Signal-Based Accent and Phrase Marking Using the Fujisaki Model

Hussein Hussein and Rüdiger Hoffmann

Laboratory of Acoustics and Speech Communication, Dresden University of Technology, 01062 Dresden, Germany
{hussein.hussein, ruediger.hoffmann}@ias.et.tu-dresden.de

Abstract

Automatic prosodic marking is very important in speech signal processing, since its results are required in many subsections, e.g. speech synthesis and speech recognition. The most important prosodic features on the linguistic level are the marking of accents and phrases. In this paper, we develop an automatic algorithm for marking accents and phrases which analyzes the F0 contour by using the quantitative Fujisaki model. The results of automatic extraction of accents and phrases have been compared to the human labeling performance. The success rate of accent and phrase marking amounts to 77.11% and 67.12%, respectively.

Index Terms: prosody, Fujisaki model, prosodic marking

1. Introduction

Prosodic marking of speech corpora is very important in speech signal processing. Manual prosodic marking is time-consuming and not an easy task. Therefore, many automatic methods have been developed [1][2].

Accents and phrases are the most important prosodic features on the linguistic level. As a result, we developed an automatic algorithm for accent and phrase marking which analyzes the F0 contour by using the well-known Fujisaki model [3].

The Fujisaki model is a quantitative intonation model which is being utilized especially in speech synthesis for intonation analysis and intonation generation. The Fujisaki parameters are: F0: speaker-individual base frequency, I: number of phrase commands, J: number of accent commands, Ap: amplitude of the pth phrase command, Aa: amplitude of the jth accent command, T0: timing of the pth phrase command, T1: onset of the jth accent command, T2: offset of the jth accent command, α: natural angular frequency of the pth phrase command, β: natural angular frequency of the jth accent command, γ: relative ceiling level of the accent commands (generally set to γ = 0.9). In Section 2, we present the speech material. Section 3 introduces a detailed concept of the accent and phrase marking algorithm. Section 4 provides an overview of the evaluation criteria. Section 5 displays the evaluation results.

2. Speech Material

A large speech database was used in order to create a statistic of the Fujisaki parameters. The corpus is a subset of the multilingual Verbomobil database [4] (only the spontaneous utterances of German) and consists of 10 hours of speech signals from multiple speakers (male and female). The Verbomobil database contains boundary labels on the phone level. The speech signals of the Verbomobil database were provided with a sampling frequency of 16 kHz and a resolution of 16 bit.

An evaluation of the performance of the automatic accent and phrase marking algorithm is usually achieved by comparing the automatically marked accents and phrases with a reference of accents and phrases. A subset of the Verbomobil database was used as reference. It contains speech signals of 8 speakers (4 male + 4 female speakers, 10 speech signals per speaker). The reference contains manually marked accents and phrases. The first author (non-native speaker of German) listened to each of the speech signals several times and annotated the space between two words in the orthographic sentence with an accent or phrase symbol.

3. Proposed Method

This section describes the proposed algorithm for prosodic marking. The algorithm is based on the extraction of Fujisaki parameters. Figure 1 shows a block diagram of the proposed algorithm. The accent and phrase marking algorithm contains the following components:

- **Orthographic Text**
- **Speech Signal**
- **Phoneme Label**
- **Hybrid PMA**
- **PM**
- **Calculation of F0 Contour**
- **F0 Contour**
- **Extraction of Fujisaki Parameter**
- **PAC**
- **Accent Annotation**
- **Phrase Annotation**
- **Accents (Major, Minor)**
- **Phrases (Major, Minor)**

Figure 1: Framework of the accent and phrase marking algorithm

3.1. Extraction of Pitch Marks

The hybrid algorithm for pitch marking [5], which combines the outputs of two speech signal-based pitch marking algorithms using Finite State Machines (FSM), was used. The algorithm is based on the alignment of pitch marks to the nearest negative peak of the speech signal and on the selection of more accurate pitch marks that yield the highest confidence score.
### 3.2. Calculation of the $F_0$ Contour

The distance between pitch marks was calculated in samples. The pitch marks were sampled with the same sampling rate of the $F_0$ contour ($F_s F_0$). The values of the $F_0$ contour are calculated as follows:

$$F_{0i} = \frac{F_S}{\text{Length of PM}_i \text{ in samples}}$$

In this equation, $F_S$ is the sampling rate of the speech signal ($F_S F_0=100$ Hz for $F_S=16$ kHz).

### 3.3. Extraction of Fujisaki Parameters

The Fujisaki parameters were automatically extracted from the smoothed $F_0$ contour. The inverse Fujisaki model was used for analyzing the $F_0$ contour as well as for breaking it down into its accent and phrase commands. The Fujisaki parameters were saved in Phrase and Accent Command (PAC) files. We applied two methods for the extraction of Fujisaki parameters, each of which uses an individual algorithm for smoothing the $F_0$ contour.

