Perceptual Wavelet Decomposition for Speech Segmentation

Mariusz Ziółko, Jakub Gałka, Bartosz Ziółko, Tomasz Drwięga

POLSKA PLATFORMA BEZPIECZEŃSTWA WEWNĘTRZNEGO
PPBW

1. Decompose signal s into spectrum $\mathbf{W} = \{\mathbf{d}_1, \mathbf{e}_2, \mathbf{f}_2, \ldots, \mathbf{e}_6, \mathbf{f}_6\}_l$ which consists of eleven levels.

2. Calculate the sum of power samples in all frequency subbands l according to rule

$$B_{l,k} = \sum_{n=(k-1)\cdot 2^{6-m}+1}^{k\cdot 2^{6-m}} \mathbf{w}_{n,l}^2, \qquad (15)$$

where k is a new discrete time index with the sampling period 4 ms, due to energy aggregation over the summation range and $\mathbf{w}_{n,l}$ are elements of $\mathbf{W}$.

3. Calculate the power envelopes as running mean values

$$B_{l,n}^{env} = \frac{1}{K} \sum_{k=n-\frac{K}{2}}^{n+\frac{K}{2}} B_{l,k}, \qquad (16)$$

where $K = 2^{-M}\Delta t_\mu f_s$ for expected mean duration Δt_μ of the speech segments. For the given $\Delta t_\mu = 0.1$ s, f_s=16 000 Hz and $M = 6$ we obtain $K = 25$ samples.

4. Generate importance matrix $\mathbf{L} = [L_{i,k}] \in \mathbb{R}^{11\times L_s}$ of frequency bands by sorting the envelopes in each time k position *i.e.*

$$\mathbf{L} = \left\{\{l_i\}_{i=1}^{11} : B_{l_1,n}^{env} \geq \cdots \geq B_{l_{11},n}^{env}\right\}_n \qquad (17)$$

where L_s depends on the length of the speech signal utterance.

5. Compute event-function

$$f(n) = \sum_{i=1}^{11} \frac{|L_{i,n+1} - L_{i,n}|}{i}. \qquad (18)$$

6. Segment border's locations can now be extracted from $f(n)$ by choosing its local maxima, which are greater than specified threshold f_{tr} and where each of them is the highest within the neighbourhood of Δt_n milliseconds.

Time-range condition rejects multiple changes related to the same border and segments shorter than Δt_n. Threshold adjusts sensitivity of the segmentation. By increasing its value we reduce the number of chosen events. It is reasonable to set its value on-line, according to the varying values of detection function

$$f_{tr}(n) = \frac{\alpha \cdot \sum_{k=-P}^{P} f(n-k)}{2P}, \qquad (19)$$

where P is an adaptation range corresponding to 100 milliseconds, and $\alpha \geq 0$ is a sensitivity factor.