Przetwarzanie mowy w czasie rzeczywistym

Akustyka mowy

opracowanie: M. Kaniewska, A. Kupryjanow, K. Łopatka

prezentuje: J.Kotus

Plan wykładu

- Zasada przetwarzania sygnału w czasie rzeczywistym
- Algorytmy zmiany czasu trwania sygnału
- Modyfikacja częstotliwości podstawowej
- Algorytmy transformacji głosu

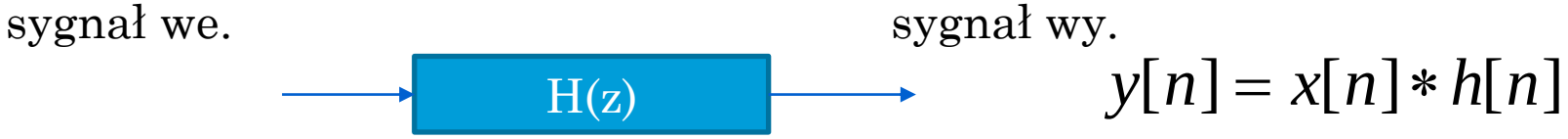
Zasada przetwarzania sygnału w czasie rzeczywistym

System czasu rzeczywistego

- Tryb **przetwarzania w czasie rzeczywistym** jest takim trybem, w którym programy przetwarzające dane napływające z zewnątrz są zawsze gotowe, a **wynik ich działania jest dostępny nie później niż po zadanym czasie**.
- Moment nadejścia kolejnych danych może być losowy (asynchroniczny) lub ściśle określony (synchroniczny)

Przetwarzanie w dziedzinie czasu

- Wykorzystanie filtru cyfrowego o transmitancji H(z) i odpowiedzi impulsowej h(n)
- Minimalne opóźnienia – działanie „z próbki na próbkę”

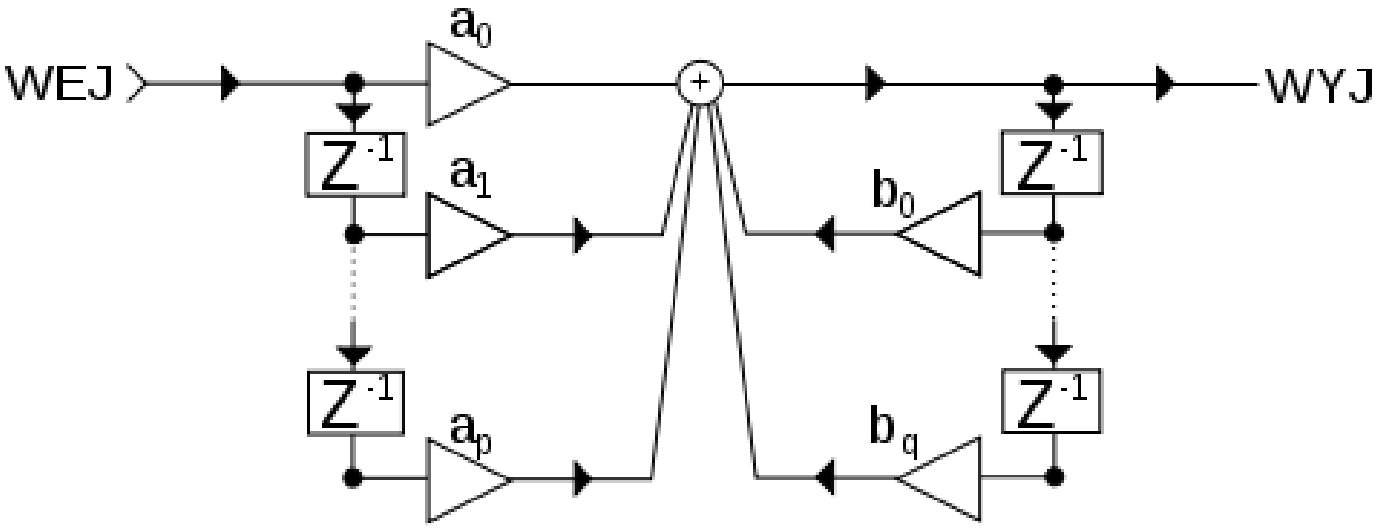


Problemy:

- trudność projektowania filtrów
- długie filtry – kosztowność obliczeniowa
- Konieczność zapewnienia stabilności

Przetwarzanie w dziedzinie czasu

$$y(k)=\sum_{n=1}^{M} b(n)x(k-n+1)-\sum_{n=2}^{N} a(n)y(k-n+1)$$



Struktura bezpośrednia filtru cyfrowego.
Moduły z^{-1} oznaczają opóźnienie sygnału o jedną próbkę, natomiast a_p oraz b_q są współczynnikami filtru.

https://pl.wikipedia.org/wiki/Filtr_o_niesko%C5%84czonej_odpowiedzi_impulsowej

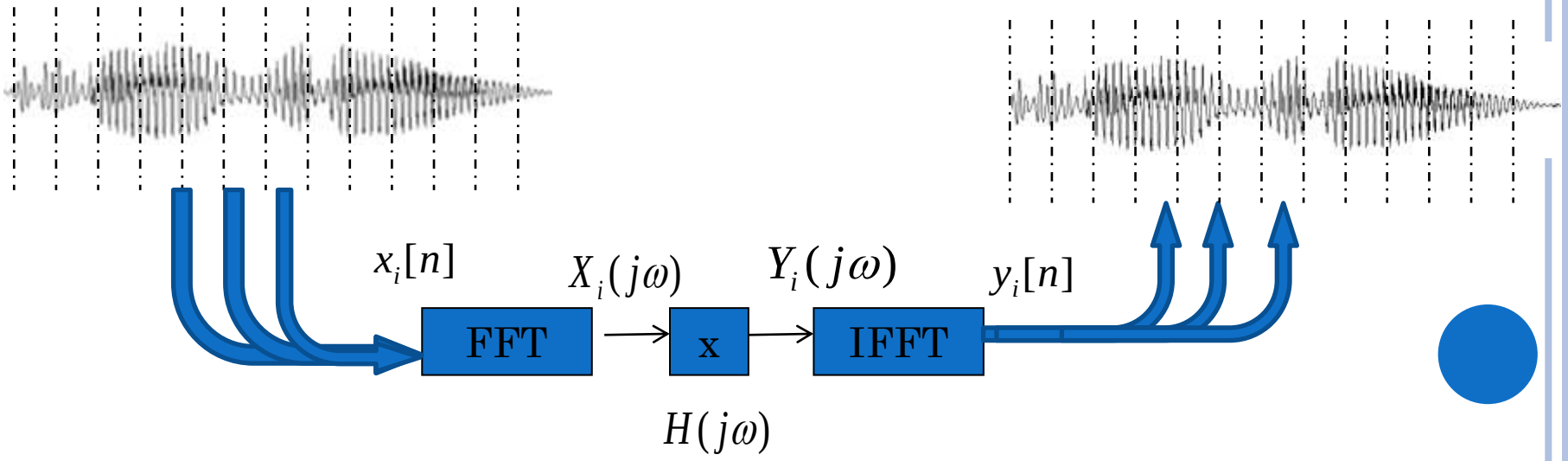
Cyfrowe przetwarzanie sygnałów, Adam Łutkowski,

Przetwarzanie w dziedzinie widma – Filtracja OLA

- Technika pozwalająca wykonać dowolną filtrację z wykorzystaniem odwracalnej transformacji (np. DFT)

$$h[n] \rightarrow H(j\omega) = |H| e^{j\angle H}$$

$$x[n] * h[n] \Leftrightarrow X(j\omega) \cdot H(j\omega)$$



Filtracja OLA

- Przy pobieraniu ramek obowiązuje zasada nakładania się (overlap) – zwykle 50 %
- Przy **resyntezie** sygnału dodajemy do siebie nakładające się ramki – **O**ver**L**ap **A**dd

Zalety:

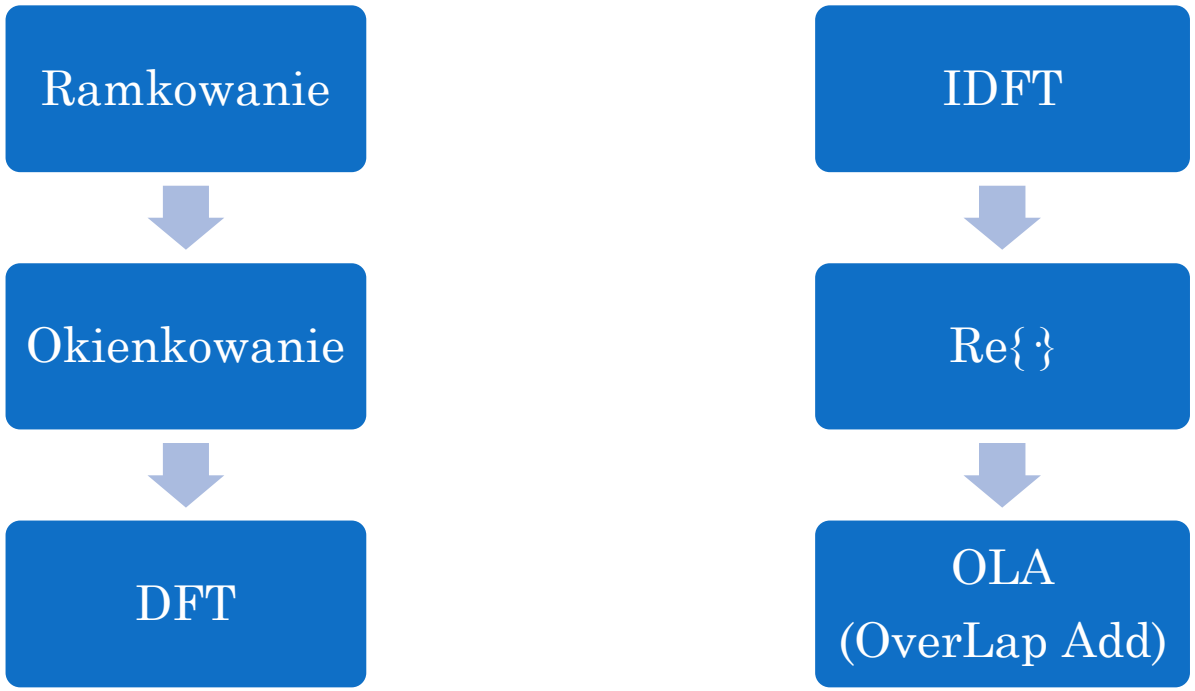
- mniejsza złożoność obliczeniowa w stosunku do filtracji splotowej przy znacznej długości filtra
- możliwość „intuicyjnego” projektowania filtrów

