# Wavelet Method of Speech Segmentation


Bartosz Ziółko, Suresh Manandhar, Richard C. Wilson and Mariusz Ziółko*

Department of Computer Science, The University of York

* Department of Electronics, AGH University of Science and Technology, Kraków



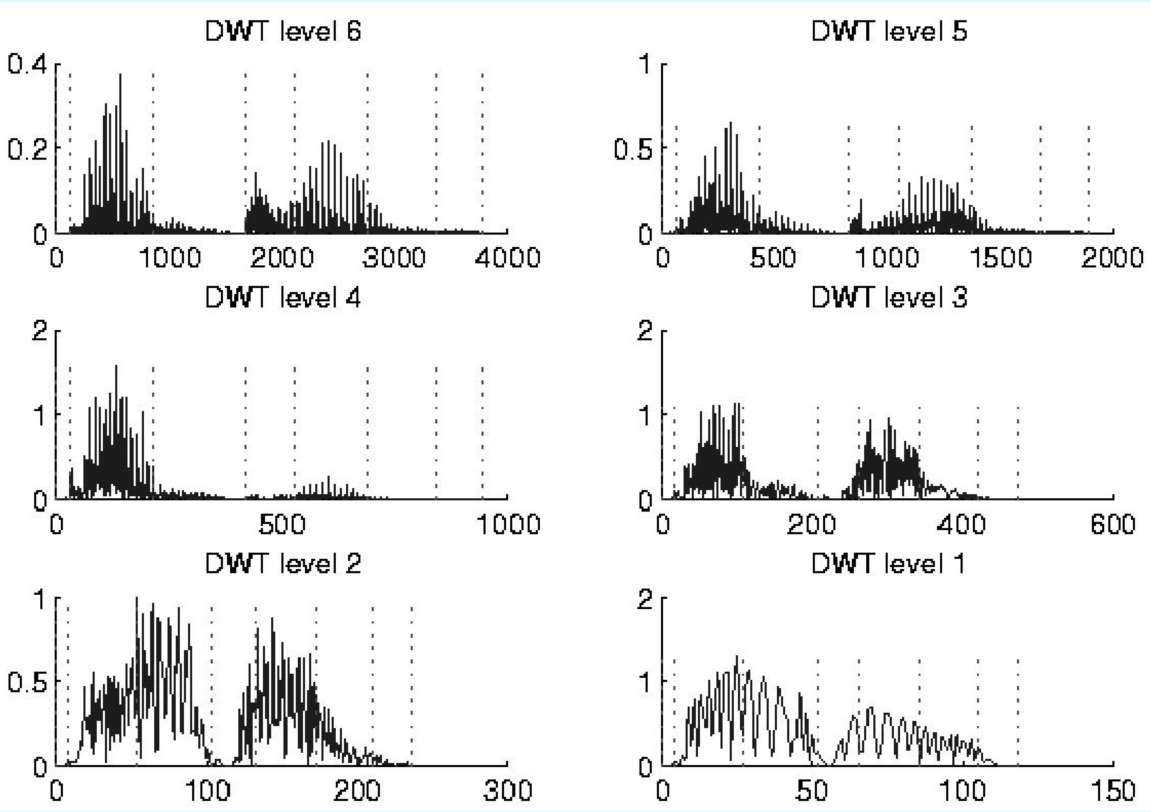
## Abstract

In this paper a new method of speech segmentation is suggested. It is based on power fluctuations of the wavelet spectrum of a speech signal. Boundaries are assigned in places where some energy of a frequency band rapidly changes. Most methods of non-constant segmentation need training for particular data or are realized as a part of modelling. In this paper we apply the DWT to analyse speech signals, the resulting power spectrum and its derivatives. This information allows us to locate the boundaries of phonemes. Additionally we present an evaluation by comparing our method with hand segmentation. The segmentation method proves effective for finding most phoneme boundaries.


