Automatic disordered sound repetition recognition in continuous speech using CWT and kohonen network

Ireneusz Codello*, Wiesława Kuniszyk-Jóźkowiak†, Elżbieta Smołka‡, Adam Kobus§

1Institute of Computer Science, Maria Curie-Skłodowska University
pl. M. Curie-Skłodowskiej 1, 20-036 Lublin, Poland

Abstract – Automatic disorders recognition in speech can be very helpful for a therapist while monitoring therapy progress of patients with disordered speech. This article is focused on sound repetitions. The signal is analyzed using Continuous Wavelet Transform with 16 bark scales. Using the silence finding algorithm, only speech fragments are automatically found and cut. Each cut fragment is converted into a fixed-length vector and passed into the Kohonen network. Finally, the Kohonen winning neuron result is put on the 3-layer perceptron. Most of the analysis was performed and the results were obtained using the authors’ program WaveBlaster. We use the STATISTICA package for finding the best perceptron which was then imported back into WaveBlaster and used for automatic blockades finding. The problem presented in this article is a part of our research work aimed at creating an automatic disordered speech recognition system.

1 Introduction

Speech recognition is a highly important branch of computer science nowadays – oral communication with a computer can be helpful in real-time document writing, language translation or simply in using a computer. Therefore the issue has been analyzed for many years by the researchers which resulted in creating many algorithms, such as the Fourier transform, Linear Prediction, spectral analysis. Disorders recognition in speech is quite a similar issue – one attempt to find where speech is not fluent instead of trying

*irek.codello@gmail.com
†jozkowiak@gmail.com
‡esmolka@tytan.umcs.lublin.pl
§kobus.adam@gmail.com
to understand the speech, therefore the same algorithms can be used. Automatically generated statistics of disorders can be used as a support for therapists in their attempts at estimating therapy progress.

Several methods for the disordered speech detection have been used by researchers for disordered speech recognition, like: Fourier Transform, third octave filters, fuzzy logic \cite{1}, Hidden Markov Models, MFCC coefficients \cite{2}, Linear Prediction \cite{3} or Kohonen networks \cite{4}. In this paper a relatively new algorithm is used – Continuous Wavelet Transform (CWT) \cite{5, 6, 7} as - by using it - the most suitable scales (frequencies) can be chosen. Fourier transform and Linear Prediction \cite{8} are not so flexible – we have to choose if we want to have more precise time scale (small window) and more precise frequencies or the opposite - for the whole spectrogram. In CWT such a decision can be made for each scale separately. The bark scales set was taken, which is, besides the Mel scales and the ERB scales, considered as a perceptually based approach\cite{9}. Using only the speech finding algorithm, the utterance fragments were found and cut (automatically). Each cut fragment was converted into the fixed-length window (which contained several vectors eq. 5) and passed into the Kohonen network which received the 3D data and produced the 2D data (see Fig. 3). Such a dimensionally reduced signal was passed to a 3-layer perceptron with a mark: containing a blockade or not.

Perceptron learning was performed by the STATISTICA’s 'Neural Network’ package and its tool – Intelligent Problem Solver. Once found, the best network was imported back again into WaveBlaster and then it was used for the automatic disorders finding. Two–result statistics were presented: learning statistics of the best perceptrons from the STATISTICA package and recognition statistics obtained by WaveBlaster using these perceptrons.

\section{Input signal processing by CWT}

\subsection{Mother wavelet}

Mother wavelet is the heart of the Continuous Wavelet Transform:

\begin{equation}
CWT_{a,b} = \sum_{t} x(t) \cdot \psi_{a,b}(t), \quad \text{where} \quad \psi_{a,b}(t) = \frac{1}{\sqrt{a}} \psi \left( \frac{t - b}{a} \right)
\end{equation}

where $x(t)$ – input signal, $\psi_{a,b}(t)$ – wavelet family, $\psi(t)$ – mother wavelet, $a$ – scale (multiplicity of mother wavelet), $b$ – offset in time. The Morlet wavelet represented by equation (2) was used \cite{10}:

\begin{equation}
\psi(t) = e^{-t^2/2} \cdot \cos(2\pi \cdot 20 \cdot t)
\end{equation}

which has the center frequency $F_C = 20$Hz. Mother wavelets have one significant feature: length of the wavelet is connected with $F_C$ which is a restraint. The Morlet wavelet is different because the length can be chosen and then its $F_C$ can be set by changing the cosines argument.
2.2 Scales

