Technologie multimedialne

Synteza mowy

opracowanie: mgr inż. Kuba Łopatka

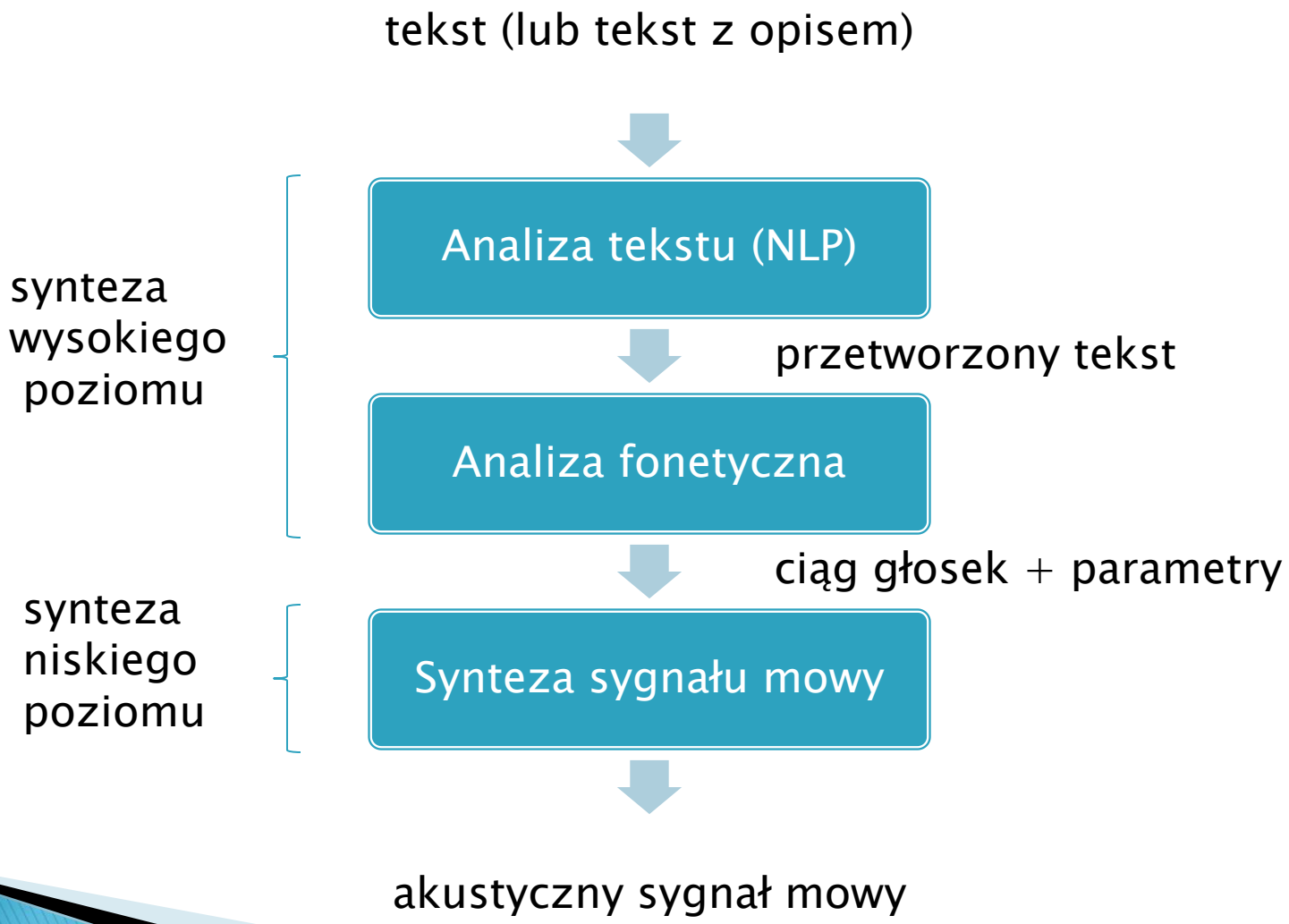
Synteza mowy

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

Podstawowe cechy (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 prozodyczna.

