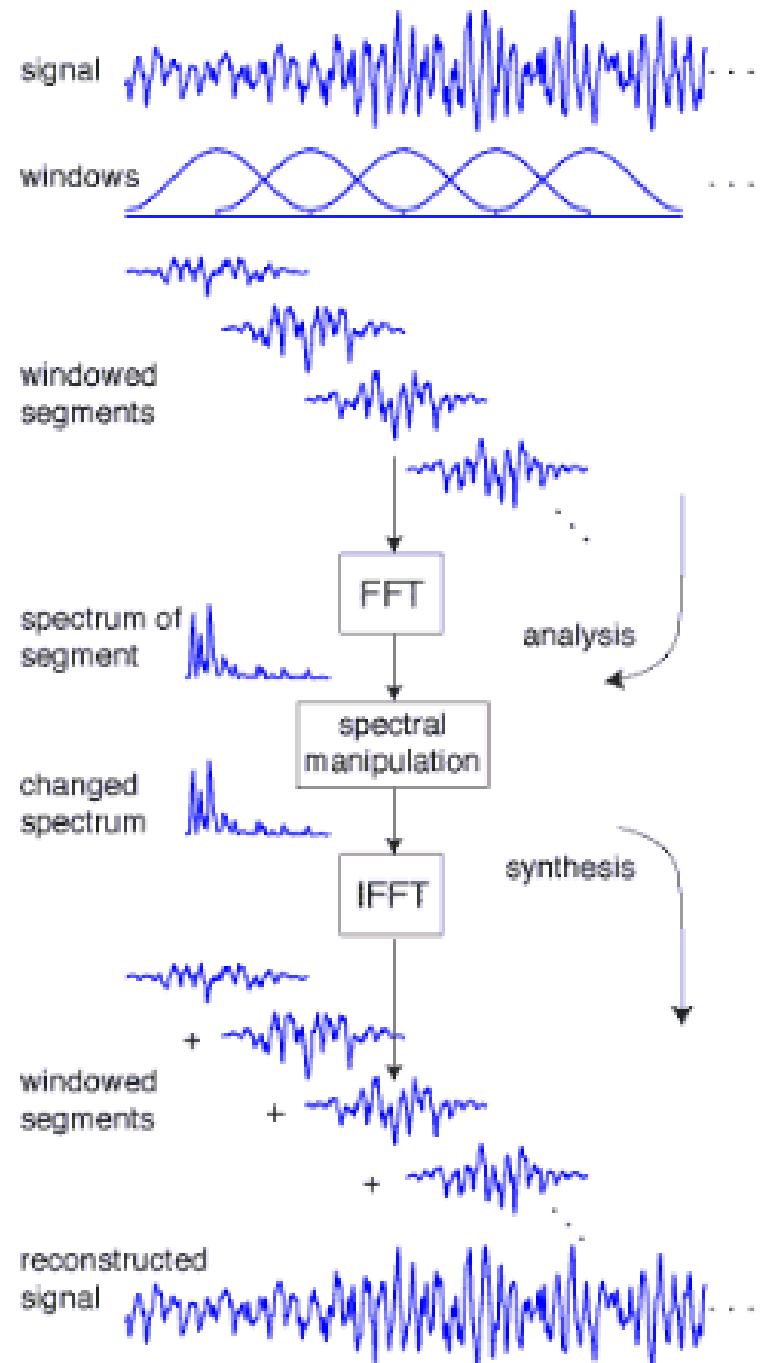
KSZTAŁTOWANIE PROZODII

Wokoder fazowy
Modyfikacja częstotliwościowa:

- Modyfikacja czasowa
- Przepróbkowanie
- Korekcja obwiedni widmowej

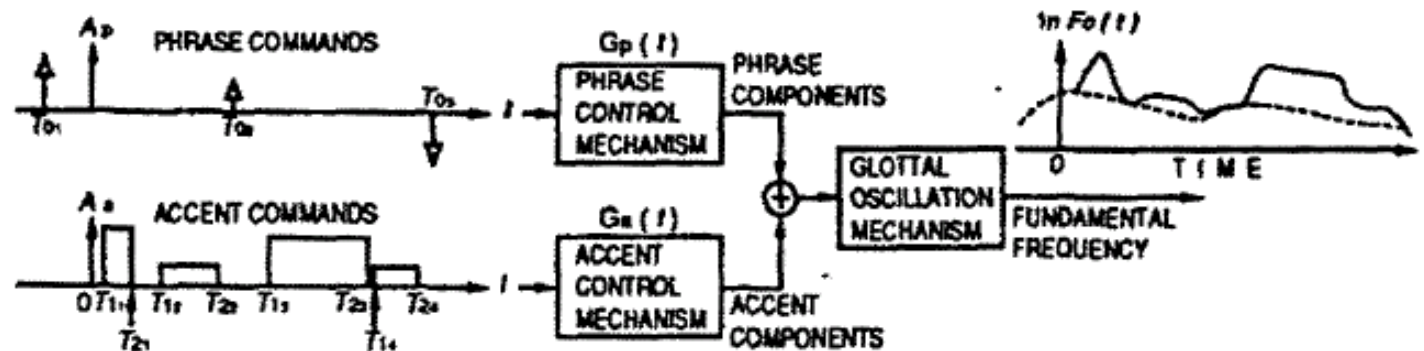
Problemy:

- Rozmywanie transjentów
- Efekt *phasera*

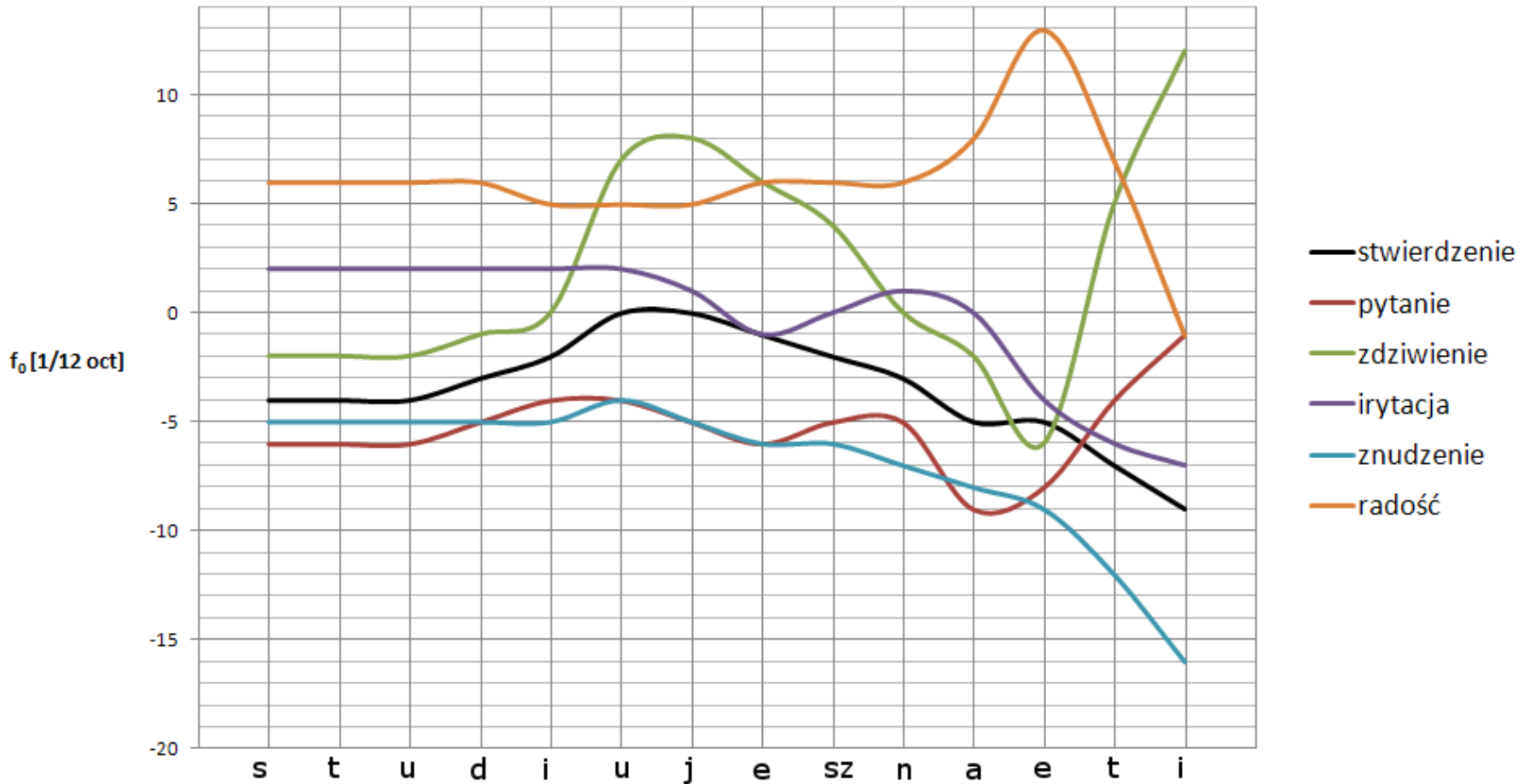


KONTUR INTONACYJNY

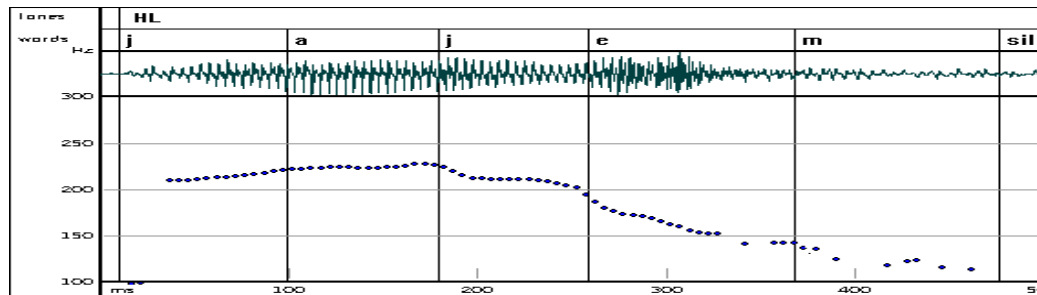
Model Fujisaki – tworzenie konturu intonacyjnego wypowiedzi jako złożenia składowych wynikających z akcentu i intonacji.



