

ANALIZA I SYNTEZA MOWY

Opracował:
Adam
Kupryjanow

Plan wykładu

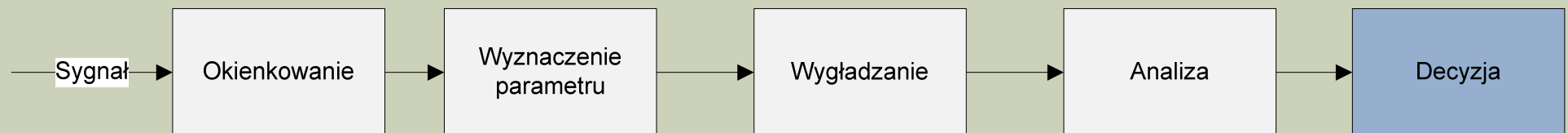
- Metody detekcji samogłosek
- Algorytmy modyfikacji czasu trwania sygnału

DETEKCJA SAMOGŁÓSEK

- Zastosowania:
 - Telefonia - kodowanie mowy
 - Analiza sygnału mowy - segmentacja
 - Rozpoznawanie samogłosek
 - Systemy rozpoznawanie mowy
 - ...

METODY DETEKCJI SAMOGŁÓSEK

- Metody statystyczne najczęściej progowe:
 - Analiza energii sygnału + liczby przejść przez zero:
 - Spectral Peaks Energy
 - Spectral Band Energy Cumulating (Sbec)
 - Peak-valley difference (PVD)
- Metody inteligentne -> parametryzacja sygnału + klasyfikator:
 - SVM
 - Sztuczne sieci neuronowe



Analiza energii sygnału

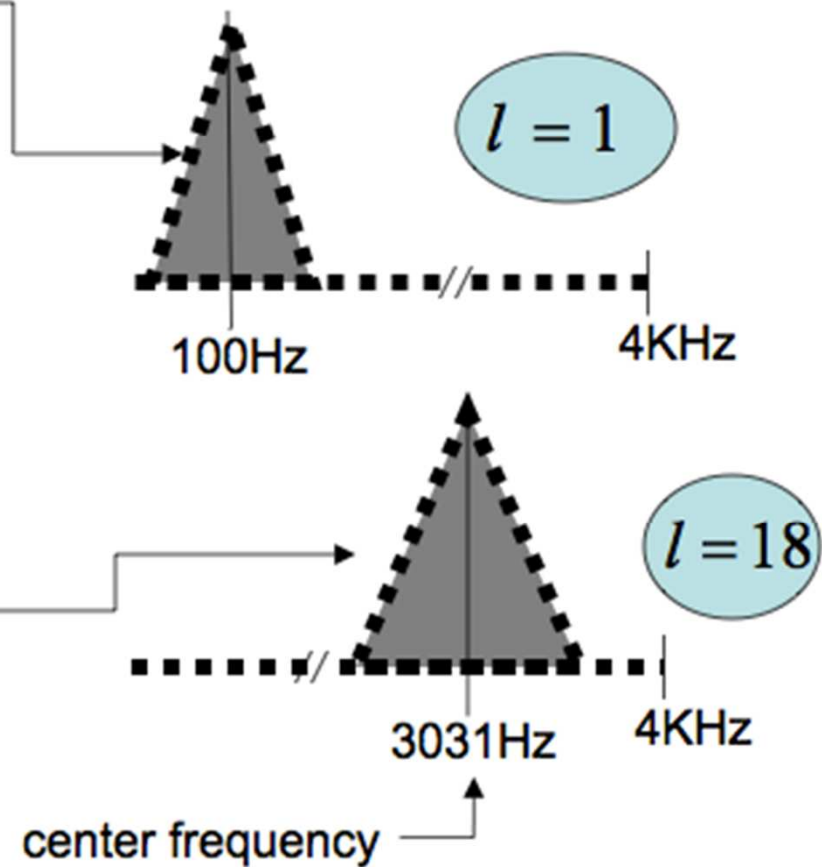
- **Samogłoski:**
 - wysoki poziom energii
 - charakterystyczne piki w widmie
 - niewielka liczba przejść przez zero
- **Spółgłoski:**
 - niski poziom energii
 - płaskie widmo
 - wysoka liczba przejść przez zero
- Prosta progowa analiza nie daje dobrych rezultatów!