Wady:

- wprowadza stałe opóźnienie wynikające z bufora DFT

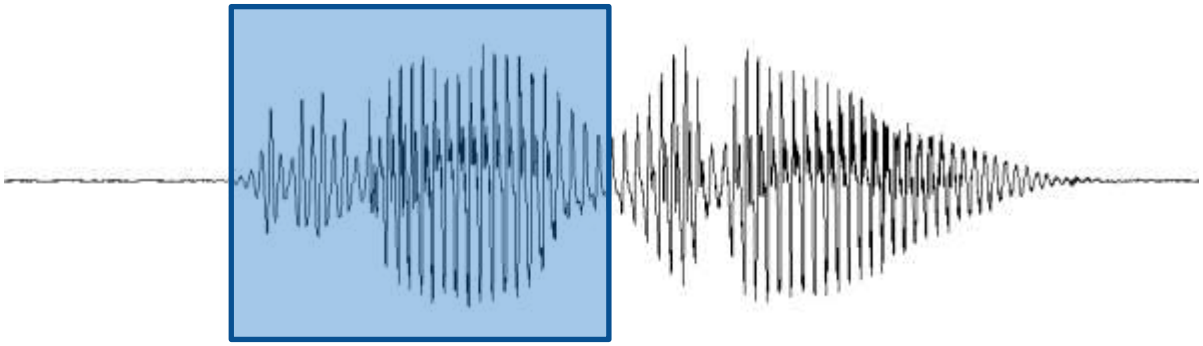
Short Time Fourier Transform

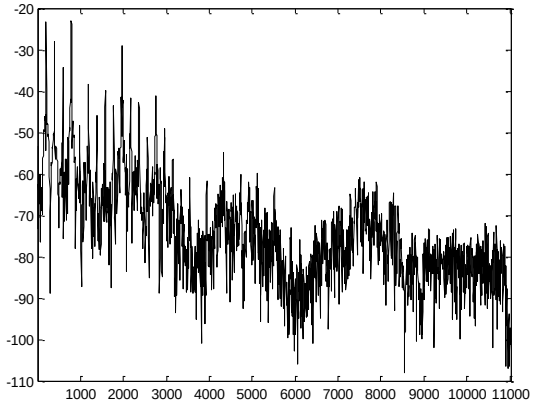
- STFT (nie mylić z FFT!) - Short-Time Fourier Transform – technika analizy sygnału - krótkoczasowe przekształcenie Fouriera – przekształcenie czasowo częstotliwościowe



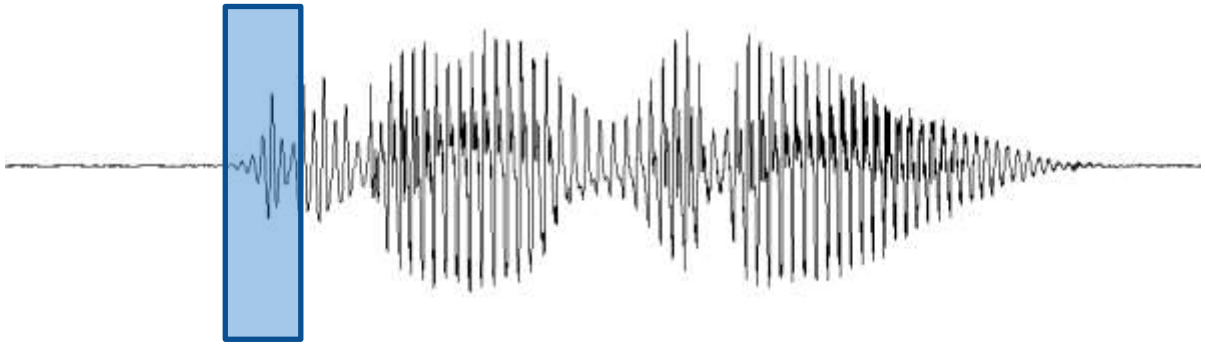
Zasada Nieoznaczoności

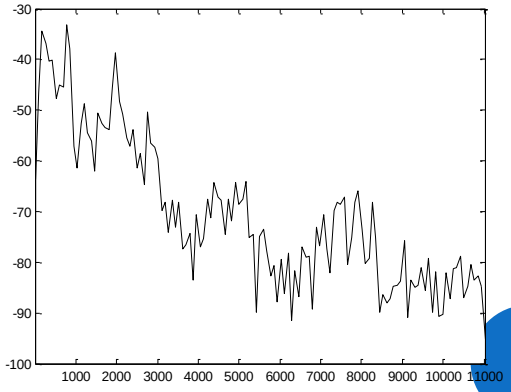
rozdzielczość czasowa <<>> rozdzielczość częstotliwościowa





długa ramka – dokładne widmo





krótka ramka – zgrubne widmo

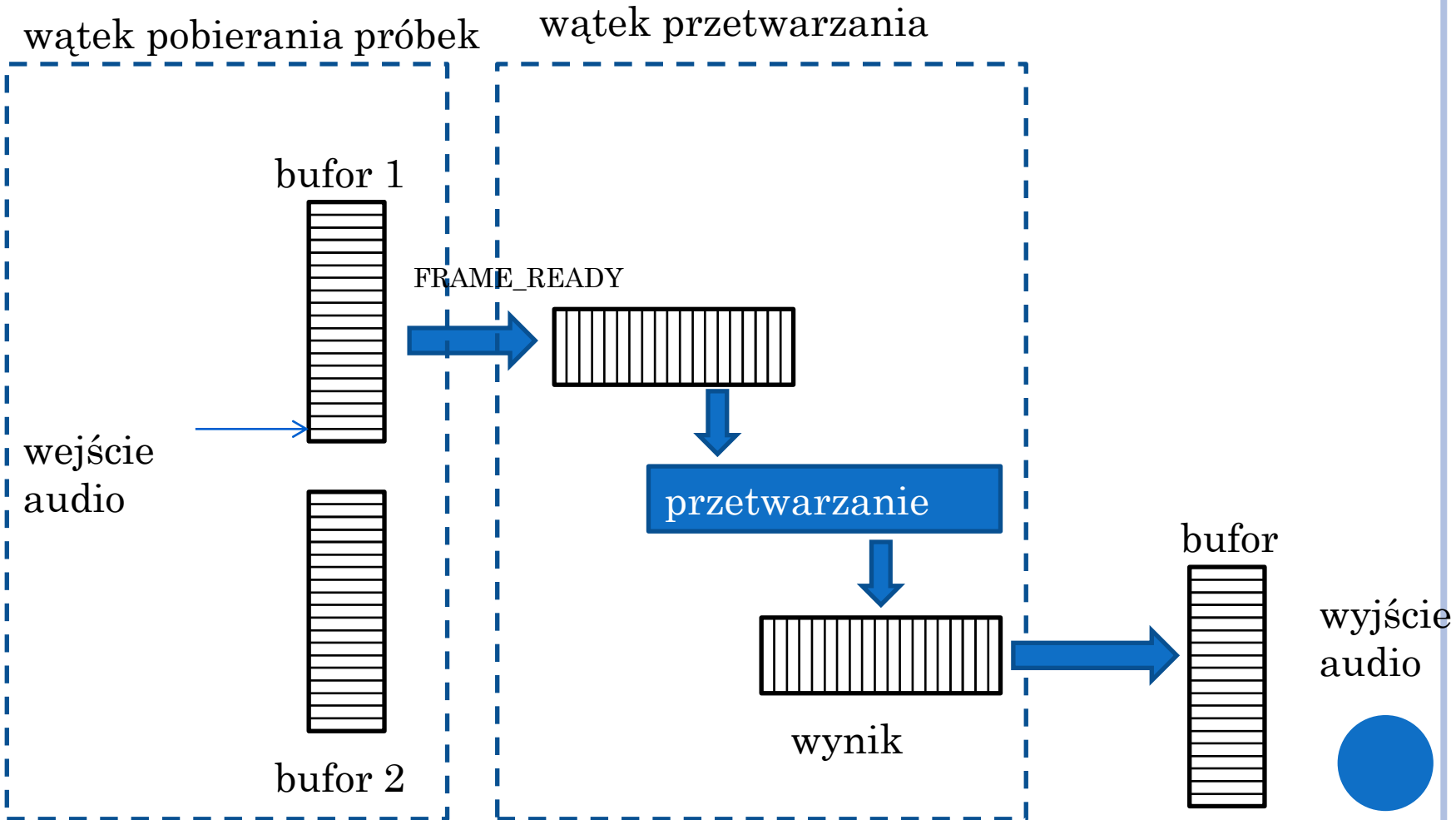
Preemfaza / Deemfaza

- Preemfaza – filtracja górnoprzepustowa 6dB/okt.
  - uwypuklenie wysokich częstotliwości

$$y[n] = x[n] - a \cdot x[n-1]$$

- Deemfaza – filtracja dolnoprzepustowa 6dB/okt.
  - odwrotność preemfazy

Przykładowa implementacja



Algorytmy zmiany czasu trwania i wysokości dźwięku

Zmiana czasu / wysokości dźwięku

operacja	zmienia	zachowuje
Zmiana szybkości odtwarzania	• czas trwania • częstotliwość • brzmienie	nic
Zmiana szybkości odtwarzania + przepróbkowanie	• częstotliwość • brzmienie	• czas trwania
Zmiana wysokości (pitch shifting)	• częstotliwość	• czas trwania • brzmienie
Zmiana czasu trwania (time stretching)	• czas trwania	• częstotliwość • brzmienie