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 zło tych pięć dzie siąt gro szy

analiza fonetyczna:

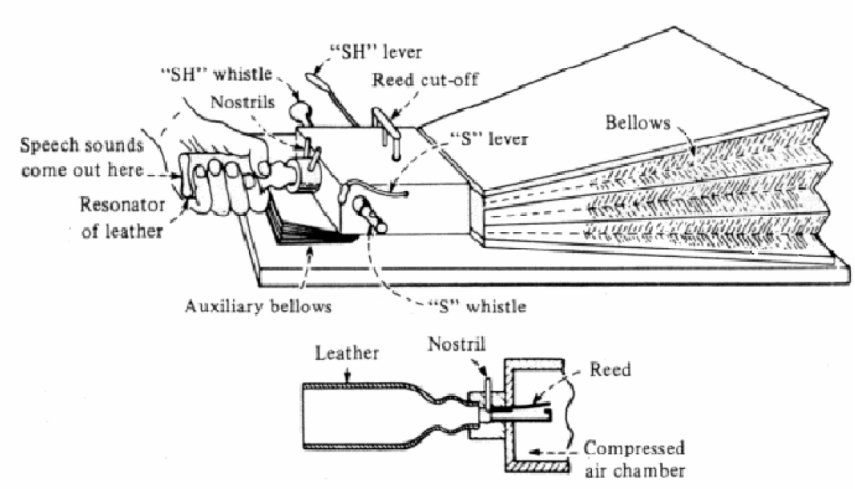
zośadałastefanowipjęćźzłotyhpjeńćźęśǫdgrošy

Synteza sygnału mowy

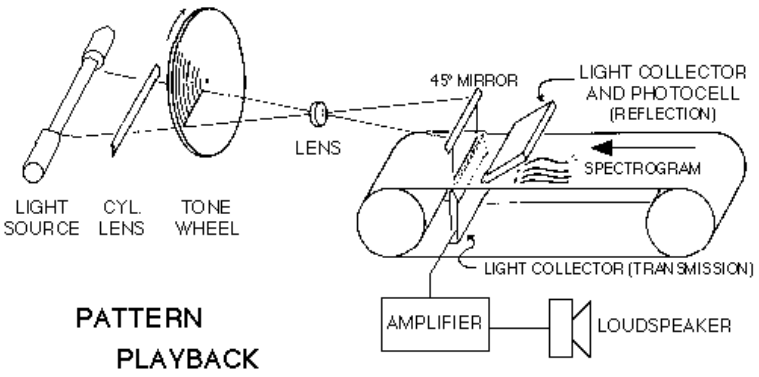
Wyróżnia się 3 podstawowe metody:

Historia

Pierwsze syntetyzery – mechaniczne (von Kempelen 1791)