Trójkątne Filtry melowe

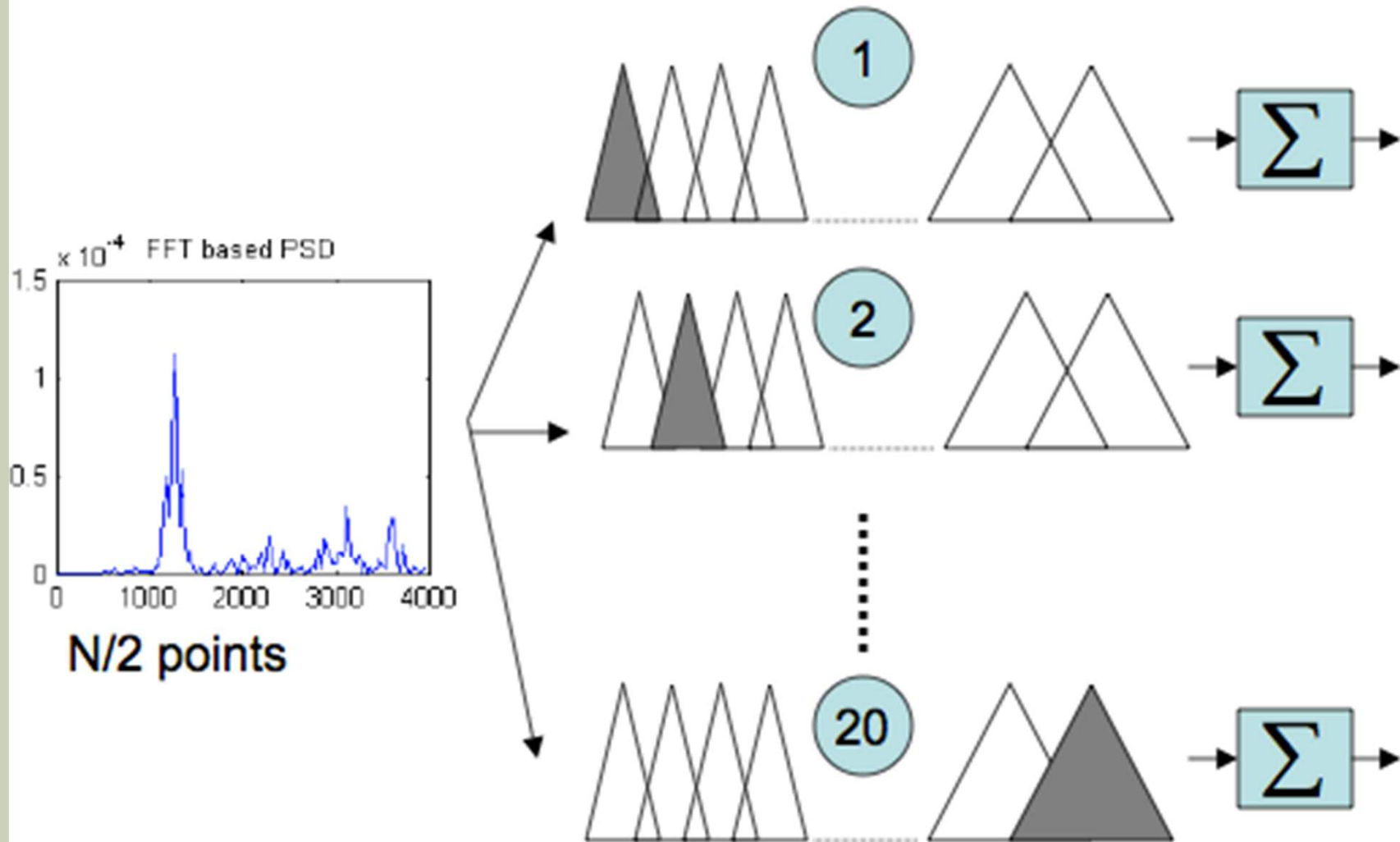
Index	Bark Scale		Mel Scale	
	Center Freq. (Hz)	BW (Hz)	Center Freq. (Hz)	BW (Hz)
1	50	100	100	100
2	150	100	200	100
3	250	100	300	100
4	350	100	400	100
5	450	110	500	100
6	570	120	600	100
7	700	140	700	100
8	840	150	800	100
9	1000	160	900	100
10	1170	190	1000	124
11	1370	210	1149	160
12	1600	240	1320	184
13	1850	280	1516	211
14	2150	320	1741	242
15	2500	380	2000	278
16	2900	450	2297	320
17	3400	550	2639	367
18	4000	700	3031	422
19	4800	900	3482	484
20	5800	1100	4000	556
21	7000	1300	4595	639
22	8500	1800	5278	734
23	10500	2500	6063	843
24	13500	3500	6964	969

$$M_l(k) \quad l = 0, 1, \dots, L-1$$

$$k = 0, 1, \dots, N/2$$



BANK filtrów Skali melowej



SPECTRAL BAND ENERGY CUMULATING (SBEC)

$$SBEC(t) = \sum_{i=1}^{24} \alpha_i \left| E_i(t) - \bar{E}_i(t) \right|$$

i – numer filtru

$E_i(t)$ – energia sygnału i -tego filtru

$\bar{E}_i(t)$ – średnia energia sygnału i -tego filtru

t – numer analizowanej ramki

α_i – współczynnik wagi i -tego filtru

SPECTRAL BAND ENERGY CUMULATING (SBEC)

- Maksima w przebiegu $SBEC(t)$ wyższe od wartości progu odpowiadają miejscom występowania głosek dźwięcznych
- Wartość progu podlega adaptacji
- W zwiększenia skuteczności analizowane są tylko fragmenty trwające dłużej niż 32 ms
- Algorytm wykazuje dużą liczbę błędów typu false-positive

REC (REDUCED ENERGY CUMULATING)

$$REC(t) = \sum_{i=1}^{24} \alpha_i |E_i(t) - \bar{E}_i(t)|$$

$$REC(t) = REC_{LF}(t) + REC_{HF}(t)$$

$REC_{LF}(t)$ – parametr $REC(t)$ wyznaczony dla częstotliwości poniżej 1 kHz

$REC_{HF}(t)$ – parametr $REC(t)$ wyznaczony dla częstotliwości powyżej 1 kHz

Warunek analizy maksimum parametru $REC(t)$:

$$\begin{cases} \frac{REC_{LF}(t)}{REC(t)} \geq 0.5 \\ \Delta t \geq 15ms \end{cases}$$

PEAK VALLEY-DIFFERENCE (PVD)