ALGORYTMY MODYFIKACJI CZASU TRWANIA SYGNAŁU

- Założenia:
  - Brak zmiany wysokości dźwięku
  - Wprowadzanie jak najmniejszej liczby zniekształceń:
    - Nieciągłości fazy i częstotliwości
    - Trzasków
    - Powtarzania transjentów
  - Osiągnięcie największego możliwego podobieństwa do sygnału wejściowego
- Zastosowania:
  - Synteza mowy
  - Dopasowanie czasu trwania wypowiedzi np. audiobooki, audycje radiowe i telewizyjne
  - Testy percepcji mowy
  - Wspomaganie procesu rozumienia mowy przez osoby z pogorszoną rozdzielczością czasową słuchu
  - Modyfikacja brzmienia mowy
  - …

ALGORYTMY MODYFIKACJI CZASU TRWANIA SYGNAŁU

- Algorytmy działające po stronie czasu:
  - OLA (Overlap and Add)
  - SOLA (Synchronous Overlap and Add)
  - PSOLA (Pitch-synchronous Overlap and Add)
  - WSOLA (Waveform Similarity Overlap and Add)
  - PAOLA (Peak Alignment Overlap and Add)
- Algorytmy działające po stronie widma:
  - FD-PSOLA
  - Wokoder-fazowy

ALGORYTMY – WSPÓŁCZYNNIK SKALI

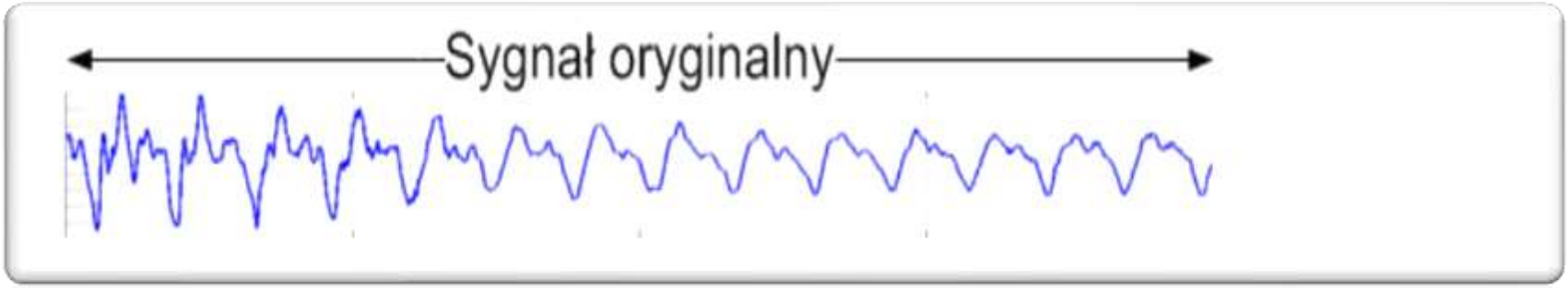
$$Ts = \alpha \cdot Ta$$

gdzie Ts – przesunięcie czasowe syntezy,

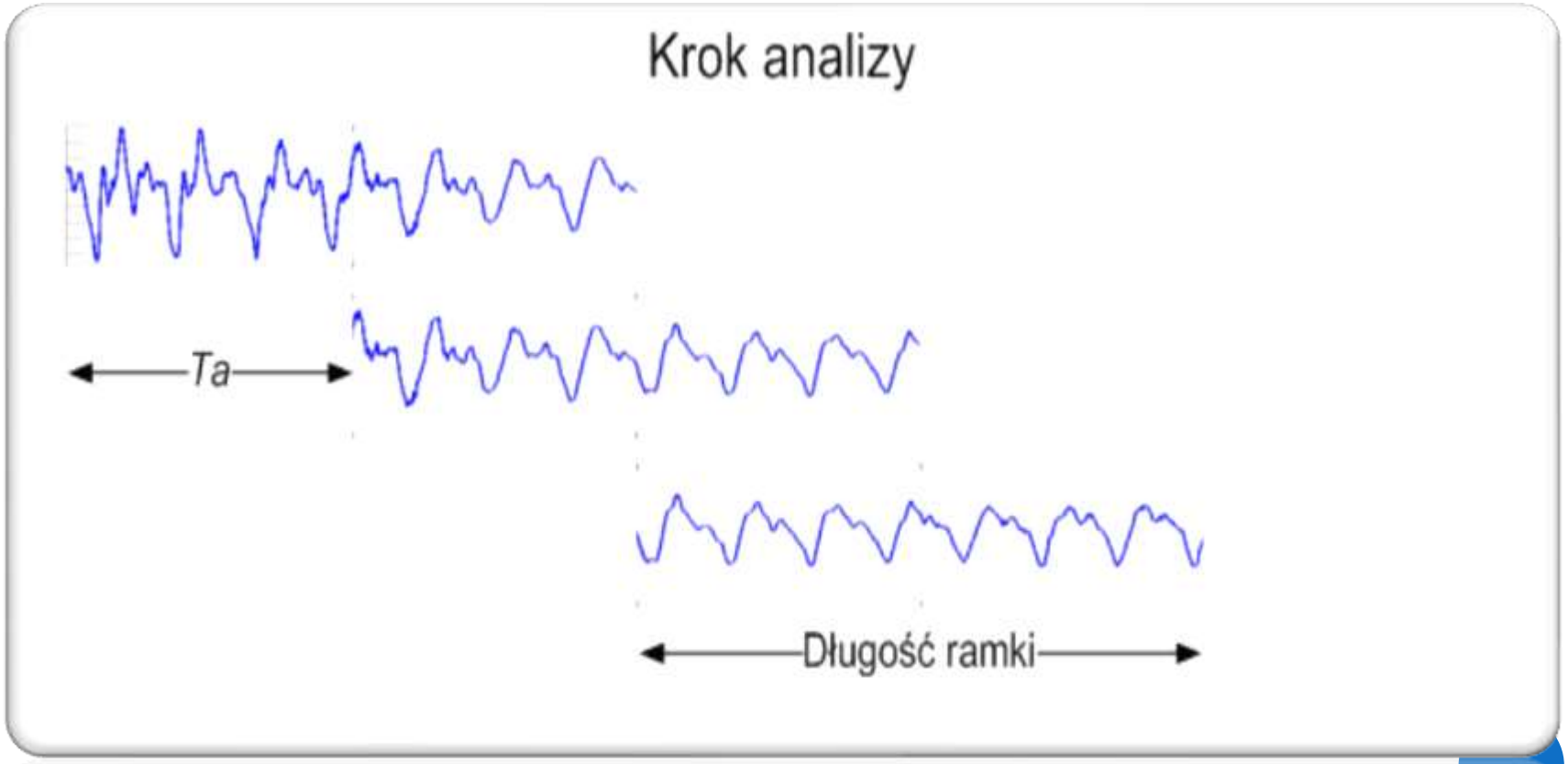
Ta – przesunięcie czasowe analizy,

α – współczynnik skali.