For frequencies of scales, a perceptually based approach was assumed – because it is considered to be the closest to the human way of hearing. The Hartmut scales were chosen [11]:

\[
B = \frac{26.81}{1 + 1960/f} - 0.53, \quad f \text{ – freq. in Hz.}
\] (3)

The frequency \( F_a \) of each wavelet scale \( a \) was computed from the equation

\[
F_a = F_C F_S / a, \quad F_S \text{ – sampling frequency.}
\] (4)

Due to the discrete nature of the algorithm, it was not always possible to match scale \( a \) with scale \( B \) perfectly (Table 1). During the research some Hartmut scales were found as insignificant in the recognition process. Therefore eventually only 16 scales were used.

Table 1. 16 scales \( a \) with the corresponding frequencies \( f \) and the bark scales \( B \).

<table>
<thead>
<tr>
<th>( a ) [scale]</th>
<th>( f ) [Hz]</th>
<th>( B ) [bark]</th>
<th>( a ) [scale]</th>
<th>( f ) [Hz]</th>
<th>( B ) [bark]</th>
</tr>
</thead>
<tbody>
<tr>
<td>57</td>
<td>7736</td>
<td>20.9</td>
<td>220</td>
<td>2004</td>
<td>13</td>
</tr>
<tr>
<td>68</td>
<td>6485</td>
<td>20.1</td>
<td>256</td>
<td>1722</td>
<td>12</td>
</tr>
<tr>
<td>83</td>
<td>5313</td>
<td>19.1</td>
<td>297</td>
<td>1484</td>
<td>11</td>
</tr>
<tr>
<td>100</td>
<td>4410</td>
<td>18</td>
<td>347</td>
<td>1270</td>
<td>10</td>
</tr>
<tr>
<td>119</td>
<td>3705</td>
<td>17</td>
<td>408</td>
<td>1080</td>
<td>9</td>
</tr>
<tr>
<td>140</td>
<td>3150</td>
<td>16</td>
<td>479</td>
<td>920</td>
<td>8</td>
</tr>
<tr>
<td>163</td>
<td>2705</td>
<td>15</td>
<td>572</td>
<td>770</td>
<td>7</td>
</tr>
<tr>
<td>190</td>
<td>2321</td>
<td>14</td>
<td>700</td>
<td>630</td>
<td>6</td>
</tr>
</tbody>
</table>

2.3 Smoothing scales

Because the CWT values are similarity coefficients between the signal and wavelet, their sign are therefore irrelevant, in all computations, the following modules are taken – \( |CWT_{a,b}| \). We went one step further and the \( |CWT_{a,b}| \) was smoothed by creating a contour (see Fig. 1) because of its good recognition ratio influence [12].

Fig. 1. Left: Cross-section of one \( CWT_{a,b} \) scale. Right: Cross-section of one \( |CWT_{a,b}| \) scale and its contour (smoothed version).
2.4 Windowing

Thus the spectrogram consists of 16 smoothed bark scales vectors. Then the spectrogram was cut into 23.2ms frames (512 samples when \( F_S = 22050 \text{Hz} \)), with a 100\% frame offset. Because each scale has its own offset – one window of fixed width (e.g. 512 samples) will contain a different number of CWT values (CWT similarity coefficients) in each scale (see Fig. 3), therefore the CWT arithmetic mean of each scale value was taken.

\[ \bar{V} = \left\{ \text{mean} (|\text{CWT}_{57,i}|), \text{mean} (|\text{CWT}_{68,i}|), \ldots, \text{mean} (|\text{CWT}_{572,i}|), \text{mean} (|\text{CWT}_{700,i}|) \right\} \]

3 Modified kohonen network algorithm

The Kohonen network ([13, 14, 15, 16, 17, 18]) (or "self-organizing map" or SOM, for short) was developed by Teuvo Kohonen. The basic idea behind the Kohonen network is to establish a structure of interconnected processing units ("neurons") which compete for the signal. While the structure of the map may be quite arbitrary, rectangular maps were used in the research.