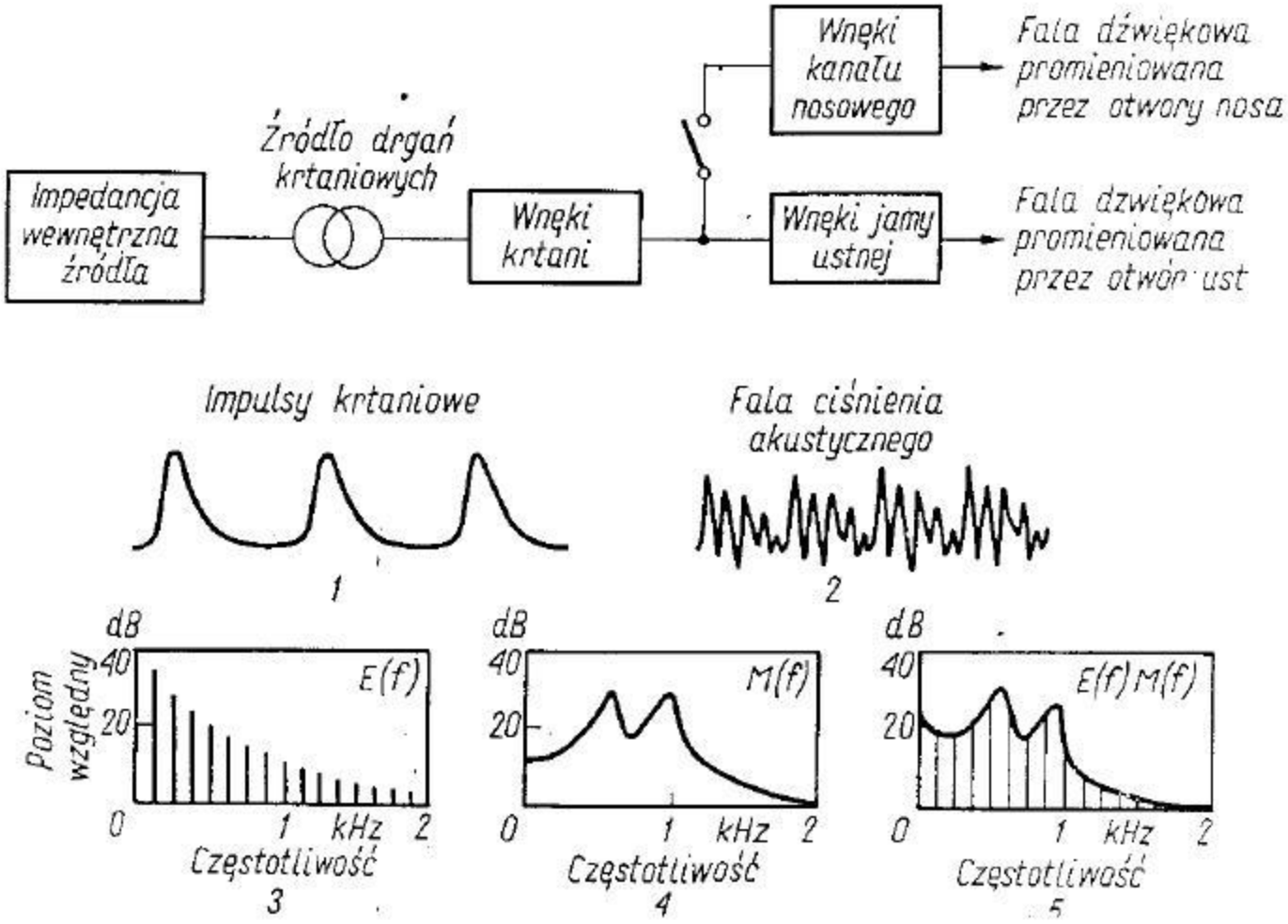
Pattern Playback – 1950 – maszyna „czytająca” spektrogram



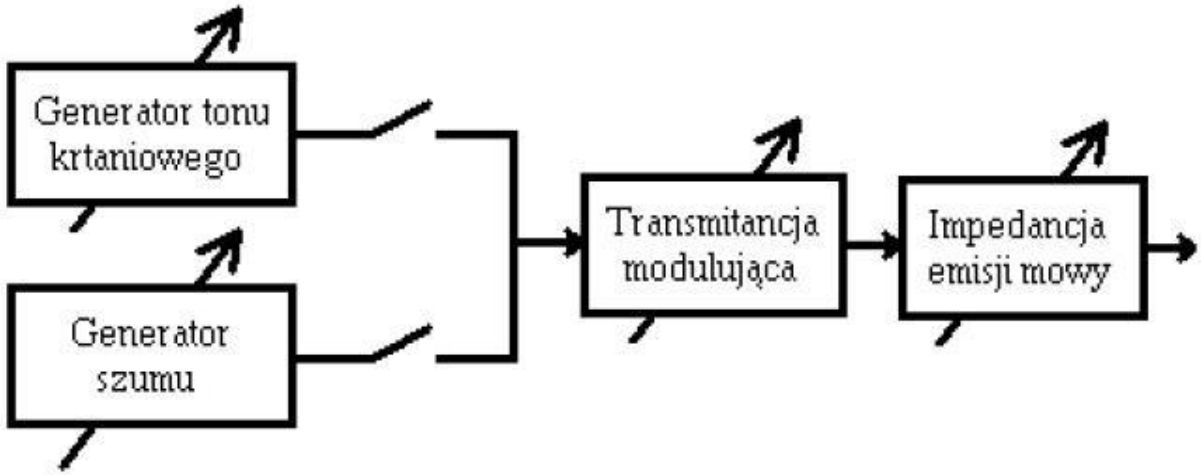
Pierwszy syntetyzer formantowy – 1964 r.