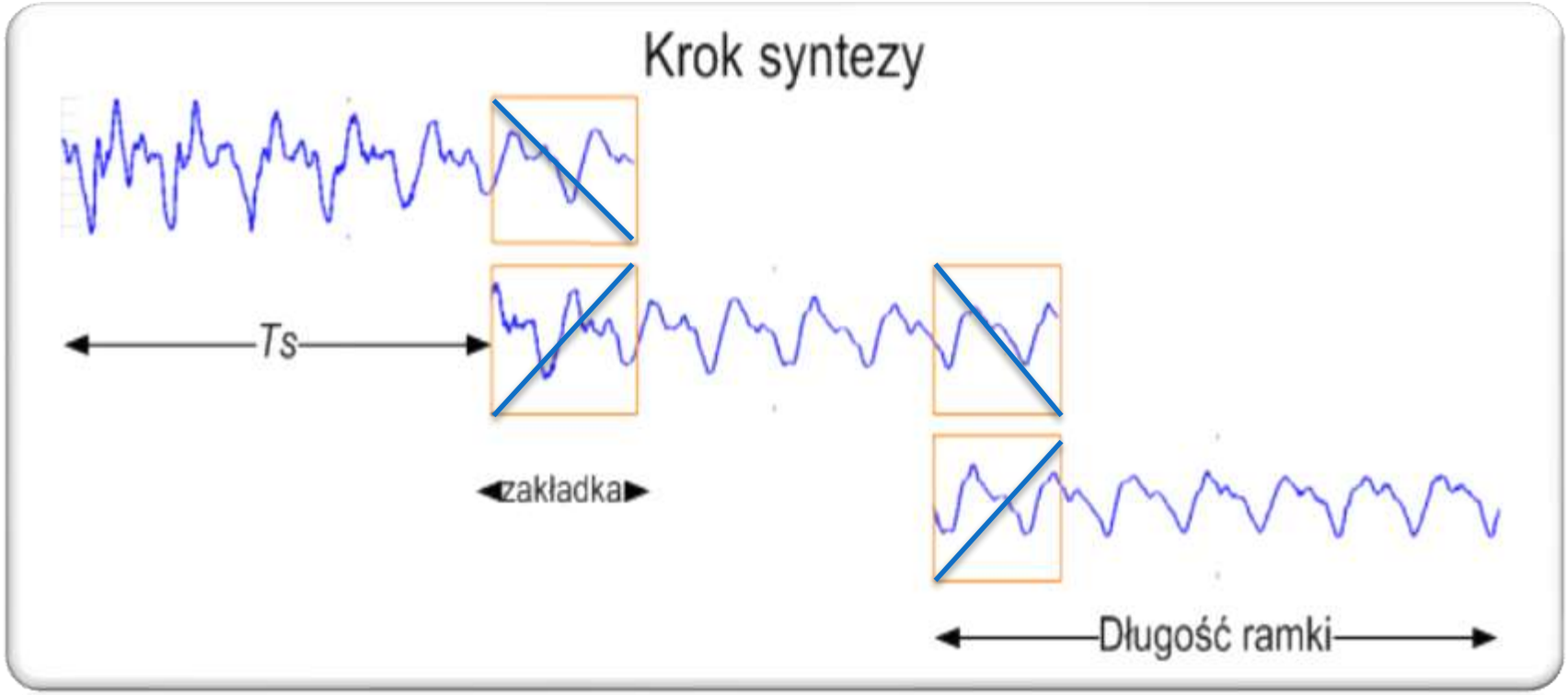
Algorytm OLA



Algorytm OLA - analiza

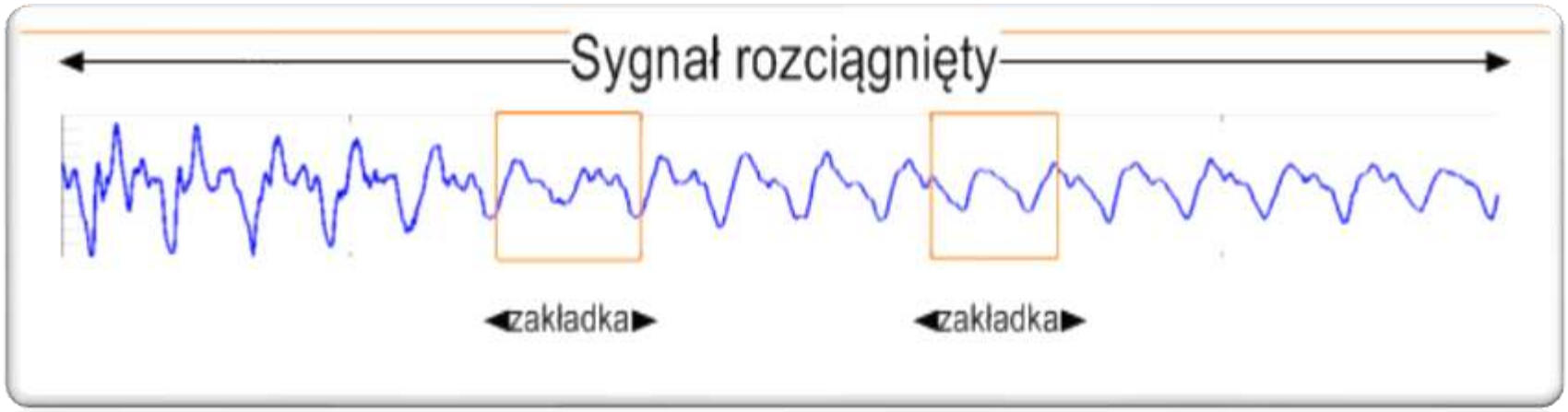


Algorytm OLA - Synteza



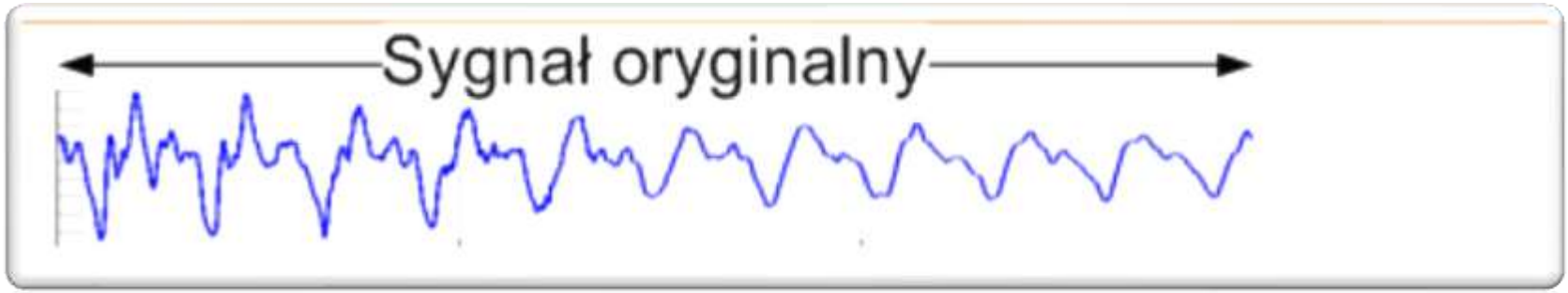
- Dla danego wsp. skali stały rozmiar zakładki
- Obszary zakładek są przemiksowywane z cross-fadem

Algorytm OLA

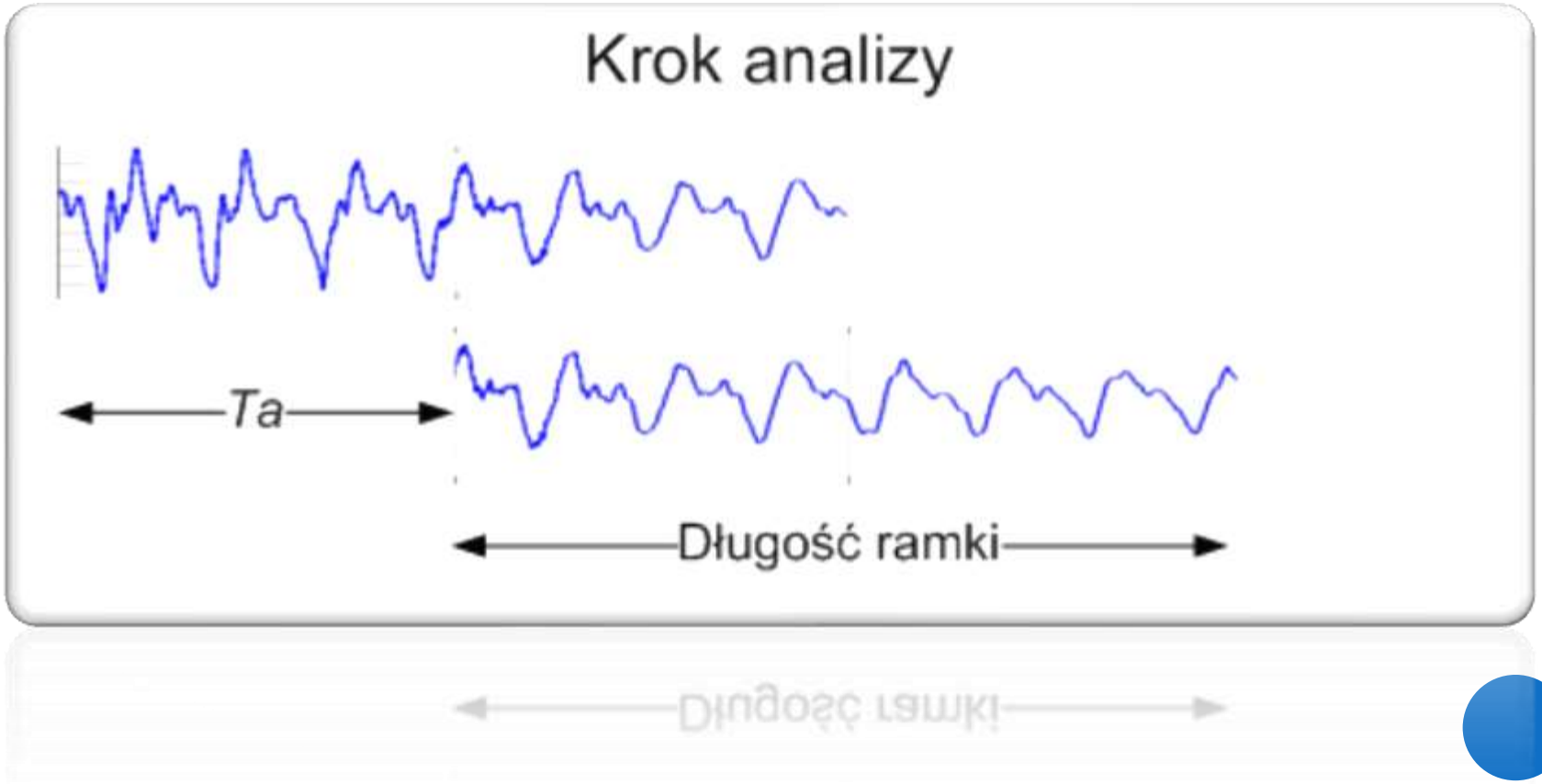


- Zalety:
  - Niewielka złożoność obliczeniowa
  - Szybki
- Wady:
  - Sygnał wynikowy jest niskiej jakości
  - Słyszalne są trzaski na łączeniach ramek
  - Występują nieciągłości fazy i częstotliwości