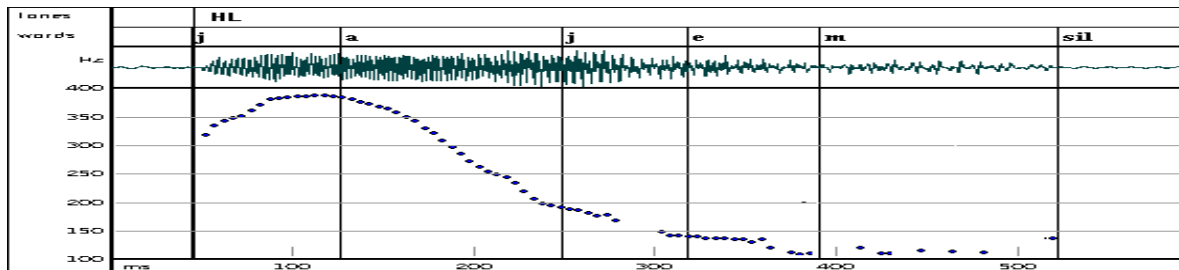
KONTUR INTONACYJNY



KSZTAŁTOWANIE PROZODII

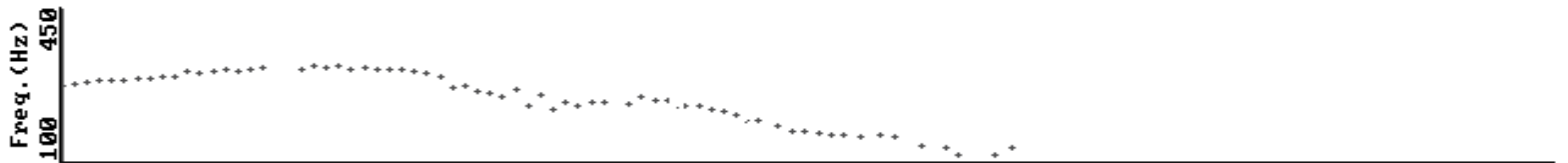
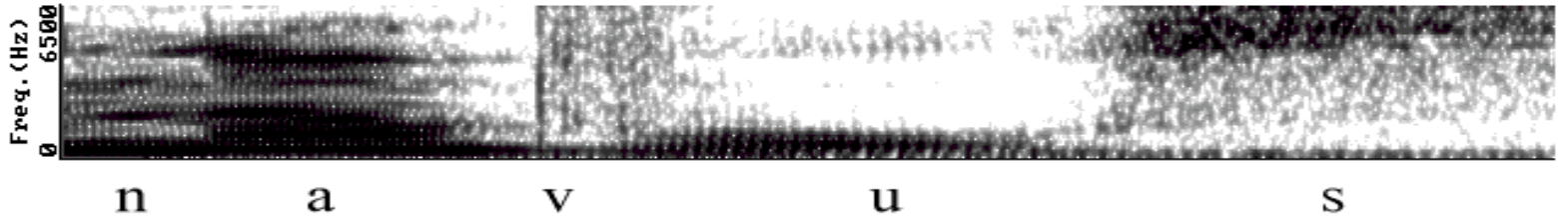
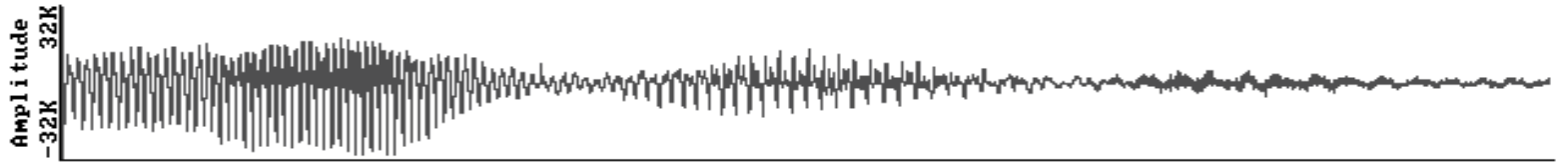