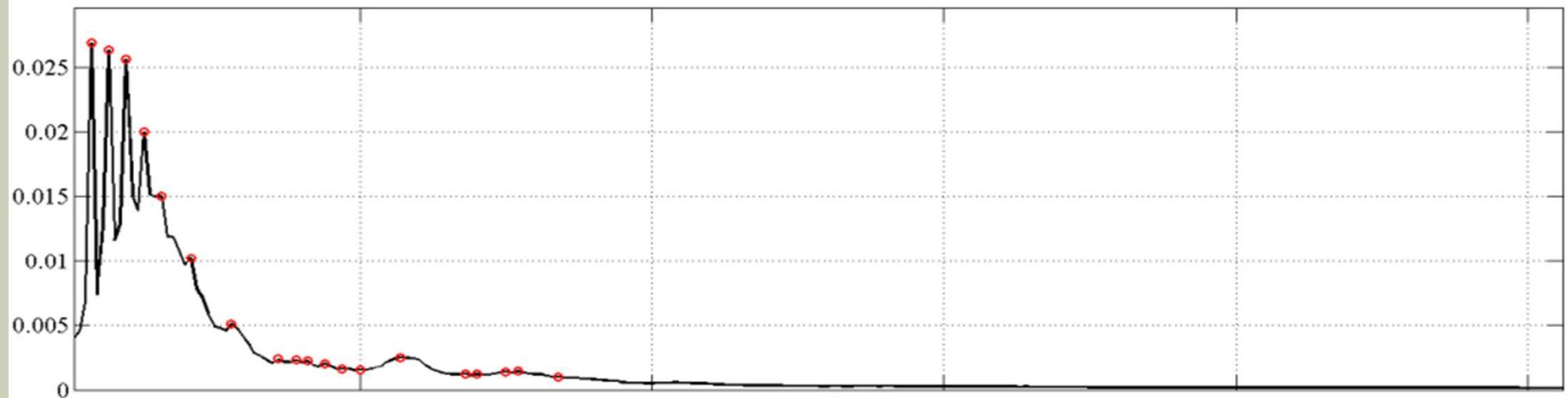
- Analiza sygnału w ramkach czasowych
- Algorytm oparty na analizie parametru PVD (peak-valley difference)

$$PVD(VM, A) = \frac{\sum_{k=0}^{N-1} (A(k) \cdot VM(k))}{\sum_{i=0}^{n-1} VM(k)} - \frac{\sum_{k=0}^{N-1} (A(k) \cdot (1 - VM(k)))}{\sum_{k=0}^{N-1} (1 - VM(k))}$$

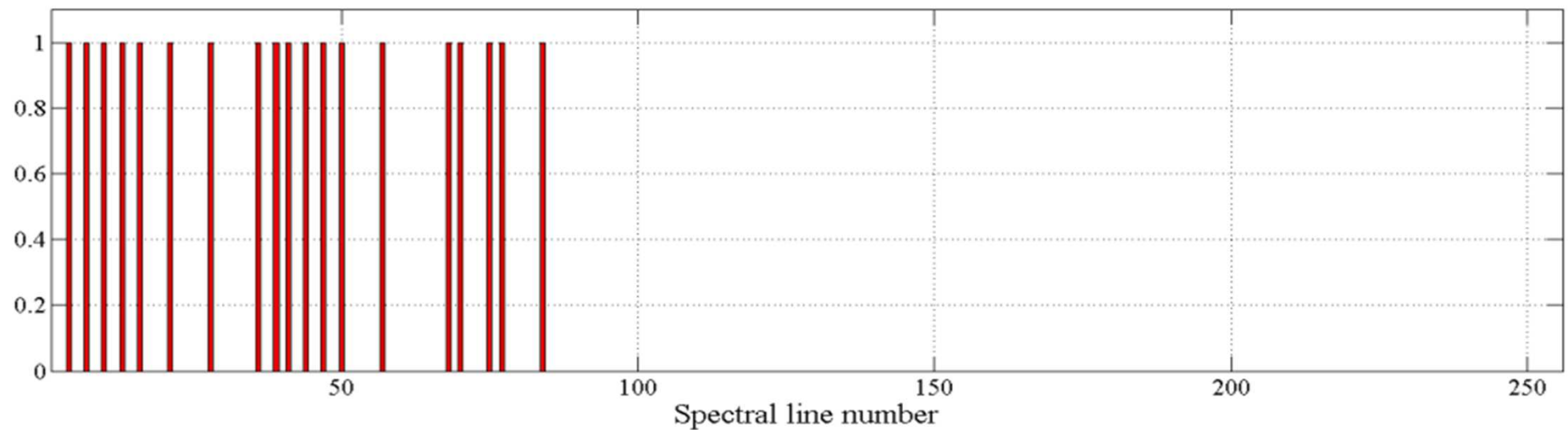
- Wyznaczenie modelu samogłoski VM:
 - wyznaczenie średniego widma amplitudowego dla zbioru samogłosek
 - znalezienie szczytów w uśrednionym widmie
 - stworzenie wektora VM zawierającego 1 w miejscach szczytów w pozostałych miejscach