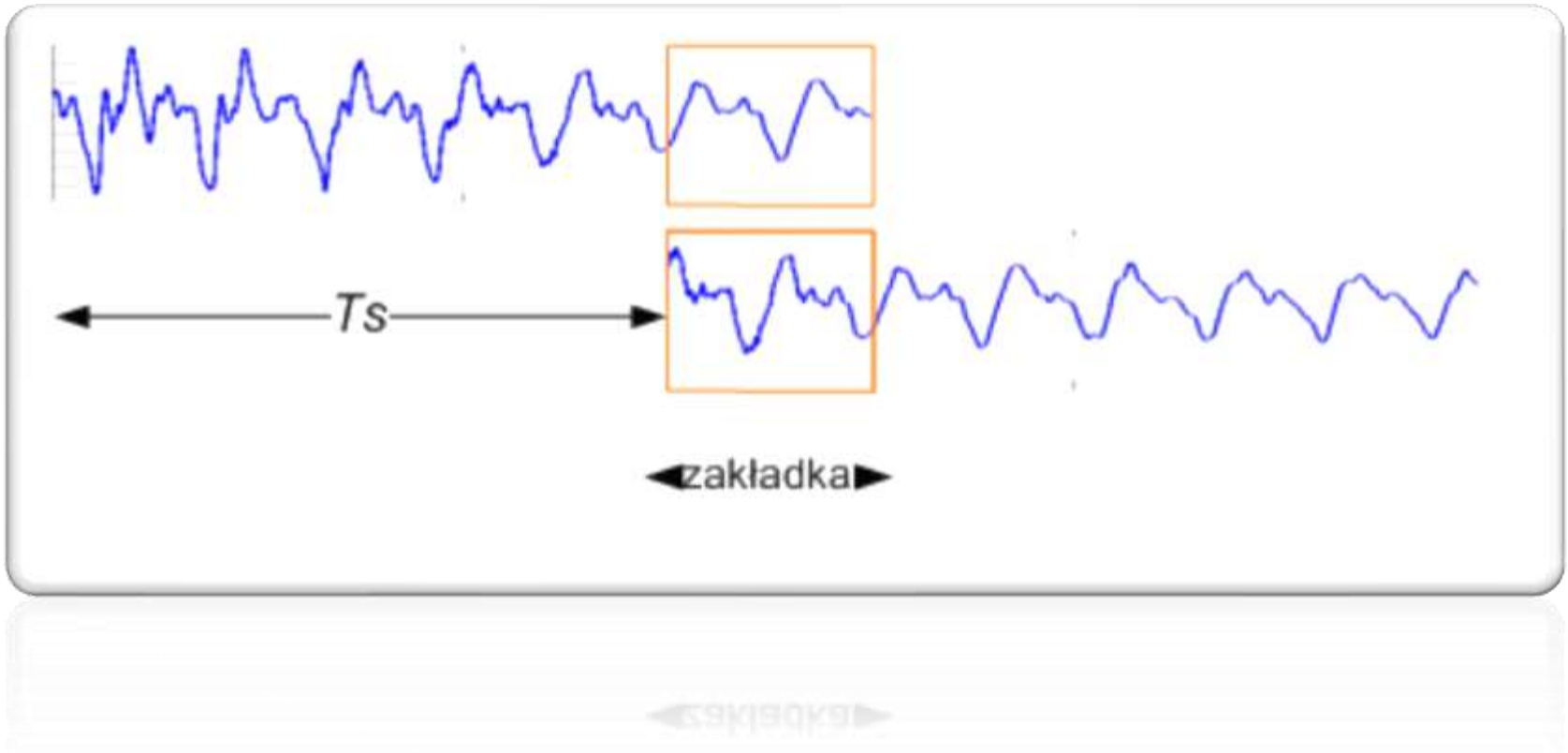
Algorytm SOLA



Algorytm SOLA - Analiza



Algorytm SOLA-Synteza

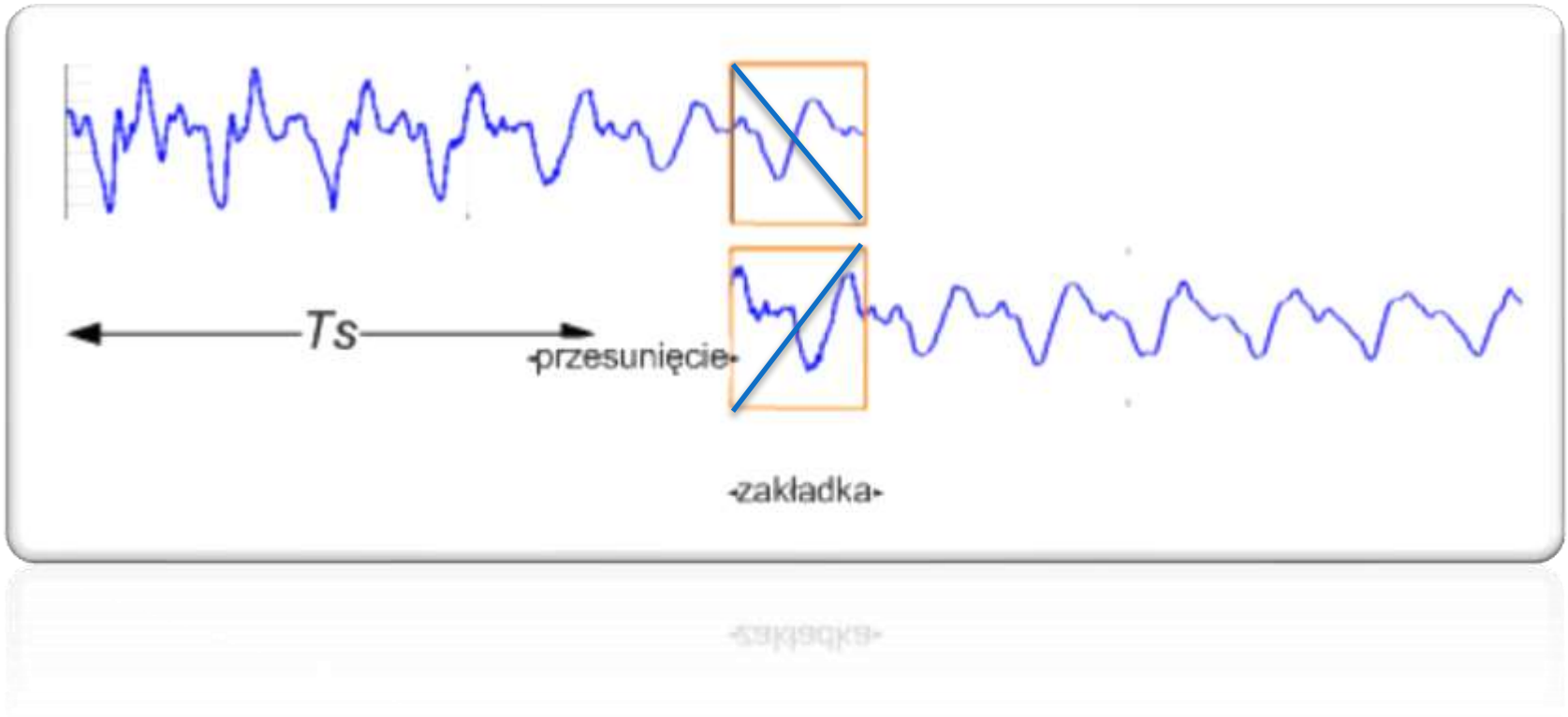


- Wyznaczanie funkcji korelacji skrośnej dla sygnałów zakładki

Algorytm SOLA - Synteza

- Znalezienie pozycji maksimum funkcji

Algorytm SOLA

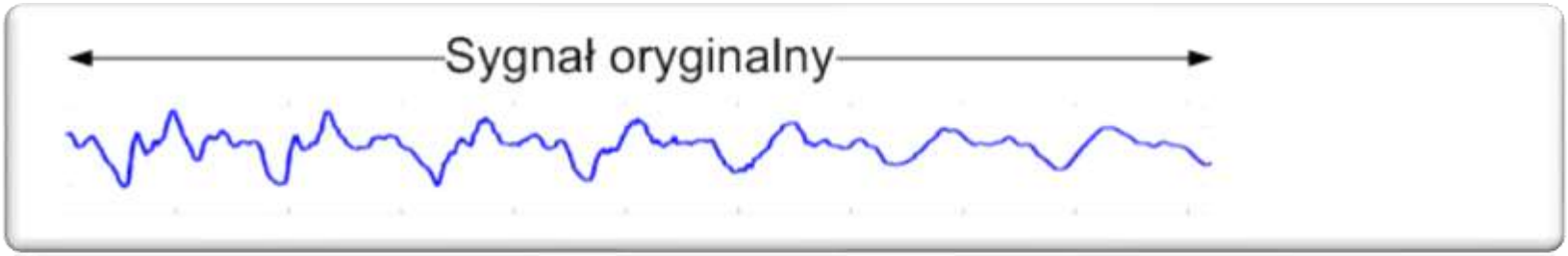


- Korekta obszaru zakładki
- Dla każdej ramki obszar zakładki jest inny

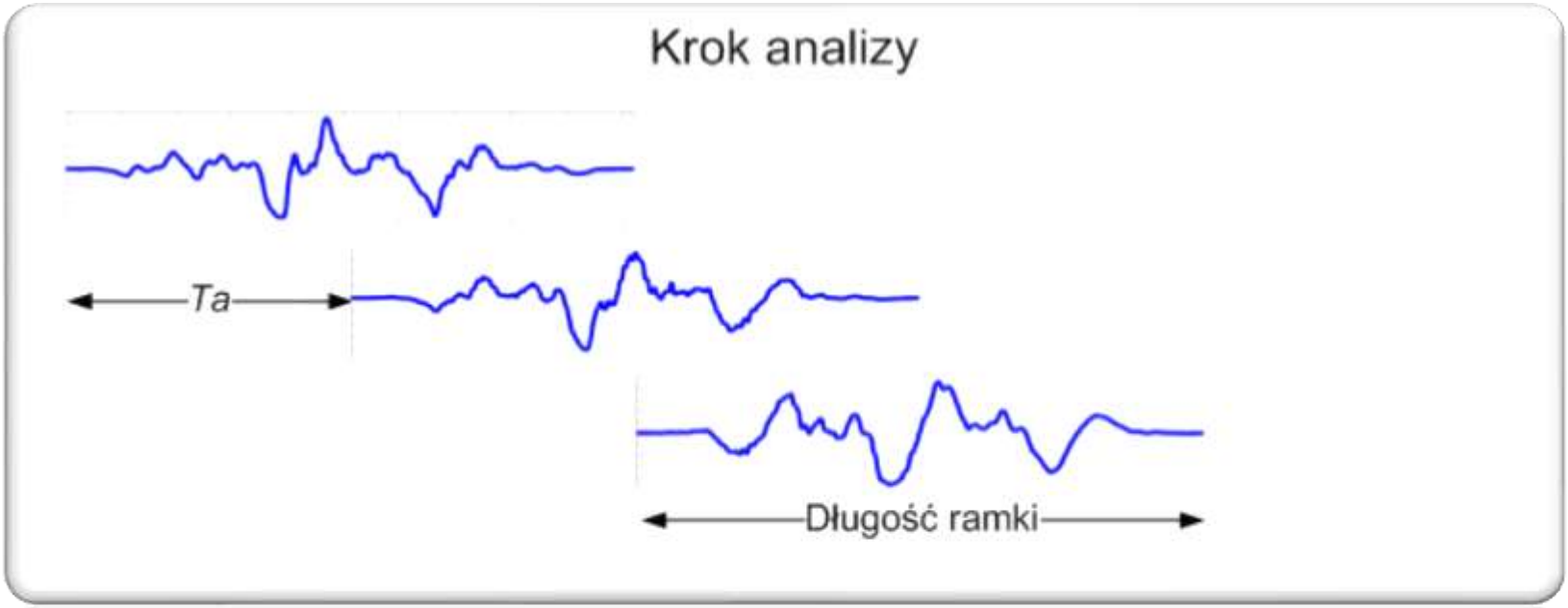
Algorytm SOLA

- Zalety:
  - Wysoka jakość zmodyfikowanego dźwięku
  - Nie słyszalne są nieciągłości w sygnale
- Wady:
  - Konieczność liczenia funkcji korelacji (wymaga wielu obliczeń)
  - Zmienna wartość współczynnika skali

Algorytm Wokodera Fazowego



Algorytm Wokodera Fazowego - Analiza



Algorytm Wokodera Fazowego-Synteza

- Okienkowanie oknem Hamminga
- Obliczanie FFT dla ramki