Let us assume that:

- Kohonen network has \( K \) neurons
- \( n \) is the dimension of each input vector \( X \)
- each element \( x_i \in X \) is connected to all \( K \) neurons, so we have \( K \times n \) connections. Each connection is represented by its weight \( w_{ij}, i = 1, \ldots, n, j = 1, \ldots, K \) which is adjusted during the training.

The Kohonen neurons were numbered by rows from the top to the bottom

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
</tr>
</tbody>
</table>
For every 2D CWT vector (see eq. (5)) one winning neuron is obtained. Therefore the Kohonen network is used to convert the 3–dimension CWT spectrogram (which consists of 2D CWT vectors situated one next to another) into the 2–dimension winning neuron contour as depicted in Fig. 3 ([4, 19]). Such reduction of data, from 3D into 2D, which is later passed on to MLP, occurred to have a positive impact on the non-fluencies recognition ratio ([4, 19]) (the whole 3D spectrogram seems to be too large for MLP to find general features).

The standard training algorithm [16, 18] was used with one modification – i.e. 0th neuron clearing [20].

4 Automatic disordered sound repetitions recognition

4.1 Input data

The Polish speech recordings of 9 stuttering persons were taken of the summary length equal to 9 min 44 s divided into 3 files: allblknn1, allblknn2, allblknn3 containing 294 disordered repetitions of the sounds: b,d,g,k,n,o,p,t. The statistics were the following:

Table 2. Disordered sound repetition fragments counts.

<table>
<thead>
<tr>
<th>file</th>
<th>b</th>
<th>d</th>
<th>g</th>
<th>k</th>
<th>n</th>
<th>o</th>
<th>p</th>
<th>r</th>
<th>sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>allblknn1</td>
<td>3</td>
<td>3</td>
<td>33</td>
<td></td>
<td></td>
<td>20</td>
<td>11</td>
<td></td>
<td>70</td>
</tr>
<tr>
<td>allblknn2</td>
<td>1</td>
<td>3</td>
<td>15</td>
<td></td>
<td></td>
<td>17</td>
<td>59</td>
<td></td>
<td>95</td>
</tr>
<tr>
<td>allblknn3</td>
<td>19</td>
<td>3</td>
<td>7</td>
<td>34</td>
<td>2</td>
<td>1</td>
<td>29</td>
<td>34</td>
<td>129</td>
</tr>
<tr>
<td>all</td>
<td>23</td>
<td>6</td>
<td>10</td>
<td>82</td>
<td>2</td>
<td>1</td>
<td>66</td>
<td>104</td>
<td>294</td>
</tr>
</tbody>
</table>

4.2 Automatic blockades cutting

Input files were automatically divided into words by a simple algorithm. We divided the CWT scalogram into 22ms windows with 11ms offset. Each window was marked as speech if it contained at least one value above the threshold: -53dB, -54dB or
Automatic disordered sound repetition recognition in...

-55dB (maximum CWT value was assigned as 0dB). Because we were looking only for disordered blockades, which are always short, words longer than 200ms were removed. Moreover, we observed that cutting algorithm was so sensitive, that it found silence in fluent words and divided them into pieces. Therefore we added the second parameter – distance: 50ms, 40ms, 30ms, 0ms. If two words were closer than the distance, then they were treated as one longer word and removed. Based on these two parameters, we created blockades cutting statistics containing a number of correctly cut blockades and a number of fluent words:

Table 3. Blockades cutting statistics (number of words) for the threshold and distance parameters.