Później – synteza artykulacyjna i konkatenacyjna

Model generowania sygnału mowy



Schemat zastępczy wytwarzania mowy



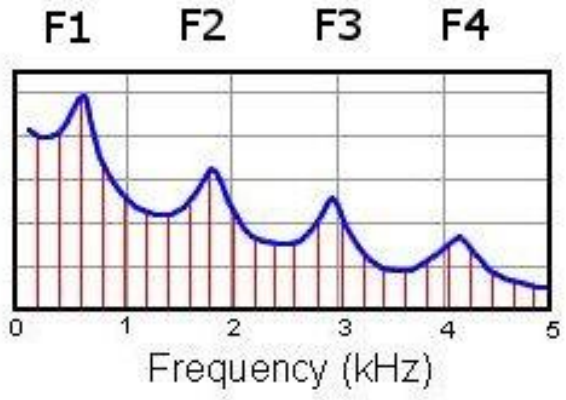
Istotą dwóch pierwszych podejść do syntezy mowy jest *zamodelowanie* pobudzenia (tonu krtaniowego) i funkcji przenoszenia (transmitancji) traktu głosowego, która je moduluje.

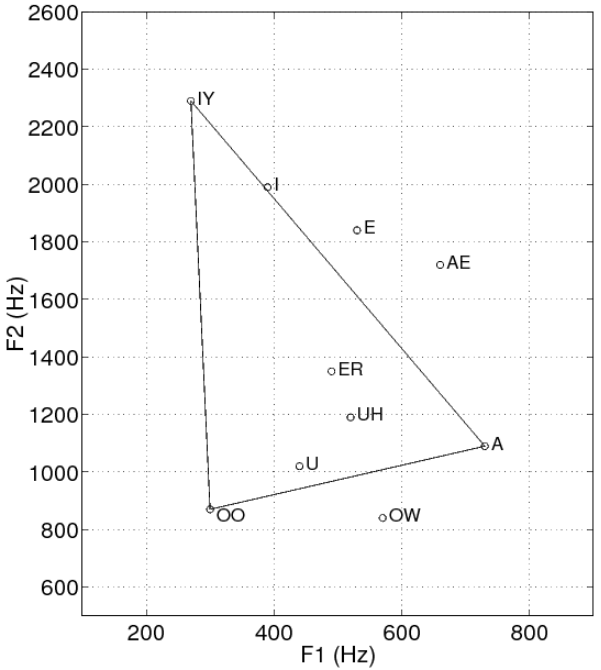
Metody syntezy

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.

Formanty

Formant – skupisko energii w widmie sygnału mowy. Rozmieszczenie i relacje między formantami (zwłaszcza pierwszymi oznaczanymi F1 i F2) mają kluczowe znaczenie dla zrozumiałości mowy.





Metody syntezy

Synteza artykulacyjna – rozwinięcie metody formantowej – próbkowanie (lub teoretyczne obliczanie) charakterystyki traktu głosowego i odwzorowanie jej za pomocą modelu matematycznego – najczęściej kodowania predykcyjnego (LPC – *Linear Predictive Coding*).

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 konkatenacyjnym:

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

dłuższe segmenty = lepsza jakość = obszerniejsza baza

Difony

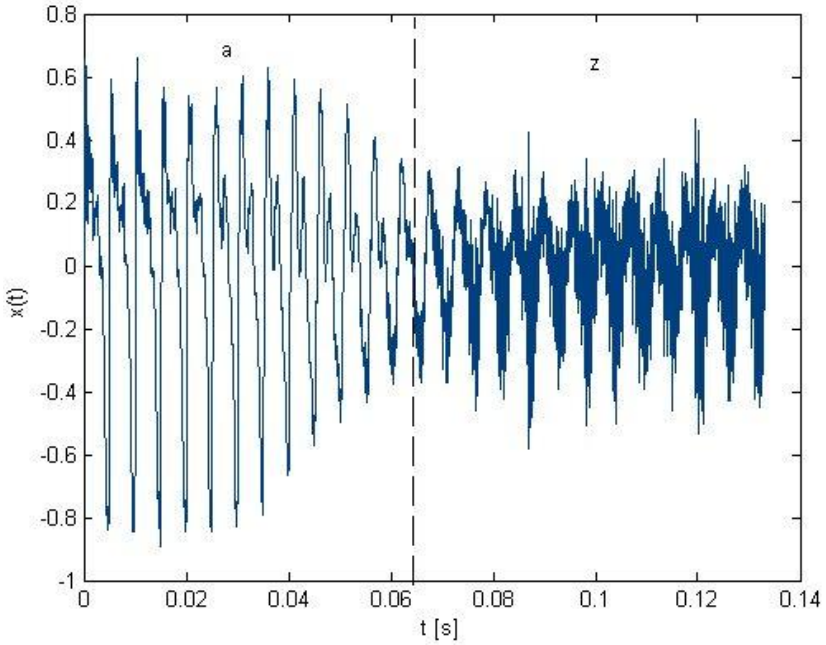
Brzmienie głoski jest bardzo mocno zależne od głosek poprzednich i następnych. Difony zawierają przejście między dwoma głoskami wraz ze stanami ustalonymi obu głosek.

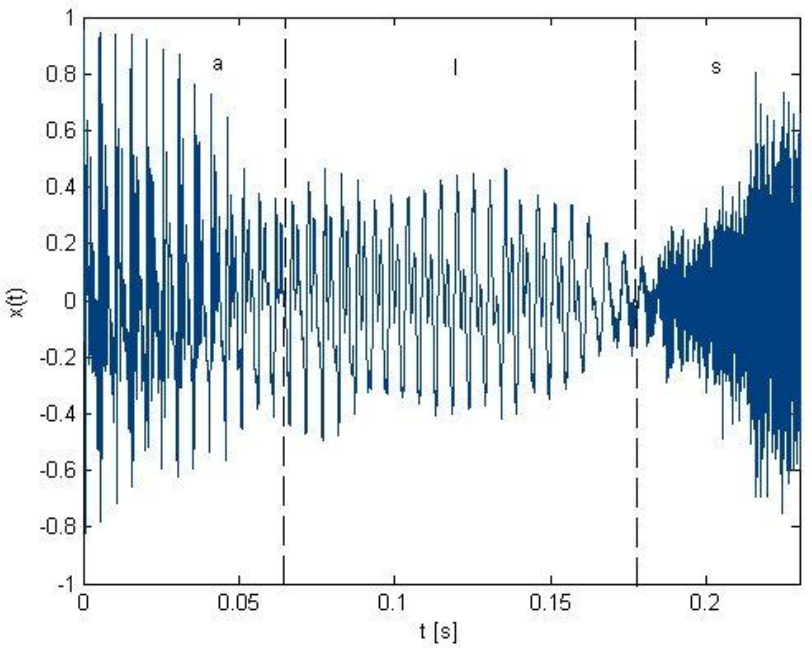
Składanie wypowiedzi z difonów:

`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-#

Przykładowe segmenty





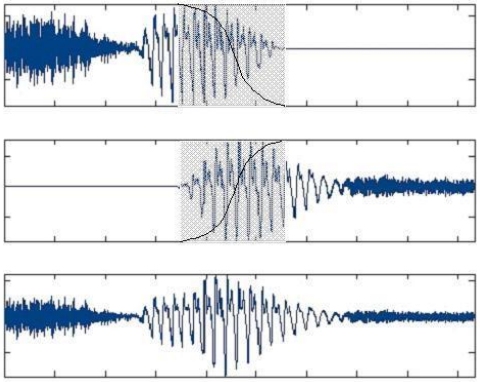
difon – połączenie dwóch głosek
liczba difonów w j. polskim – $37^2=1369$