1. **Mixdorff (2000):** The $F_0$ contour is interpolated and smoothed by employing the popular Momel method [6]. A high-pass filter is used to extract the high frequency contour (HFC), which contains the accent commands, from the smoothed $F_0$ contour. The HFC is subtracted from the smoothed contour, yielding a low frequency contour (LFC) from which the phrase commands are extracted. In this algorithm, parameters $\alpha_i$ and $\beta_i$ are constant (2 and 20, respectively) [7].

2. **Kruschke (2003):** The preprocessing algorithm that is described in [8] was used for smoothing the $F_0$ contour. The $F_0$ contour is stylized by piecewise polynomial approximation (see figure 2,b). $F_0$ is subtracted from the logarithmic $F_0$ contour, resulting in $F_{0,\text{rest}}1(t)$. A continuous wavelet transform (CWT) using a Mexican hat wavelet is applied to the $F_{0,\text{rest}}1(t)$. This way, the accent commands are detected and optimized. A new $F_0$ contour is generated from accent commands and subtracted from the $F_{0,\text{rest}}1(t)$ contour. The resulting contour is $F_{0,\text{rest}}2(t)$. Again, the CWT is applied to the $F_{0,\text{rest}}2(t)$ for detecting phrase commands. All phrase commands are detected and optimized. Finally, the parameters of all phrase and accent commands are optimized together [9]. The Fujisaki parameters of this method are shown in figure (2.e).

### 3.4. Accent and Phrase Annotation

Two levels of accents and phrases (major and minor) as in the EU project "Technology and Corpora for Speech to Speech Translation" (TC-STAR) [10] were automatically marked by using the Fujisaki model. The marking of accent and phrase is based on the word level. Therefore, the word boundaries were calculated from the phoneme boundaries. We analysed the Fujisaki parameters which were extracted with [9] from 10 hours of speech signals (see figures 3 and 4). The mean and standard deviations (SD) of the word duration were calculated at 0.299 sec and 0.218 sec, respectively.

#### 3.4.1. Automatic Extraction of Prosodic Accents

The syllable carries the accent of the word [11]. The following prosodic accent labels were defined: Major accent $<a>$ (emphatic) and Minor accent $<a>$ (normal).
3.4.2. Automatic Extraction of Prosodic Phrases

Two types of phrase labels were distinguished: Major breaks $<\text{sS}>$ (full intonational phrase) and Minor breaks $<\text{s}>$ (intermediate intonational phrase).

The following steps were taken in phrase marking:

- Phrase commands at the beginning of an utterance ($t < 0$) were marked at the beginning of a sentence.
- Phrase commands at the end of an utterance were not marked.
- Phrase commands within a pause led to a phrase that began with the following word.
- If phrase commands occurred within a word and if there was a pause before or after this word, a phrase was marked as starting at the beginning of the word which came after the pause.
- Phrase commands within a word led to a phrase that started at the beginning of the next word if it was not preceded or followed by a pause. These were marked as minor phrases.
- Phrases that feature very small amplitudes ($Ap < 0.01$) were deleted.
- Phrase commands with large amplitudes ($Ap > 0.5$) and large variables ($\alpha > 2$) or phrase commands with long pauses ($t > 700\text{ms}$) were marked as major phrases [14].

Figure (2,d) shows the modified accent and phrase commands after accent and phrase marking. The resulting sentence is: “$<\text{s}>$ Sonntag $<\text{a}>$ der einunddreißigste $<\text{a}>$ Juli $<\text{a}>$ doch das ginge $<\text{a}>$ bei mir auch ich denke $<\text{a}>$ $<\text{s}>$ da ließe sich s$. ... ich denke, da ließe sich was machen.” (“Sunday, the thirty-first of July. Certainly, this would work for me as well. I think we can work something out.”)
4.1.1. Success rate (SR\%)  
Correct prosodic marks take the same positions as the according reference prosodic marks. The success rate (SR\%) of the prosodic marking algorithm is defined as follows:

$$SR_\% = \frac{|\{x | x \in Text \land x \in ReRef\}|}{|ReRef|} \times 100\% \quad (2)$$

where $|ReRef|$ represents the set of all manually corrected prosodic marks (reference) and $|Text|$ represents the set of automatically generated prosodic marks.

4.1.2. Accuracy (ACC\%)  
To consider potentially substituted, deleted and inserted prosodic marks, the accuracy of the prosodic marking algorithm is calculated as follows:

$$ACC_\% = \frac{|ReRef| - |Sub| - |Del| - |Ins|}{|ReRef|} \times 100\% \quad (3)$$

where $|Sub|$ is the number of substituted prosodic marks, $|Del|$ is the number of deleted prosodic marks and $|Ins|$ is the number of inserted prosodic marks.