- Modyfikacji fazy zgodnie ze wzorem:

$$f(n)_{ni} = f(n)_i + \Delta f(n) a$$

gdzie n = {1,2,…,*N*},

$f(n)_{ni}$ - nowa wartość fazy

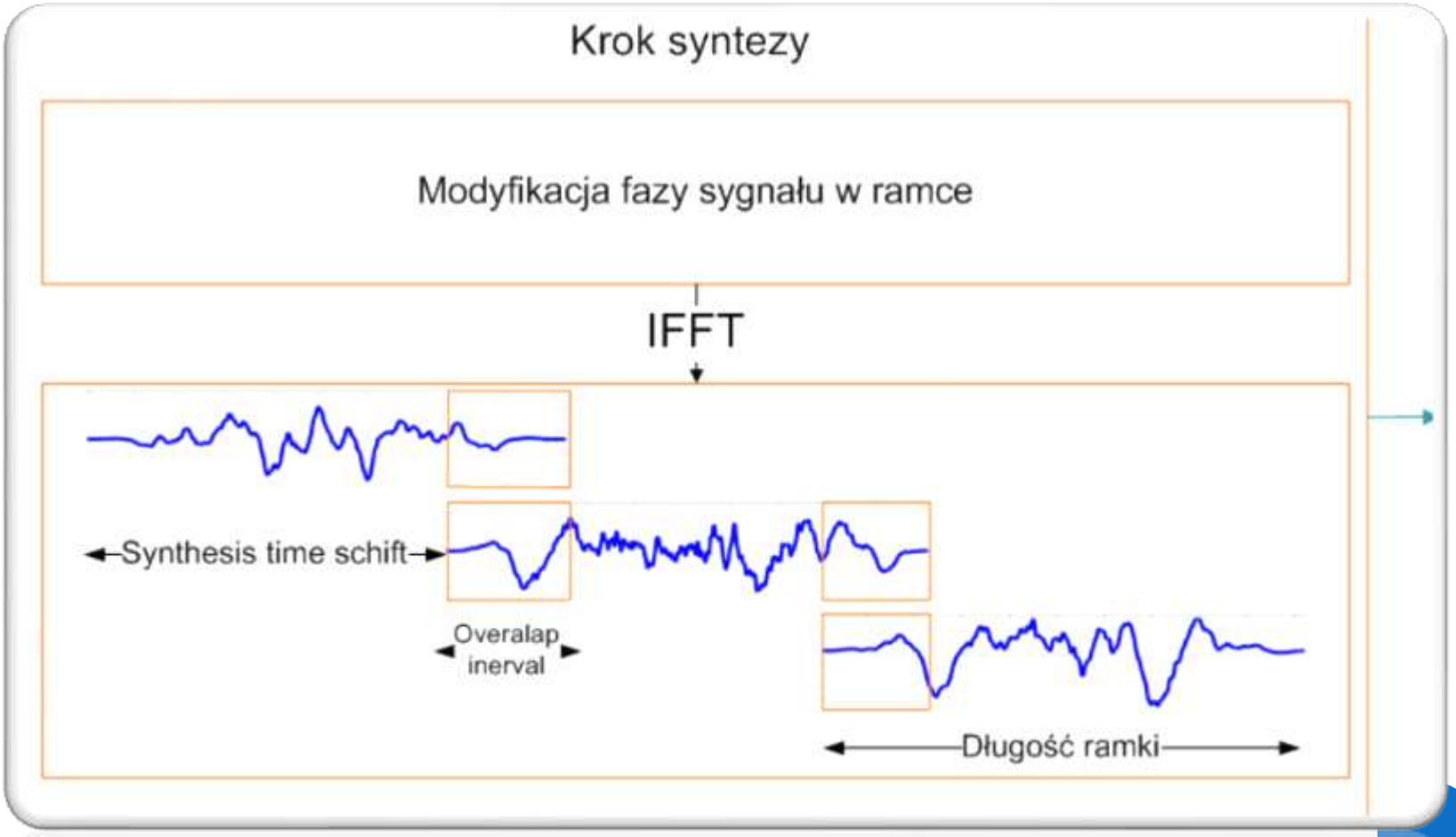
$f(n)_i$ - stara wartość fazy

$\Delta f(n)$ - parametr zależny od zmian $f(n)_i$

a - współczynnik skali

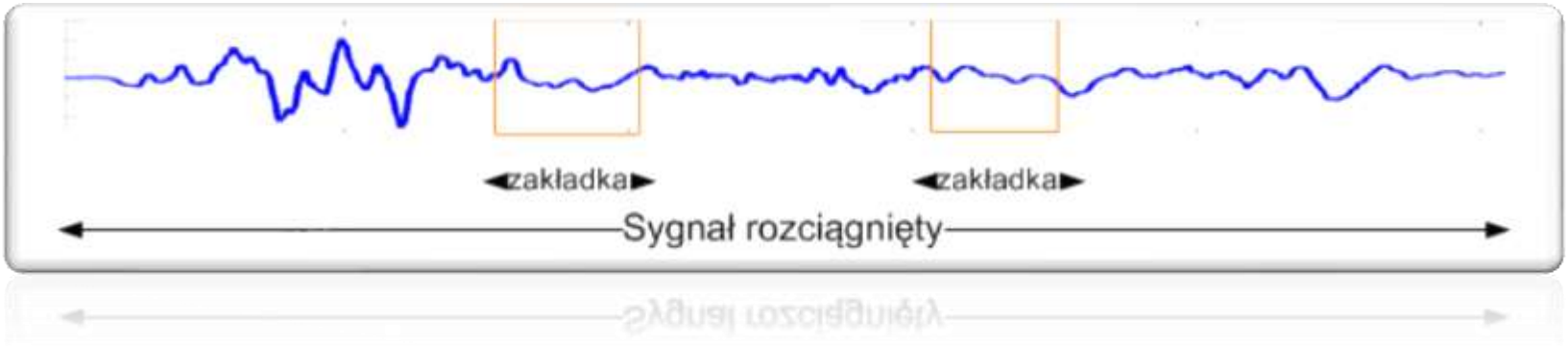
- Modyfikacja fazy pozwala zachować jej ciągłość

Algorytm Wokodera Fazowego-Syntеza



- Sumowanie okien bez cross-fade

Algorytm Wokodera Fazowego



- Zalety:
  - Zachowanie ciągłości fazy
  - Dość dobra jakość dźwięku
  - Niewielka złożoność obliczeniowa
- Wady
  - W sygnale wynikowym słyszalny jest efekt metalicznego „brzęczenia”

Modyfikacja częstotliwości podstawowej

Modyfikacja częstotliwości podstawowej

Skalowanie częstotliwości – zmiana charakterystyki widmowej sygnału przy zachowaniu tempa wypowiedzi