Przebieg częstotliwości podstawowej w wypowiedzi *ja jem.*

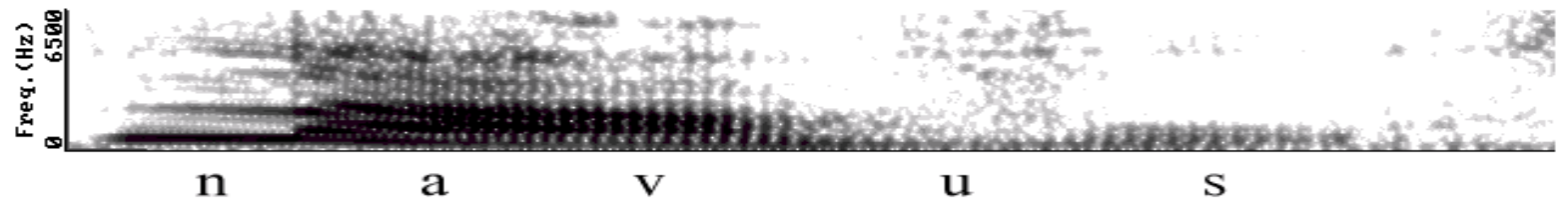


Przebieg częstotliwości podstawowej w wypowiedzi *jajem.*

źródło – Demenko G. – System syntezy mowy polskiej...



nawóz



na wóz

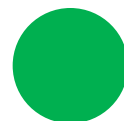
KSZTAŁTOWANIE ENERGII ARTYKULACYJNEJ

energia artykulacyjna \neq głośność
wzmocniony szept \neq krzyk

Wyznaczniki zwiększonej energii artykulacyjnej:

- zwiększenie amplitudy
- wydłużenie samogłosek
- podwyższenie częstotliwości podstawowej (5Hz/dB)
- przesunięcie pierwszego formantu w górę (3,5Hz/dB)

Aby uwzględnić wszystkie zmiany w widmie można obliczyć funkcję przejścia między poziomami głośności (np. głośnym i cichym) i wykorzystać ją do zmiany energii artykulacyjnej.



KSZTAŁTOWANIE PROZODII

Kształtowanie akcentu Akcent w języku polskim jest z reguły paroksytoniczny. Ma charakter mieszany toniczno-dynamiczny. W związku z tym akcentowanie sylaby wiąże się z:

- podniesieniem (lub obniżeniem) tonu
- wydłużeniem głoski
- wzmocnioną energią artykulacyjną

SYNTEZA KORPUSOWA

Synteza korpusowa – (ang. *unit selection*) wariant syntezy konkatenacyjnej. W bazie przechowywane są segmenty o różnej długości (np. temat i końcówka słowa). Do konkatenacji wypowiedzi wybierane są możliwie najdłuższe segmenty. Dzięki temu możliwe jest uzyskanie bardzo wysokiej jakości dla często występujących w języku słów.

CECHY DOBREGO SYNTETYZERA

- stuprocentowa zrozumiałość
- płynna mowa bez „zająknięć” i słyszalnych niedopasowań,
- poprawna normalizacja tekstu – zamiana skrótów, cyfr itp. na odpowiednie słowa,
- poprawność fonetyczna, także z uwzględnieniem wyjątków,
- zróżnicowanie wypowiedzi pod względem prozodycznym, poprawny akcent, intonacja,
- miły dla ucha głos lektora.

ZASTOSOWANIA SYNTEZY MOWY

- o urządzenia dla osób niewidomych: mówiące telefony, palmtopy itp.,
- o mówiące awatary na stronach internetowych, czasem prowadzące dialog z użytkownikiem,
- o urządzenia i programy edukacyjne,
- o udźwiękowienie stron WWW, aplikacji, filmów z napisami itp.



Syntetyzery anglojęzyczne:

NeoSpeech, TextAloud, eSpeak, Linguatec, Real Speak, Loquendo

Syntetyzery polskie:

IVONA, długo, długo nic... DANT, Spiker, SYNTALK





DZIĘKUJĘ ZA UWAGĘ!