MODEL VM

Mean amplitude spectrum

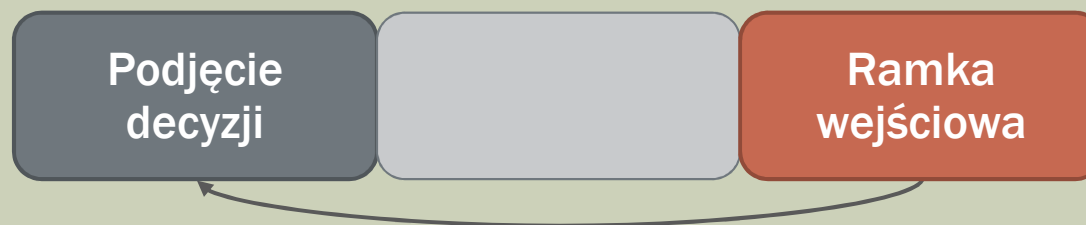


Vowel model vector



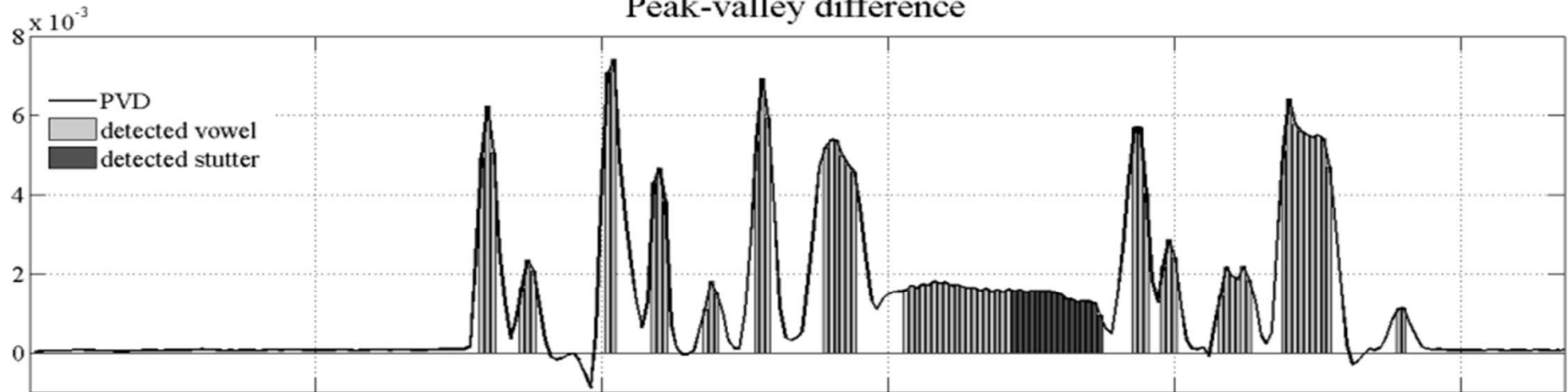
PVD – ALGORYTM DETEKCJI

- Wygładzenie wartości PVD
- Znalezienie szczytów w przebiegu PVD
- Samogłoski występują w ramach dla których wartość PVD jest większe od 70% najbliższego szczytu

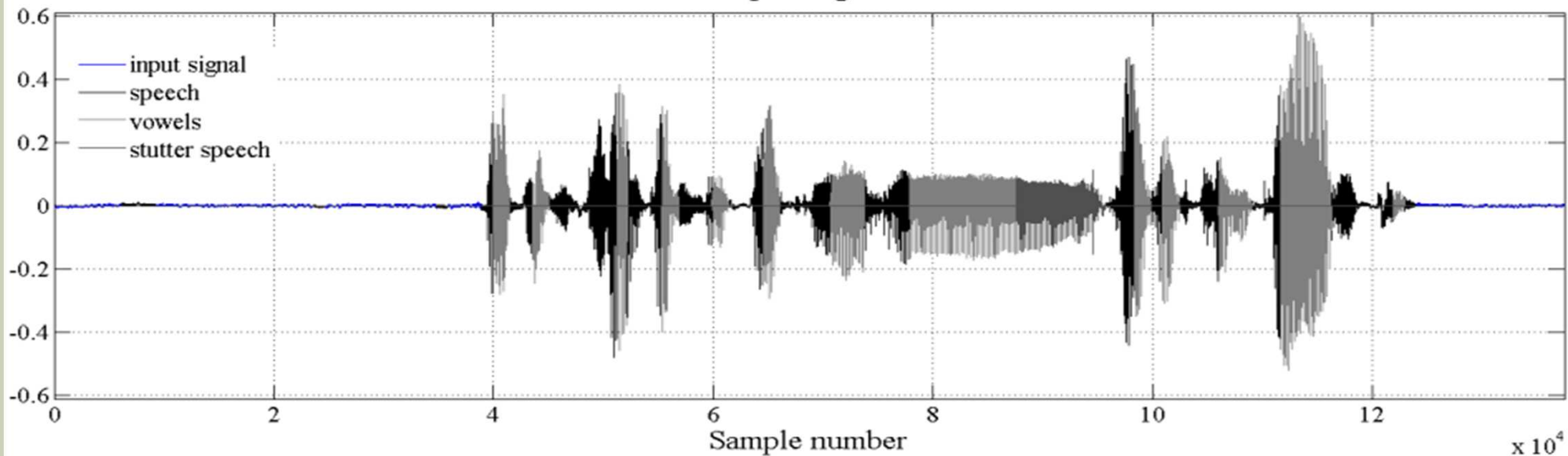


PVD – ALGORYTM DETEKCJI

Peak-valley difference



Input signal



ALGORYTMY MODYFIKACJI CZASU TRWANIA SYGNAŁU

- Założenia:
 - Brak zmiany wysokości dźwięku
 - Wprowadzanie jak najmniejszej liczby niekształceń:
 - Nieciągłości fazy i częstotliwości
 - Trzasków
 - Powtarzania transjentów
 - Osiągnięcie największego możliwego podobieństwa sygnału wejściowego
- Zastosowania:
 - Synteza mowy
 - Dopasowanie czasu trwania wypowiedzi np. audio booki, 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