<table>
<thead>
<tr>
<th>file</th>
<th>blk 55dB 50ms</th>
<th>fluent 55dB 50ms</th>
<th>blk 54dB 50ms</th>
<th>fluent 54dB 50ms</th>
<th>blk 53dB 50ms</th>
<th>fluent 53dB 50ms</th>
</tr>
</thead>
<tbody>
<tr>
<td>allbkn1</td>
<td>50 71%</td>
<td>95 66 94%</td>
<td>123 65 93%</td>
<td>139</td>
<td></td>
<td></td>
</tr>
<tr>
<td>allbkn2</td>
<td>74 78%</td>
<td>92 83 87%</td>
<td>129 82 86%</td>
<td>139</td>
<td></td>
<td></td>
</tr>
<tr>
<td>allbkn3</td>
<td>98 76%</td>
<td>58 121 94%</td>
<td>132 116 90%</td>
<td>139</td>
<td></td>
<td></td>
</tr>
<tr>
<td>all</td>
<td>222 76%</td>
<td>245 270 92%</td>
<td>384 263 89%</td>
<td>417</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>file</th>
<th>blk 55dB 40ms</th>
<th>fluent 55dB 40ms</th>
<th>blk 54dB 40ms</th>
<th>fluent 54dB 40ms</th>
<th>blk 53dB 40ms</th>
<th>fluent 53dB 40ms</th>
</tr>
</thead>
<tbody>
<tr>
<td>allbkn1</td>
<td>56 80%</td>
<td>134 68 97%</td>
<td>144 67 96%</td>
<td>162</td>
<td></td>
<td></td>
</tr>
<tr>
<td>allbkn2</td>
<td>77 81%</td>
<td>126 84 88%</td>
<td>146 85 89%</td>
<td>154</td>
<td></td>
<td></td>
</tr>
<tr>
<td>allbkn3</td>
<td>104 81%</td>
<td>77 121 94%</td>
<td>146 116 90%</td>
<td>153</td>
<td></td>
<td></td>
</tr>
<tr>
<td>all</td>
<td>237 81%</td>
<td>337 273 93%</td>
<td>436 268 91%</td>
<td>469</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>file</th>
<th>blk 55dB 30ms</th>
<th>fluent 55dB 30ms</th>
<th>blk 54dB 30ms</th>
<th>fluent 54dB 30ms</th>
<th>blk 53dB 30ms</th>
<th>fluent 53dB 30ms</th>
</tr>
</thead>
<tbody>
<tr>
<td>allbkn1</td>
<td>56 80%</td>
<td>134 70 100%</td>
<td>165 69 99%</td>
<td>189</td>
<td></td>
<td></td>
</tr>
<tr>
<td>allbkn2</td>
<td>77 81%</td>
<td>126 91 96%</td>
<td>168 89 94%</td>
<td>178</td>
<td></td>
<td></td>
</tr>
<tr>
<td>allbkn3</td>
<td>104 81%</td>
<td>77 122 95%</td>
<td>161 117 91%</td>
<td>174</td>
<td></td>
<td></td>
</tr>
<tr>
<td>all</td>
<td>237 81%</td>
<td>337 283 96%</td>
<td>494 275 94%</td>
<td>541</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>file</th>
<th>blk 55dB 0ms</th>
<th>fluent 55dB 0ms</th>
<th>blk 54dB 0ms</th>
<th>fluent 54dB 0ms</th>
<th>blk 53dB 0ms</th>
<th>fluent 53dB 0ms</th>
</tr>
</thead>
<tbody>
<tr>
<td>allbkn1</td>
<td>59 84%</td>
<td>170 70 100%</td>
<td>199 70 100%</td>
<td>226</td>
<td></td>
<td></td>
</tr>
<tr>
<td>allbkn2</td>
<td>87 92%</td>
<td>188 93 98%</td>
<td>197 91 96%</td>
<td>209</td>
<td></td>
<td></td>
</tr>
<tr>
<td>allbkn3</td>
<td>98 76%</td>
<td>111 122 95%</td>
<td>188 118 91%</td>
<td>199</td>
<td></td>
<td></td>
</tr>
<tr>
<td>all</td>
<td>244 83%</td>
<td>449 285 97%</td>
<td>584 279 95%</td>
<td>634</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Based on these statistics we decided to get only the configurations: 50ms-55dB, 50ms-54dB, 30ms-54dB.