Oryginalny kontur częstotliwości podstawowej (okresu): $P(t)$

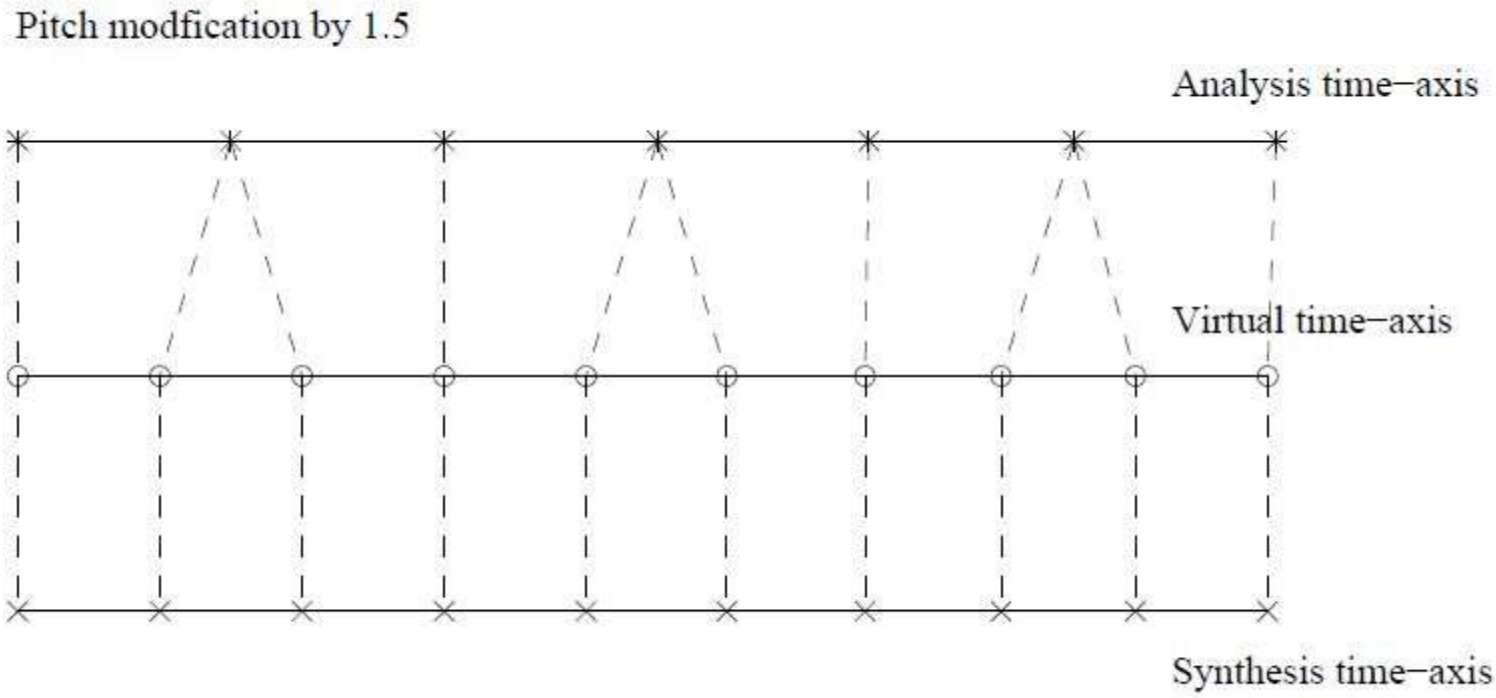
Funkcja skalująca: $\alpha(t)$

Zakładając: $t_a^{i+1} = t_a^i + P(t_a^i)$

Mapowanie: $t_a^i \leftrightarrow t_s^i$

$$t_s^{i+1} - t_s^i = \frac{1}{t_s^{i+1} - t_s^i} \int_{t_s^i}^{t_s^{i+1}} \frac{P(t)}{\alpha(t)} dt$$

Modyfikacja częstotliwości podstawowej



Modyfikacje częstotliwości podstawowej i tempa wypowiedzi

- Algorytmy działające w dziedzinie czasu i częstotliwości
- Najczęściej stosowane – algorytmy PSOLA (Pitch Synchronous OverLap-Add)

  - TD-PSOLA – w dziedzinie czasu

  - FD-PSOLA – w dziedzinie częstotliwości
- Podobne algorytmy:

  - SOLA

  - WSOLA

  - MBROLA

Skalowanie czasu i częstotliwości PSOLA

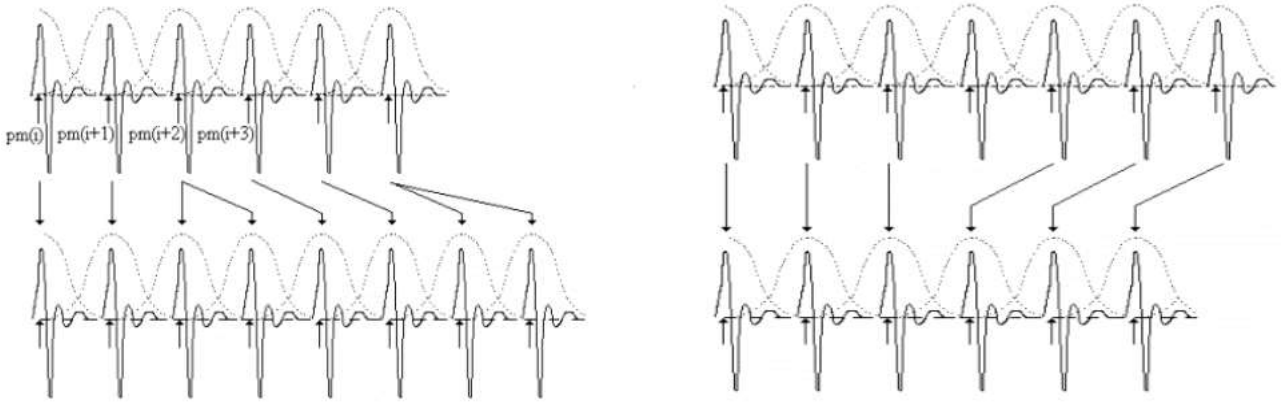
- Przetwarzanie sygnału mowy w krótkich segmentach, a następnie odpowiednie ich połączenie
- Aby uniknąć nieciągłości w miejscach łączenia segmentów stosuje się nakładkowanie oraz odpowiednie okna
- Dla sygnałów okresowych (lub prawie okresowych) sensownym jest dopasowanie długości okna do długości okresu
- W algorytmie PSOLA (opracowanym specjalnie dla przetwarzania mowy), długość kolejnych okien dobierana jest zgodnie z wartością estymowanej częstotliwości podstawowej
- Najlepszymi znacznikami początków ramek byłyby chwile zamknięcia głośni, jednak ze względu na trudność ich wyznaczenia stosuje się inne znaczniki (np. CoG)

Skalowanie czasu i częstotliwości
TD-PSOLA

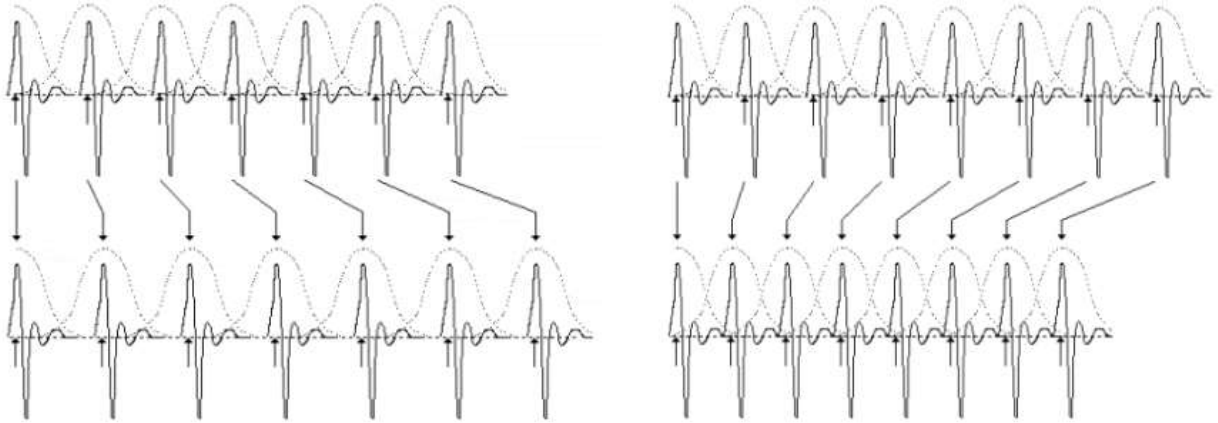
- Estymacja częstotliwości podstawowej
- Podział sygnału mowy na segmenty synchronicznie z estymowaną częstotliwością podstawową (dla bezdźwięcznych fragmentów mowy długość segmentów jest z góry określona i stała).
- Modyfikacja sygnału.
- Rekonstrukcja sygnału poprzez złożenie segmentów z zastosowaniem zakładek.

Skalowanie czasu i częstotliwości
TD-PSOLA

Skalowanie czasu:



Skalowanie częstotliwości:



Skalowanie częstotliwości
FD-PSOLA

- Podział sygnału mowy na segmenty synchronicznie z częstotliwością tonu krtaniowego (nie jest konieczne wyznaczenie dokładnych chwil początków okresów, wystarczy dobry algorytm estymacji częstotliwości podstawowej).
- Obliczenie krótkookresowego widma sygnału, estymowanie obwiedni widma, obliczenie pobudzenia (w tym kroku istotnym jest, by widmo uzyskanego pobudzenia było płaskie, należy więc skorzystać z dobrego algorytmu wyznaczania obwiedni widma).
- Modyfikacje częstotliwości podstawowej.
- Rekonstrukcja sygnału (przejście z dziedziny częstotliwości do dziedziny czasu - może się okazać, że po dokonaniu transformacji nie istnieje sygnał rzeczywisty, który odpowiadałby uzyskanemu widmu - należy wyznaczyć widmo, dla którego istnieje sygnał rzeczywisty, a które jest jak najbardziej zbliżone do widma syntetycznego).