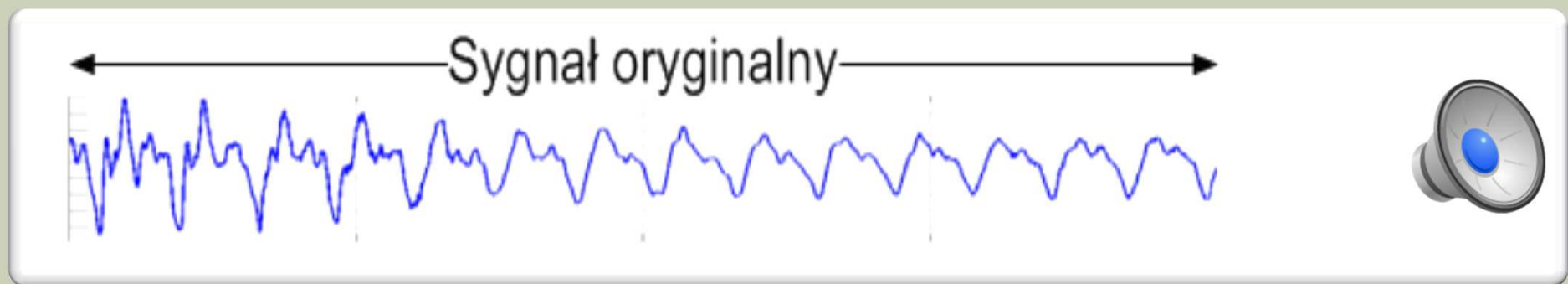
- Algorytmy działające po stronie czasu:
 - OLA (Overlap and Add)
 - SOLA (Synchronous Overlap and Add)
 - PSOLA (Pitch-synchronous Overlap and Add)
 - WSOLA (Waveform Similarity Overlap and Add)
 - PAOLA (Peak Alignment Overlap and Add)
- Algorytmy działające po stronie widma:
 - FD-PSOLA
 - Wokoder-fazowy

ALGORYTMY – WSPÓŁCZYNNIK SKALI

$$T_s = \alpha \cdot T_a$$

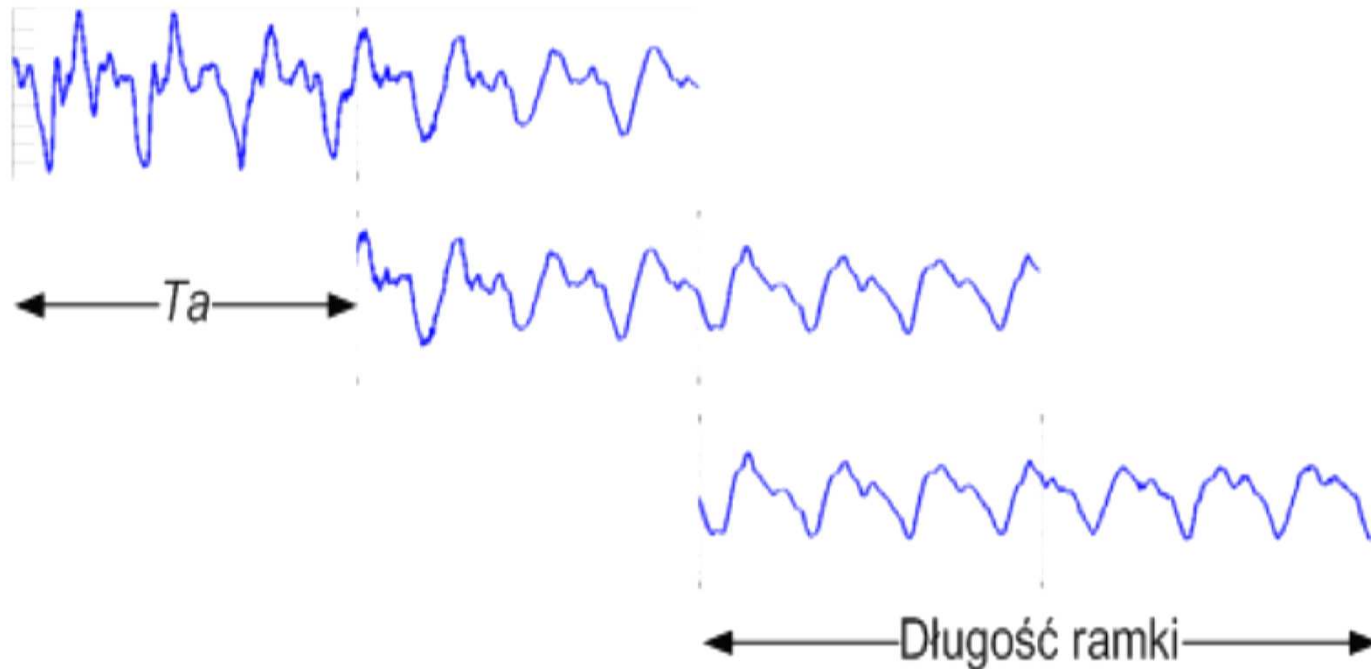
gdzie T_s – przesunięcie czasowe syntezy,
 T_a – przesunięcie czasowe analizy,
 α – współczynnik skali.