## DWT subband power of speech

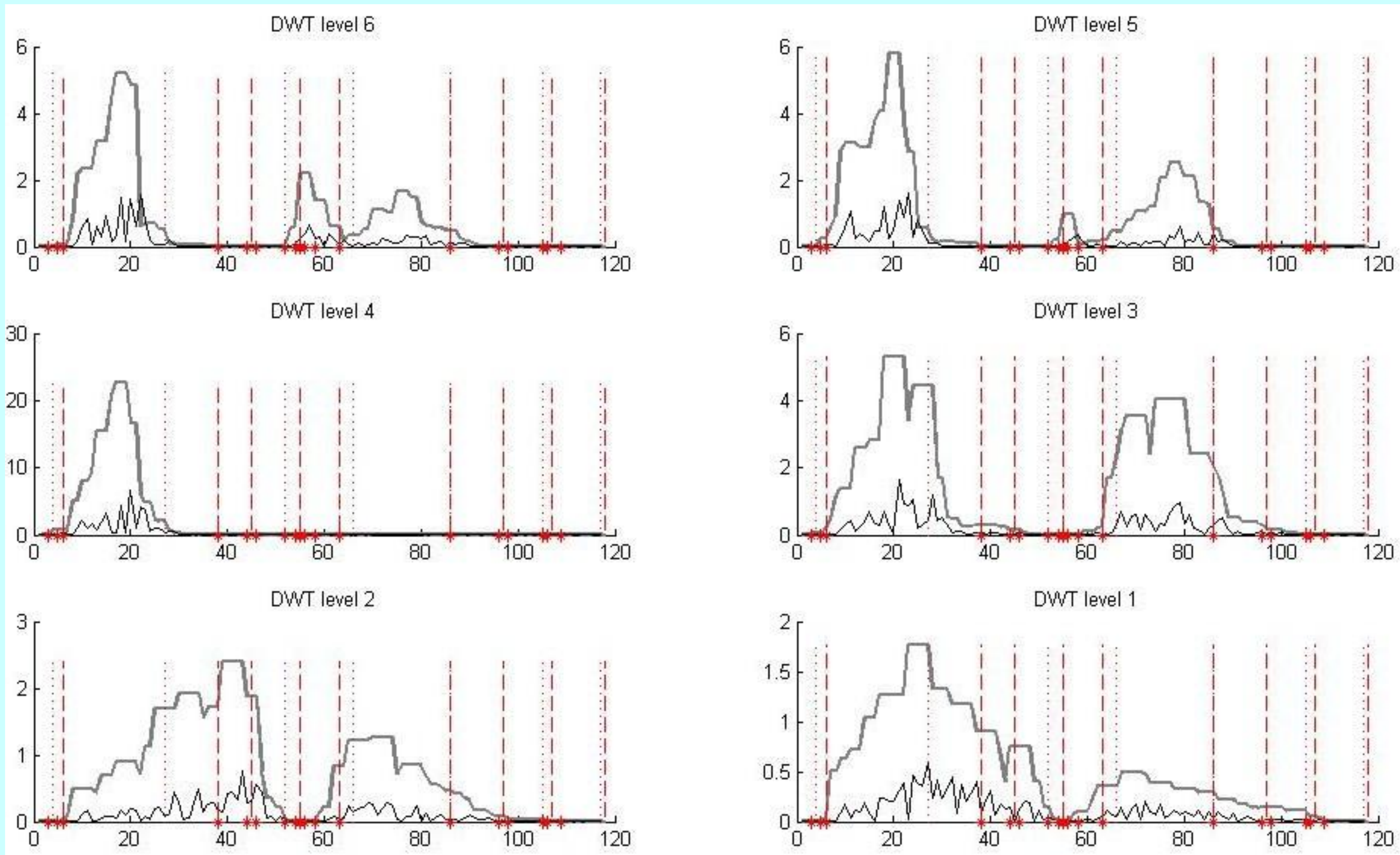
The DWT subband power shows rapid variations. The first order differences in the power are inevitably noisy, and so we calculate the envelopes $p'_n$ for power fluctuations in each subband by choosing the highest values of $p_n$ in a window of given size ω to obtain a power envelope. Additionally we use a smoothed differencing operator. The subband power $p_n$ is convolved with the mask [1,2,-2,-1] to obtain smoothed rate-of-change information $r_n(i)$.

$$p_n(i) = \sum_{j=1}^{2^{n-1}} d^2_{n,j+2^{n-1}i} \qquad \text{where } i = 0, \ldots, 2^{-M}N - 1,$$




## Phoneme segmentation

The start of a phoneme should be marked by an initially small but rapidly rising power level in one or more of the DWT levels. In other words, we should expect the power to be small and the derivative to be large. We can detect phoneme boundaries searching for $i$-points for which the inequality $p \geq |\beta|r_n(i)| - p'_n(i)|$ holds for the phoneme boundaries, where constant $p$ is a value of threshold which accounts for the time scale and sensitivity of the crossing points. Rate-of-change function $r_n$ is multiplied by scaling factor β approximately equal to 1. In practice we seek indexes for which the smoothed power and rate-of-change function approach close to each other and not necessarily cross them.