Skalowanie częstotliwości
FD-PSOLA

Sposoby modyfikacji częstotliwości:

- usuwanie (dla obniżenia) lub dodawanie (dla podwyższenia) harmonicznych w widmie sygnału - konieczna jest dokładna estymacja częstotliwości podstawowej dla wyznaczenia harmonicznych, muszą być zachowane zależności fazowe między poszczególnymi harmonicznymi;
- kompresja/ekspansja widma pobudzenia - oryginalna oś częstotliwości jest „zawijana” (ang. *warping*) z wykorzystaniem współczynnika skalującego β (wprowadza zniekształcenia, jednak jeśli obwiednia widma została wyznaczona poprawnie, będą one mniejsze niż w przypadku usuwania/dodawania harmonicznych)

$$Y(k_s) = (1-\alpha)X(k_v) + \alpha X(k_v + 1)$$

$$\alpha = k_s - \frac{k}{\beta}$$

Dyskusja

Algorytmy TD:

- Szybkie, wymagają małych mocy obliczeniowych – sprawdzają się w systemach czasu rzeczywistego
- Bardzo dobre rezultaty przy małych współczynnikach skalowania
- Problem powtarzania transjentów

Algorytmy FD

- Bardziej złożone obliczeniowo
- Najczęściej wymagają obliczenia parametrów modelu (wyższa jakość)
- Przewyższają algorytmy TD w przypadku dużych współczynników skalowania.

Transformacje głosu

Modyfikacja charakterystyki traktu głosowego

Skorzystanie z algorytmu PSOLA – przepróbkowanie segmentów mowy przed ich ponownym złożeniem

- Aby uzyskać podniesienie częstotliwości środkowych formantów (przeskalowanie przez $\gamma>1$) należy zmniejszyć częstotliwość próbkowania γ razy.
- Segmenty są dodawane z oryginalną częstotliwością zmienia się więc położenie formantów, ale częstotliwość podstawowa i tempo wypowiedzi pozostają niezmienione.
- Mała złożoność obliczeniowa, ale częstotliwości środkowe formantów można zmieniać tylko liniowo.

Modyfikacja charakterystyki traktu głosowego

- Estymacja charakterystyki traktu głosowego (obwiedni widma sygnału mowy)
- Zamodelowanie charakterystyki traktu głosowego
- Modyfikacja zgodnie z założonymi regułami
- Synteza

Modyfikacja charakterystyki traktu głosowego

Modyfikacja obwiedni widma sygnału mowy

- Różne techniki estymacji obwiedni widma
- Aby podnieść częstotliwości środkowe formantów γ-krotnie należy zmodyfikować widmo sygnału zgodnie ze wzorem

$$Y(t,\Omega_k) = X(t,\Omega_k)E(t,\Omega_k/\gamma)/E(t,\Omega_k)$$

- Możliwość transformacji nieliniowych (czynnik γ zmienny w czasie)

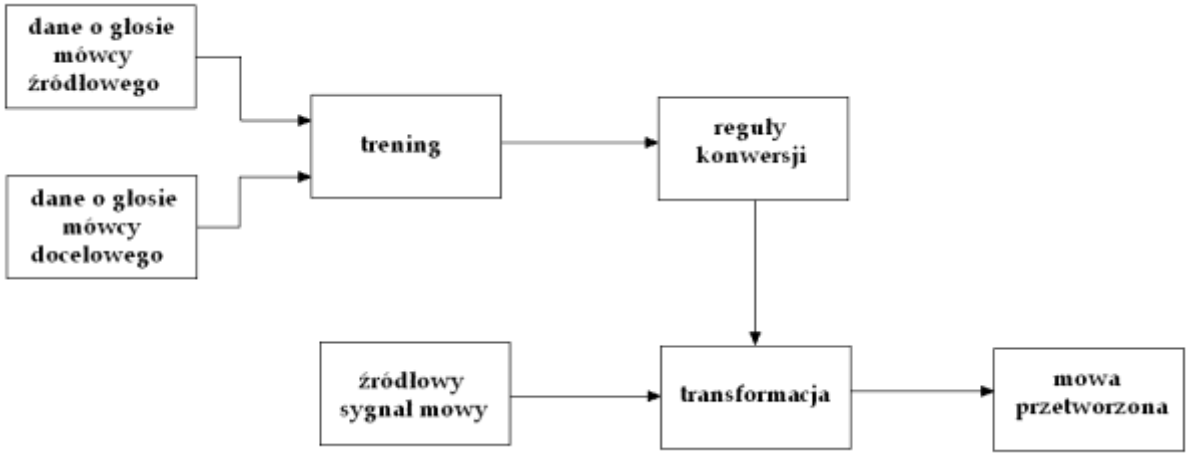
Modyfikacja charakterystyki traktu głosowego

Modyfikacja biegunów filtru modelującego trakt głosowy

- Transmitancja filtru estymowana jest za pomocą predykcji liniowej.
- Bieguny transmitancji zapisywane są w postaci biegunowej $re^{j\varphi}$
- Zmiana kąta φ powoduje przesunięcie formantu na osi częstotliwości.
- Zmiana promienia r powoduje zwężenie lub poszerzenie zajmowanego przez formant pasma.
- Dokonując transformacji należy pamiętać, by bieguny znajdowały się zawsze wewnątrz okręgu jednostkowego.

KONWERSJA GŁOSU

Automatyczna transformacja głosu mówcy źródłowego do głosu mówcy docelowego z zachowaniem treści wypowiedzi.



System konwersji gromadzi dane o głosach mówcy źródłowego i docelowego (odpowiednie próbki głosów) i na ich podstawie automatycznie generuje reguły konwersji w procesie treningu. Reguły te są następnie wykorzystywane w procesie transformacji głosu źródłowego tak, by odpowiadał charakterystyce głosu docelowego.

Konwersja głosu

Istnieje wiele systemów konwersji mowy, opierających się na różnych modelach mowy i metodach modyfikacji, jednak w każdym podejściu należy rozwiązać trzy podstawowe problemy:

1. wyodrębnienie cech charakterystycznych mówców z przebiegów akustycznych mowy,
2. opracowanie metody mapowania cech mówców źródłowego i docelowego,
3. modyfikacja charakterystyki głosu mówcy źródłowego, tak by brzmiał jak głos mówcy źródłowego, z wykorzystaniem opracowanego wcześniej schematu mapowania.

Konwersja głosu

- Przed analizą z reguły należy również zgromadzić odpowiednią bazę danych wypowiedzi mówcy źródłowego i docelowego (najczęściej te same wypowiedzi) oraz odnaleźć odpowiadające sobie ramki sygnału w wypowiedziach (za pomocą ukrytych modeli Markova lub nieliniowej transformacji czasu).
- Dla mapowania charakterystyk mówców, czyli znajdowania funkcji zależności między cechami mówcy źródłowego i docelowego, większość systemów konwersji korzysta z trzech podstawowych narzędzi: kwantyzacji wektorowej VQ (ang. *Vector Quantization*), liniowej kombinacji rozkładów normalnych GMM (ang. *Gaussian Mixture Model*) i sztucznych sieci neuronowych ANN (ang. *Artificial Neural Networks*).

KLONOWANIE GŁOSU

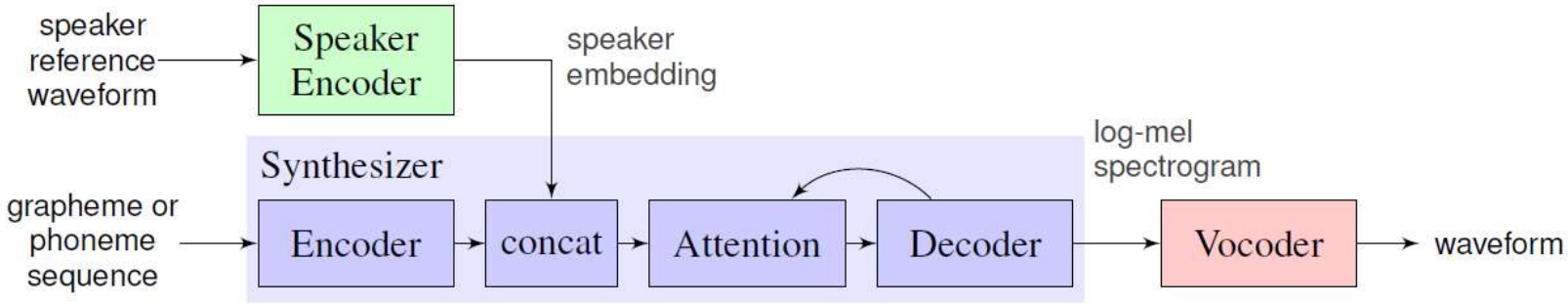


Figure 1: Model overview. Each of the three components are trained independently.

Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis, https://arxiv.org/pdf/1806.04558.pdf

Ta sztuczna inteligencja klonuje Twój głos po nasłuchiwaniu przez 5 sekund https://www.youtube.com/watch?v=0sR1rU3gLzQ

Mission: Impossible 3 (2006) - Seeing Double Scene (5/8) | Movieclips https://www.youtube.com/watch?v=CgX4uJSj00Y

Morphing głosu

- Analogia do morphingu obrazów.
- Stopniowe przechodzenie od głosu mówcy źródłowego do głosu mówcy docelowego .
- Potrzebne są takie same zdania wypowiadane przez obu mówców.
- Konieczne jest odnalezienie odpowiadających sobie segmentów w wypowiedziach obu mówców (fonosegmentacja, nieliniowa transformacja czasu DTW).
- Oddzielnie przeprowadzana jest modyfikacja pobudzenia (np. za pomocą technik SOLA) i charakterystyki traktu głosowego (transformata Fouriera, współczynniki LPC, PARCOR…)

Dziękuję za uwagę!