ALGORYTM OLA



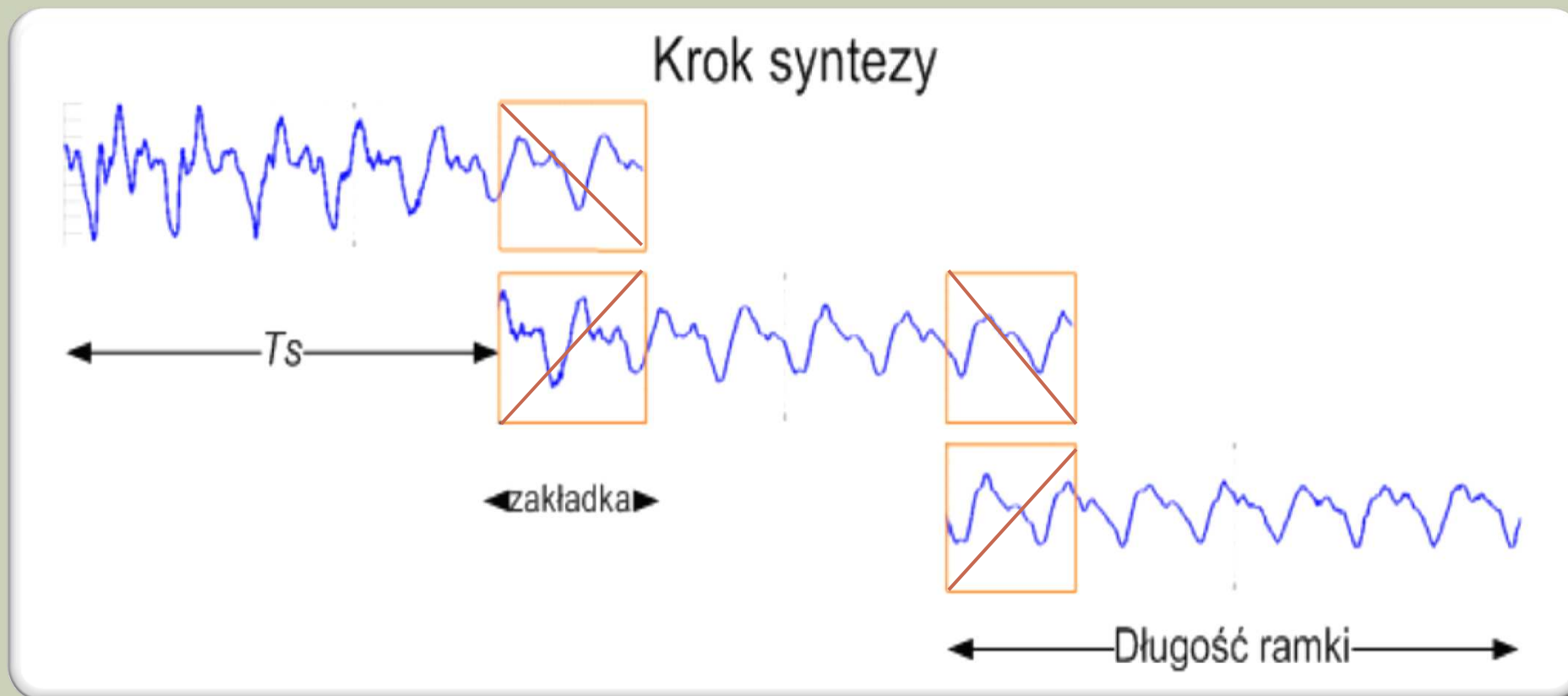
ALGORYTM OLA - ANALIZA

Krok analizy



Długość ramki

ALGORYTM OLA - SYNTEZA



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

ALGORYTM OLA



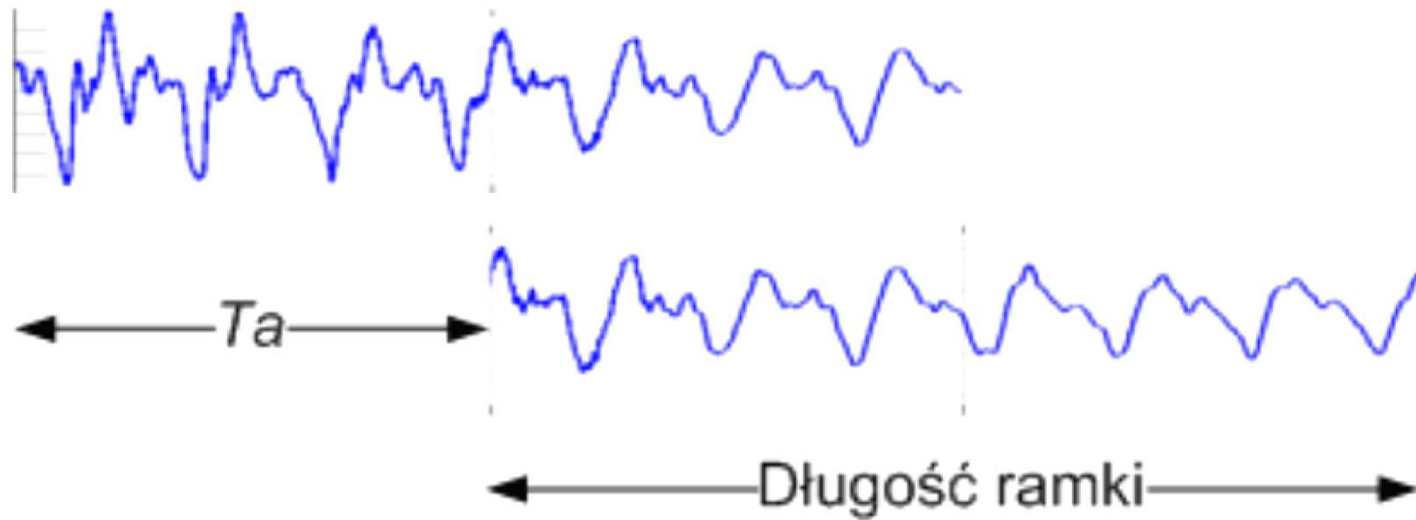
- **Zalety:**
 - Niewielka złożoność obliczeniowa
 - Szybki
- **Wady:**
 - Sygnał wynikowy jest niskiej jakości
 - Słyszalne są trzaski na łączeniach ramek
 - Występują nieciągłości fazy i częstotliwości