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

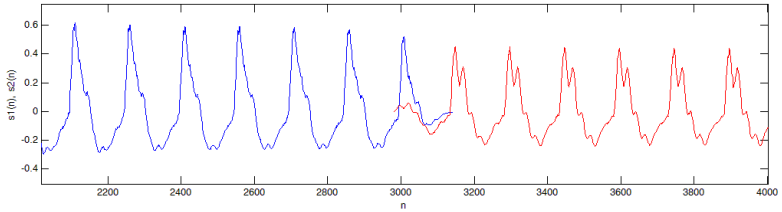
Konkatenacja

Metody konkatenacji difonów:

- przemiksowanie (*cross-fade*),



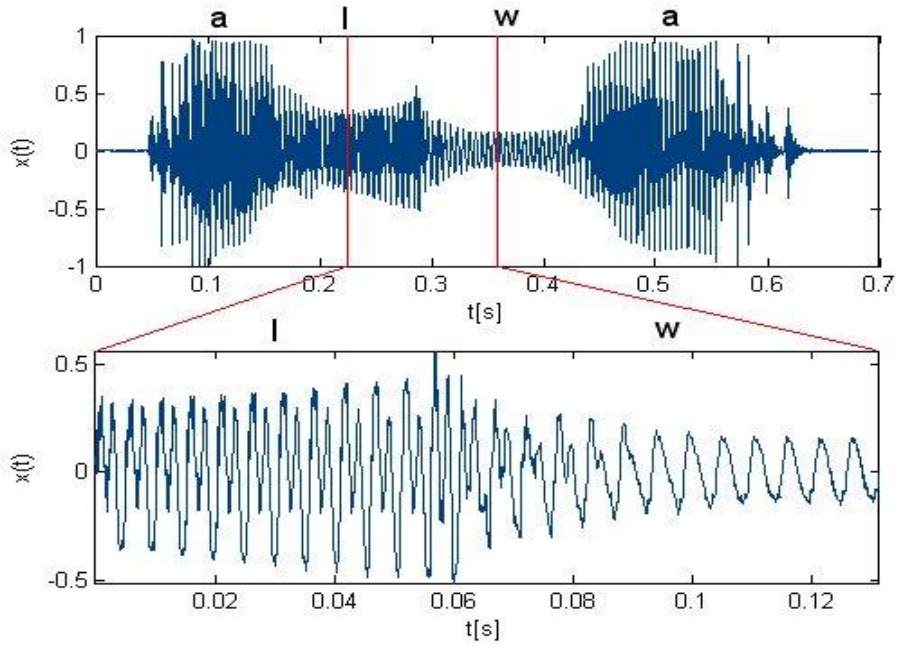
- PSOLA (*Pitch-Synchronous OverLap and Add*) – połączenie zgodnie z okresem podstawowym – zapewnia ciągłość tonu,



- MBROLA (*Multi-Band Resynthesis OverLap and Add*) – stosowanie dodatkowego przetwarzania segmentów w celu uzyskania lepszego brzmienia.

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.