5. Results

The results are calculated for both extractors of Fujisaki parameters (Mixdorff and Kruschke). Results which are based on the Kruschke algorithm are better than those of the Mixdorff algorithm. Table 1 shows the results of accent marking. The results of minor accents are better than those of major accents, but still marginal. Therefore, we combined the minor and major accents on one level (minor and major accent). Success rate and accuracy of the combined minor and major accents (SR=77.11% and ACC=71.46\%) are better than both the results of major and minor accents individually. The results of phrase marking are presented in table 2. The accuracy of major phrases is better than that of minor phrases, but the success rate is lower. The combination of minor and major phrases on one level of a phrase enhances the success rate (SR=67.12\%).

The results are better than both the results of major and minor accents individually. The success rate of minor and major phrase marking is better than minor phrase and major phrase marking individually. The parameters in the algorithm were optimised to yield the best results for both success rate and accuracy. Results indicate that manual correction of automatic labeling output is still necessary to achieve a high quality speech database.

6. Conclusion

An algorithm for accent and phrase marking using the Fujisaki-model has been presented. It is based on analyzing the $F_0$ contour as well as on breaking it down into its components (accent and phrase commands). To evaluate the algorithmic performance, we used a manually marked subset of the Verbmobil database as a reference for accents and phrases. Experimental results of accent marking indicate that the combination of minor and major accents on one level (minor and major accents) yields better results than both major and minor accents individually. The success rate of minor and major phrase marking is better than minor phrase and major phrase marking individually. The results in the algorithm were optimised to yield the best results for both success rate and accuracy. Results indicate that manual correction of automatic labeling output is still necessary to achieve a high quality speech database.

7. References


<table>
<thead>
<tr>
<th>Accent Level</th>
<th>Method</th>
<th>Test</th>
<th>Correct</th>
<th>Sub</th>
<th>Del</th>
<th>Ins</th>
<th>SM</th>
<th>ACC%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minor Accent (&lt;&lt;)</td>
<td>Mixdorff</td>
<td>120</td>
<td>594</td>
<td>255</td>
<td>57</td>
<td>108</td>
<td>130</td>
<td>60.71</td>
</tr>
<tr>
<td></td>
<td>Kruschke</td>
<td>128</td>
<td>502</td>
<td>276</td>
<td>44</td>
<td>120</td>
<td>213</td>
<td>65.71</td>
</tr>
<tr>
<td>Major Accent (&gt;&gt;)</td>
<td>Mixdorff</td>
<td>17</td>
<td>71</td>
<td>13</td>
<td>0</td>
<td>4</td>
<td>23.53</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Kruschke</td>
<td>17</td>
<td>71</td>
<td>13</td>
<td>0</td>
<td>4</td>
<td>23.53</td>
<td>0.00</td>
</tr>
<tr>
<td>Minor and Major Accent</td>
<td>Mixdorff</td>
<td>137</td>
<td>655</td>
<td>128</td>
<td>-</td>
<td>109</td>
<td>137</td>
<td>72.00</td>
</tr>
<tr>
<td></td>
<td>Kruschke</td>
<td>137</td>
<td>655</td>
<td>128</td>
<td>-</td>
<td>109</td>
<td>137</td>
<td>72.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Phrase Level</th>
<th>Method</th>
<th>Test</th>
<th>Correct</th>
<th>Sub</th>
<th>Del</th>
<th>Ins</th>
<th>SM</th>
<th>ACC%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minor Phrase (&lt;&lt;)</td>
<td>Mixdorff</td>
<td>209</td>
<td>247</td>
<td>120</td>
<td>7</td>
<td>82</td>
<td>108</td>
<td>57.42</td>
</tr>
<tr>
<td></td>
<td>Kruschke</td>
<td>208</td>
<td>226</td>
<td>135</td>
<td>3</td>
<td>71</td>
<td>98</td>
<td>64.59</td>
</tr>
<tr>
<td>Major Phrase (&gt;&gt;)</td>
<td>Mixdorff</td>
<td>13</td>
<td>14</td>
<td>7</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>53.85</td>
</tr>
<tr>
<td></td>
<td>Kruschke</td>
<td>13</td>
<td>14</td>
<td>7</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>53.85</td>
</tr>
<tr>
<td>Minor and Major Phrase</td>
<td>Mixdorff</td>
<td>222</td>
<td>247</td>
<td>149</td>
<td>-</td>
<td>73</td>
<td>98</td>
<td>67.12</td>
</tr>
<tr>
<td></td>
<td>Kruschke</td>
<td>222</td>
<td>247</td>
<td>149</td>
<td>-</td>
<td>73</td>
<td>98</td>
<td>67.12</td>
</tr>
</tbody>
</table>