ALGORYTM SOLA



ALGORYTM SOLA - ANALIZA

Krok analizy



Długość ramki

ALGORYTM SOLA-SYNTTEZA



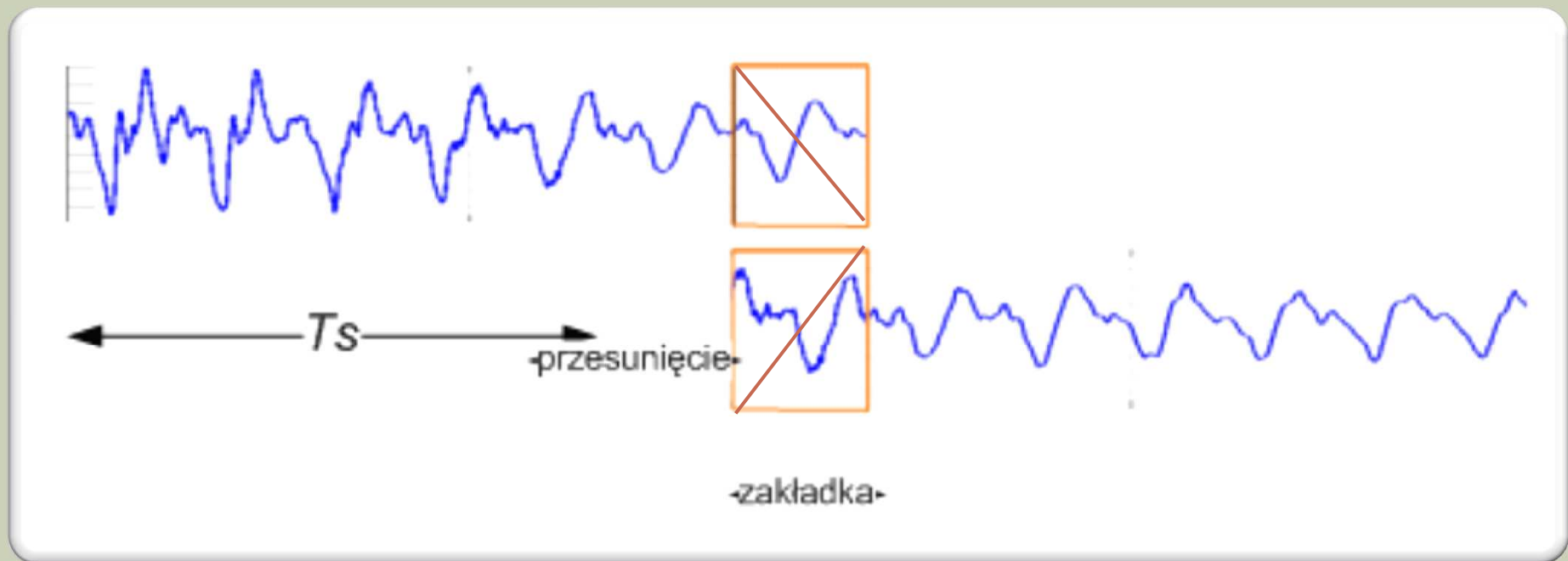
- Wyznaczanie funkcji korelacji skrośnej dla sygnałów zakładki

ALGORYTM SOLA - SYNTEZA



- Znalezienie pozycji maksimum funkcji

ALGORYTM SOLA



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

ALGORYTM SOLA



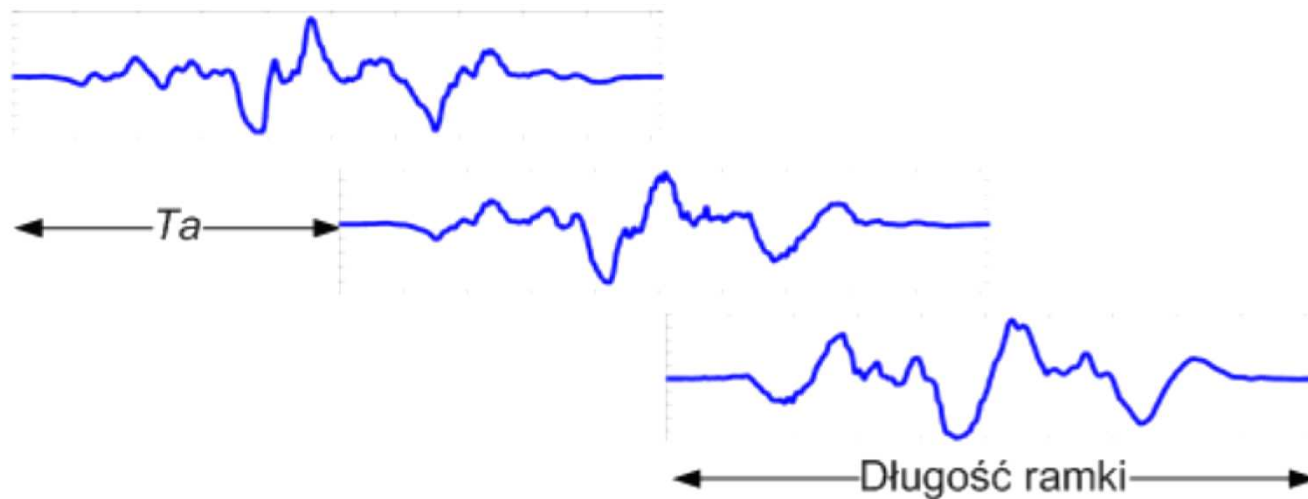
- Zalety:
 - Wysoka jakość zmodyfikowanego dźwięku
 - Nie słyszalne są nieciągłości w sygnale
- Wady:
 - Konieczność liczenia funkcji korelacji (wymaga wielu obliczeń)
 - Zmienna wartość współczynnika skali