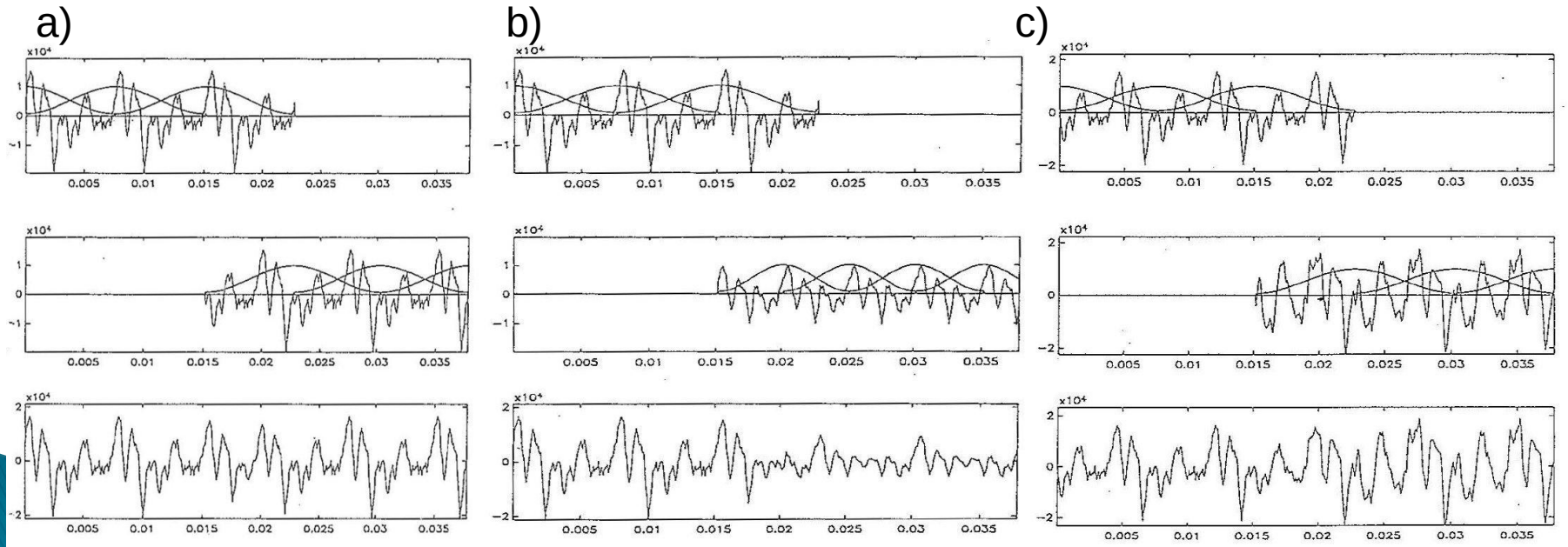
Difony – niedopasowanie

Po połączeniu difonów możliwe jest niedopasowanie:

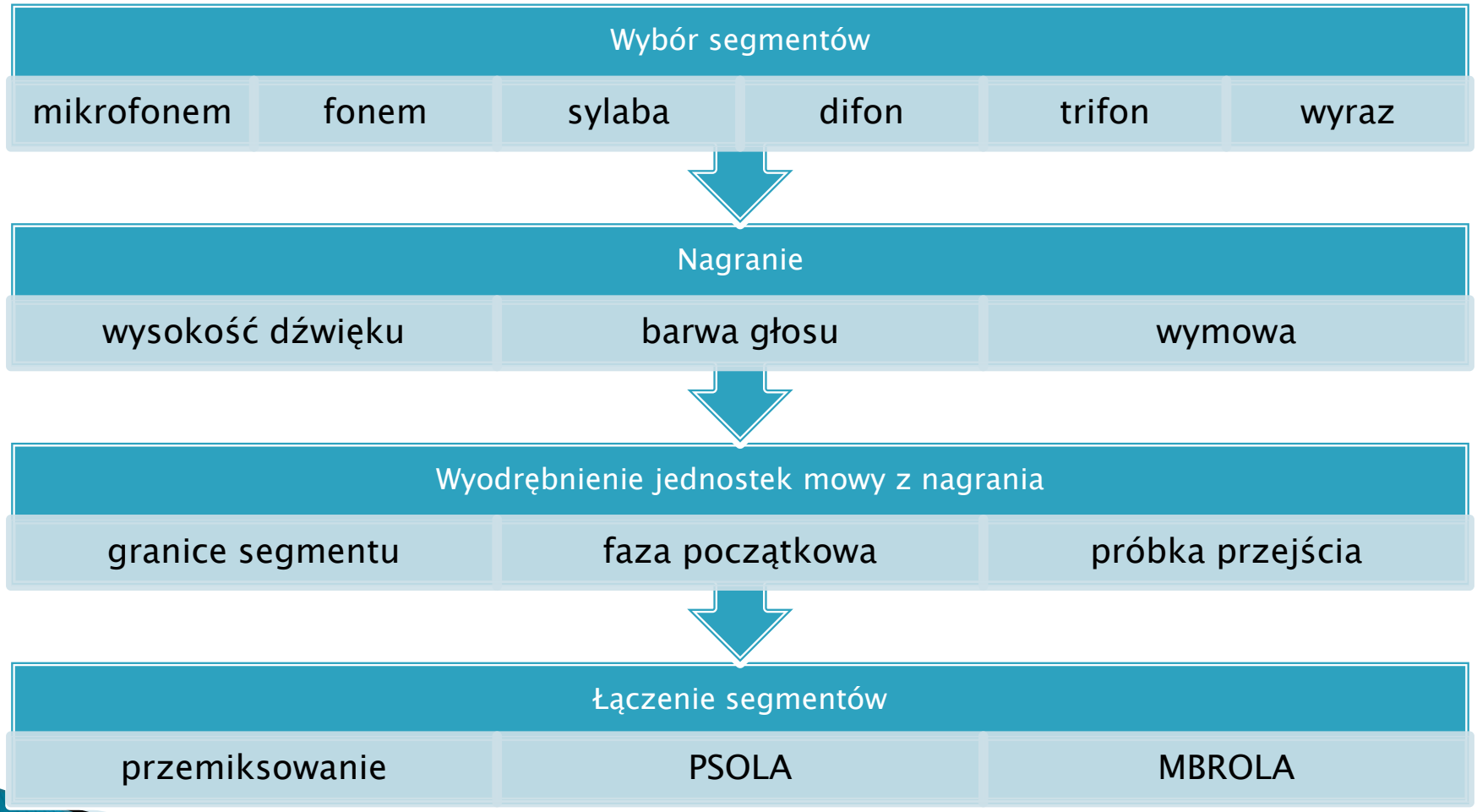
a) fazy (różne fazy)