4.3 Training algorithm

The procedure of finding sound repetitions in the file was the following:
1. The CWT spectrogram of the continuous speech was computed.
2. The CWT signal was divided into 22ms windows with 50% offset, and only the words that match criteria (see 4.2) were chosen. The distance and threshold parameters
were applied to the algorithm (see 4.2), and the most suitable ones were used: 50ms-55dB, 50ms-54dB, 30ms-54dB.

3. If the speech fragment passed the above verification, it was cut with a surrounding according to the **window length** parameters: 700ms, 1000ms, 1500ms, 2000ms, 2500ms, 3000ms (each window always contained the 500ms prefix, speech fragment and the postfix of variable length so that we would obtain a desired window length).

4. Each window which consisted of 16-element vectors was automatically passed into the Kohonen network. After the training process a winning neuron graph was obtained (Fig 3). The 5x5 Kohonen network was used with the following parameters: 100 epochs, learning coefficient 0.20-0.10, and neighbour distance 2.5-0.5.

5. Each graph was marked as fluent or non-fluent (this information was ‘the teacher’ in the perceptron learning algorithm).

6. Using STATISCTICA, the perceptron with the best recognition ratio was found. The input vectors were divided randomly by STATISTICA into teaching set (50%), verifying set (25%) and testing set (25%). (Only the allblk2n2 and allblkn3 files were passed to the STATISTICA). The best perceptron (see Table 4) was imported back into WaveBlaster and all three files took part in the finding process.

4.4 Finding algorithm

1. Steps 1–4 were repeated from the previous paragraph.
2. The obtained Kohonen vector was passed into the perceptron (imported from STATISTICA) and its output was checked.
3. Based on the output the speech fragment was marked as fluent/non-fluent.

5 Results

The recognition ratio was calculated with the use of these formulas:

$$sensitivity = \frac{P}{A}; \quad predictability = \frac{P}{P+B}$$

where \(P\) is the number of correctly recognized disorders, \(A\) is the number of all disorders and \(B\) is the number of fluent sections mistakenly recognized as disorders.

6 Conclusions

As we can see in Table 4 all perceptrons distinguish blockades really well (97%-100%), even in verification and testing sets (test vectors do not take part in teaching at all). That is because of speech cutting algorithm – on the perceptron only speech fragments that begin with the utterance were passed, therefore the perceptron does not have to straggle with fragments that have sometimes blockade in the middle and sometimes at the end. Such results would suggest that this method of cutting blockades is very good.
Unfortunately our speech cutting algorithm has a weakness – it misses some of the blockades and by making it more sensitive, it cuts disproportionately more fluent fragments (see Table 3). Maybe a more complex and more smart algorithm should be used.

As for automatic blockades recognition results in the fluent speech (see Table 5), we need to remember that they can only be as good as speech cutting efficiency. Nets 1-6 work on the set that has only 71%-78% blockades cut (see Table 3 blk 55dB 50ms section) so their results are significantly lower than sets 7-12 having 87%-94% blockades cut (see Table 3 blk 54dB 50ms section) or sets 13-18 having 95%-100% blockades cut (see Table 3 blk 54dB 30ms section). Files allblkn2 and allblkn3 have very good results. Of course these files were used in teaching the perceptron but we should remember that only 50% of fragments took direct part in teaching (learning set) while 25% of the fragments were not used at all (testing set).

We tested one file that was not used in teaching at all – allblkn1. As we can see the results are significantly lower but still good. After closer investigation it occurred that the file has a few series blockades that occur very fast one after another (like “p p p publication”). Though the cutting algorithm cut them correctly, perceptron decided...
that they were so close to each other that it had to be one fluent word. Because such a decision was applied to all blockades in one series (not only one), this lowered the recognition ratio heavily.

The last conclusion is connected with the result for the file allblknn1. Nets 7-12 which received 71%-78% blockades had better results than those 13-18 which received 95%-100% blockades. This means that perceptron cannot receive too many fluent fragments (nets 7-12 received 123 and nets 13-18 received 165) because it makes more mistakes though it has more blockade patterns to learn on.

### References


Automatic disordered sound repetition recognition in...