ALGORYTM WOKODERA FAZOWEGO



ALGORYTM WOKODERA FAZOWEGO - ANALIZA

Krok analizy



Długość ramki

ALGORYTM WOKODERA FAZOWEGO- SYNTEZA

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

$$\phi(n)_{ni} = \phi(n)_i + \Delta\phi(n)\alpha$$

gdzie $n = \{1, 2, \dots, N\}$,

$\phi(n)_{ni}$ - nowa wartość fazy

$\phi(n)_i$ - stara wartość fazy

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

α - współczynnik skali

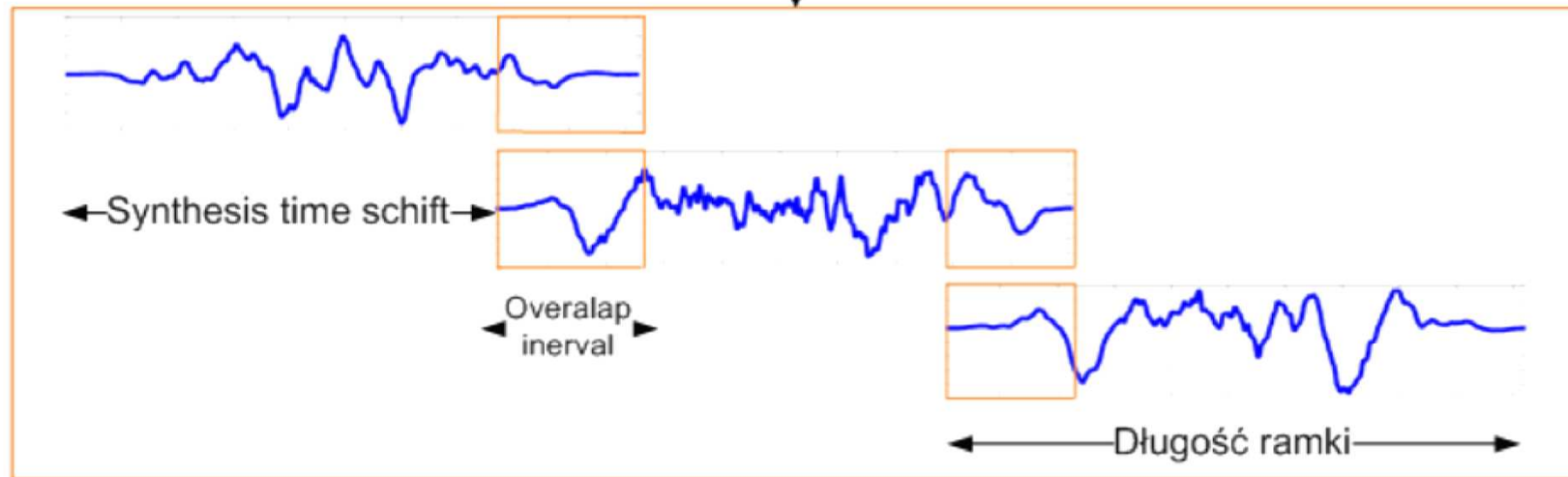
- Modyfikacja fazy pozwala zachować jej ciągłość

ALGORYTM WOKODERA FAZOWEGO-SYNTeza

Krok syntezy

Modyfikacja fazy sygnału w ramce

IFFT



- 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”

OCENA JAKOŚCI ZMODYFIKOWANEGO SYNGAŁU

- Subiektywna – wykonanie testów z udziałem grupy eksperckiej
- Obiektywna – wyznaczenie parametru opisującego jakość nagrania po modyfikacji

$$Dm = \frac{\sum_{u=P}^{U-P-1} \sum_{k=0}^{N-1} [|Y(\alpha Ta_u, \omega_k)| - |X(Ta_u, \omega_k)|]^2}{\sum_{u=P}^{U-P-1} \sum_{k=0}^{N-1} |X(Ta_u, \omega_k)|^2}$$

$X(Ta_u, \omega_k)$ – widmo amplitudowe jednej ramki sygnału wejściowego $x(n)$

$Y(\alpha Ta_u, \omega_k)$ – widmo amplitudowe jednej ramki sygnału zmodyfikowanego $y(n)$

u – numer ramki

P – numer pierwszej i ostatniej ramki, które są wyłączone z procesu analizy w celu wyeliminowania błędów

Ta_u – przesunięcie czasowe syntezy dla ramki numer u

α – współczynnik skali

BIBLIOGRAFIA

- Pellegrino F., Andre-obreht R., *From vocalic detection to automatic emergence of vowel systems*, Proc. ICASSP'97, p. 1651-1652.
- Dorran, D., Lawlor, R., Coyle, E. (2003). High quality time-scale modification of speech using a peak alignment overlap-add algorithm (PAOLA).
- Ergoul, O., Karagoz, I. (1997). Time-scale modification of speech signals for language-learning impaired children.
- Grofit, S., Lavner, Y. (2008). Time-Scale Modification of Audio Signals Using Enhanced WSOLA With Management of Transients, IEEE Trans. On audio, speech, and language processing, vol. 16, no. 1.
- Laroche, J. (1999). Improved Phase Vocoder Time-Scale Modification of Audio, IEEE Trans. On audio, speech, and language processing, vol. 7 no. 3.
- Nejime, Y., Aritsuka, T., Imamura, T., Ifukube, T., Matsushima J. (1996). A portable digital speech-rate converter for hearing impairment, IEEE Trans. Rehabil. Eng., vol. 4, no. 2, pp. 73-83.
- Zolzer, U. (2005). DAFX Digital Audio Effects, Wiley.