An example of the segmentation of a name 'Andrzej' /ʌːn dʒɛɪ/ . 6 DWT levels are presented. Dotted lines are hand segmentation boundaries; dashed lines are automatic segmentation boundaries, bold lines are power envelopes and thin grey lines are absolutes of power envelope first derivative. Asterisks are indexes with fulfilled condition for boundary candidate (see the algorithm).

| | Segment boundaries positions | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Auto | 0 | 6 | 38 | 45 | 55 | 63 | 86 | 97 | 107 | 118 |
| Hand | 0 | 4 | 27 | | 52 | 66 | 86 | | 105 | 118 |

## Phoneme detection algorithm

1. Normalise a speech signal by dividing by its maximum value.
2. Decompose a signal into six levels of the DWT.
3. Calculate the sum of power samples in all frequency sub-bands according to the table to obtain the power representations $p_n(i)$ of the $n$th subband.
4. Calculate the envelopes $p'_n$ for power fluctuations in each subband by choosing the highest values of $p_n$ in a window of a given size ω.
5. Calculate the rate-of-change function $r_n(i)$ by filtering $p_n(i)$ with [1,2,-2,-1] mask.
6. Given a threshold $p$ of the distance between $r_n(i)$ and $p'_n(i)$ and a threshold $p_{min}$ of minimal $p'_n$, find indexes for which $|\beta|r_n(i)|- p'_n(i)| < p$ AND $(|\beta|r_n(i+1)|- p'_n(i+1)| > p$ OR $|\beta|r_n(i-1)|-p'_n(i-1)| > p)$ AND $p'_n(i) > p_{min}$, where $\beta = 1$. Write such indexes in one vector (marked as asterisks in the figure).
7. Find and group indexes where there is no space between neighbouring ones longer than attribute α.
8. Calculate an average index value (rounded to the nearest integer) for each group found in the previous step as the representative of a group. They are indexes of phonemes' boundaries in indexing of DWT level 1.

| Method | av. $\varepsilon_n$ | av. $\varepsilon_p$ | Overall error |
|---|---|---|---|
| Const 23.2 ms | 2.9018 | 5.6380 | 20.1472 |
| Const 92.8 ms | 0.0796 | 5.2479 | 5.6459 |
| Meyer | 0.1602 | 3.2325 | 4.0334 |
| db2 | 0.2325 | 2.8531 | 4.0157 |
| db6 | 0.1927 | 3.0752 | 4.0385 |
| db20 | 0.1716 | 3.2724 | 4.1305 |
| sym6 | 0.1816 | 3.0581 | 3.9660 |
| haar | 0.2663 | 2.8783 | 4.2099 |

| DWT Level | Frequency band (Hz) | Number of samples in compare with level 1 | Window size ω |
|---|---|---|---|
| 6 | 2756–5512 | 32 | 3 |
| 5 | 1378–2756 | 16 | 3 |
| 4 | 689–1378 | 8 | 3 |
| 3 | 345–689 | 4 | 5 |
| 2 | 172–345 | 2 | 5 |
| 1 | 86–172 | 1 | 5 |