b) tonu podstawowego (różna wysokość)

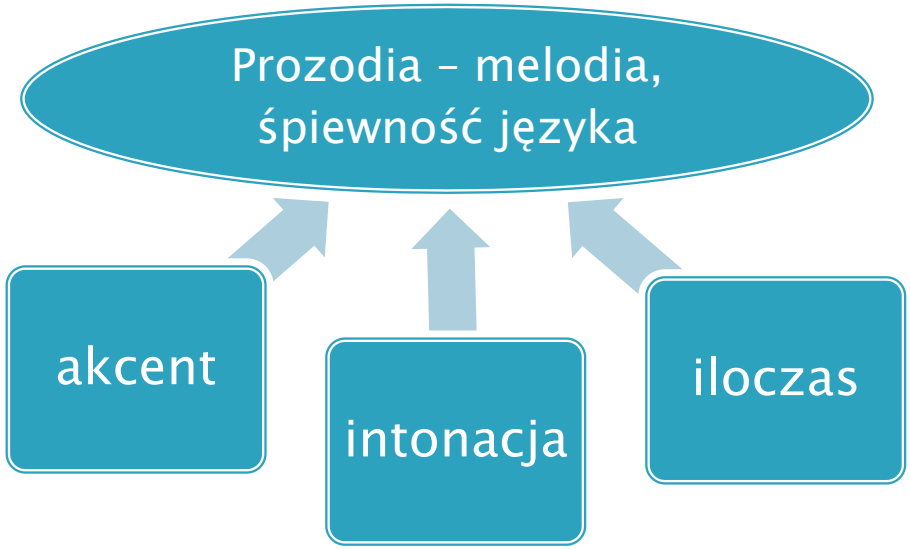
c) obwiedni widmowej (różne brzmienia głosek)



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

Możliwe jest kształtowanie prozodii wypowiedzi przez zastosowanie odpowiednich algorytmów przetwarzania sygnału:

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

Akcent	Intonacja	Iloczas
• Zmiana f_0 – podwyższenie lub obniżenie tonu • Zmiana amplitudy – zwiększona intensywność • Zmiana czasu trwania – wydłużenie samogłoski	• Zmiana f_0 – np. obniżenie tonu na końcu zdań oznajmujących i podniesienie na końcu zdań pytających	• Zmiana czasu trwania – przyspieszenie lub zwolnienie tempa wypowiedzi, wydłużenie akcentowanych samogłosek

Synteza korpusowa

Synteza korpusowa – 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

- urządzenia dla osób niewidomych: mówiące telefony, palmtopy itp.,
- mówiące avatary na stronach internetowych, czasem prowadzące dialog z użytkownikiem,
- urządzenia i programy edukacyjne,
- 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