## Fuzzy Recall and Precision for speech segmentation evaluation

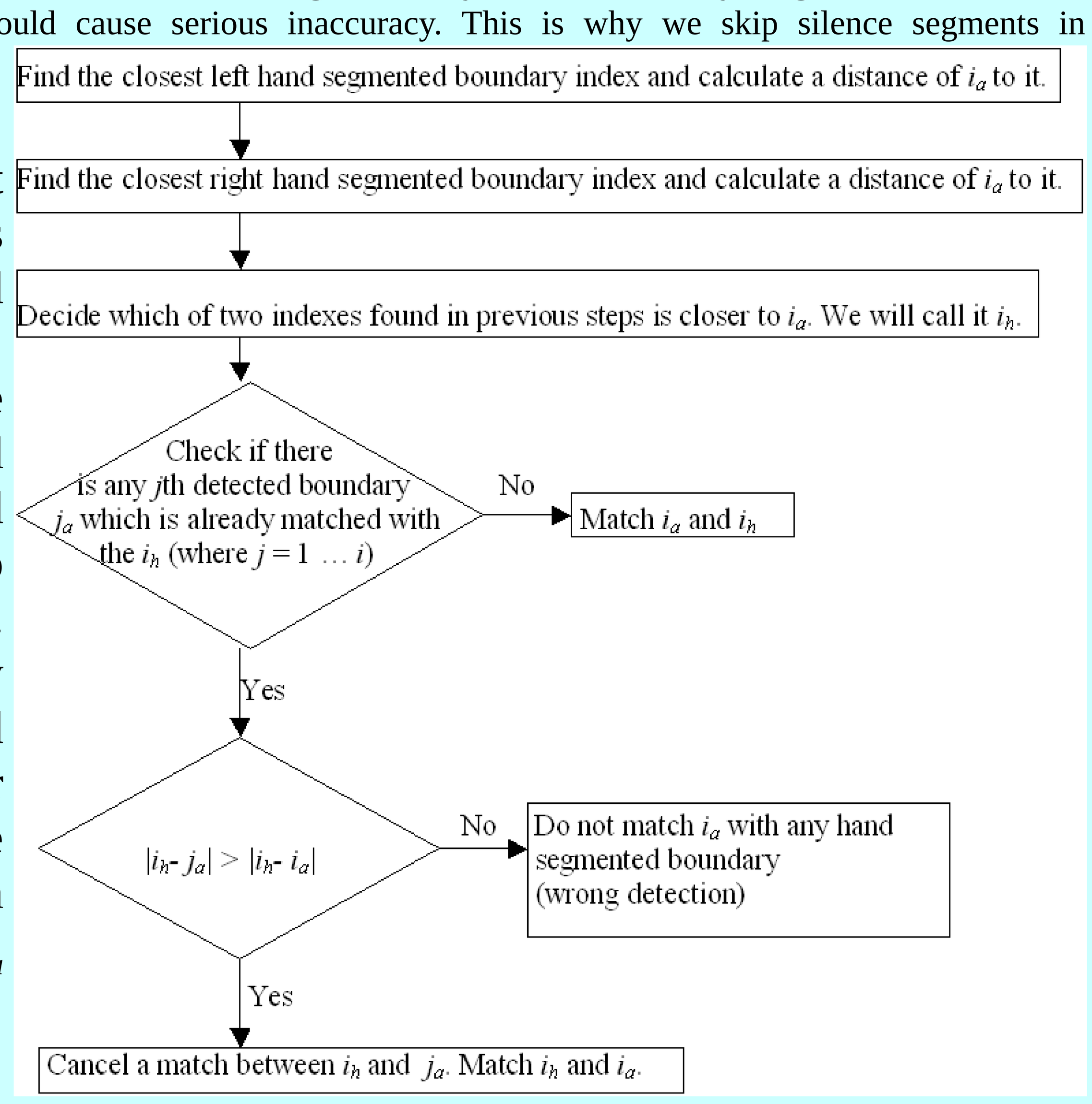
(developed after submitting the paper)

Assumptions:

- Hand segmentation is presented as a set of narrow ranges. Neighbouring phonemes overlap each other in these ranges. Detected boundaries are represented as a set of single indexes.
- We assume the perfect detection of silence. Silence segments may be of almost any length. Due to this fact including them in evaluation would cause serious inaccuracy. This is why we skip silence segments in evaluation.

The algorithm:

- Assign first and last detected boundaries as hand segmented boundaries.
- Start with matching the closest detected and hand segmented boundaries. We need to match them in pairs. Each boundary may have only one matched boundary from the other set. Do steps as in the diagram for each $i$th detected boundary $i_a$ starting from 1.




3. Calculate grades of being relevant and retrieved. All matched pairs, and all non-matched detected and hand segmented boundaries are elements of two fuzzy sets. One of them is the set of relevant elements. The other is the set containing retrieved boundaries. As the sets are fuzzy ones each element has two probability factors $A$ (representing being relevant information) and $B$ (representing being retrieved information).
   - Each hand segmented boundary not matched with any detected boundary has values $A=1$ and $B=0$.
   - Each detected boundary not matched with any hand segmented boundary has values $A=0$ and $B=1$.
   - There are two cases for matched pairs. If the detected boundary is inside the hand segmented boundary range the $A=1$ and $B=1$. Otherwise it is a fuzzy case and $A=B=a$-$b/a$ where $a$ stands for the half of the length of the phoneme which the boundary was detected (take the phoneme in which the detected boundary is situated) and $b$ stands for the distance between hand segmented boundary and the detected one.
4. The product $A \cup B$ of $A$ and $B$ has to be calculated according to fuzzy logic. Use an algebraic product $A(x) \cup B(x)$ for each element $x$.
5. Precision = $A \cup B / \sum(B)$.
6. Recall = $A \cup B / \sum(A)$.