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 F_0 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

10.21437/Interspeech.2010-368

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 F_0 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: F_b: speaker-individual base frequency, I: number of phrase commands, J: number of accent commands, Ap_i: amplitude of the ith phrase command, Aa_j: amplitude of the jth accent command, $T0_i$: timing of the ith phrase command, $T1_j$: onset of the jth accent command, $T2_j$: offset of the jth accent command, α_i: natural angular frequency of the ith phrase command, β_j: natural angular frequency of the jth accent command, γ: relative ceiling level of the accent commands (generally set to $\gamma = 0.9$).

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 Verbmobil database [4] (only the spontaneous utterances of German) and consists of 10 hours of speech signals from multiple speakers (male and female). The Verbmobil database contains boundary labels on the phone level. The speech signals of the Verbmobil 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 Verbmobil 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:

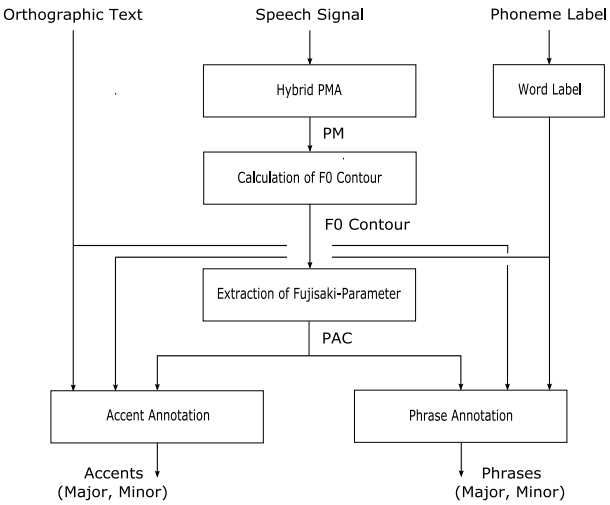

Figure 1: Framework of the accent and phrase marking algorithm

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:

$$F_{0i} = \frac{F_S}{\text{Length of PM}_j \text{ in samples}} \quad (1)$$

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

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 α_i and β_j 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_b is subtracted from the logarithmic F_0 contour, resulting in $F_0 rest1(t)$. A continuous wavelet transform (CWT) using a Mexican hat wavelet is applied to the $F_0 rest1(t)$. This way, the accent commands are detected and optimized. A new F_0 contour is generated from accent commands and subtracted from the $F_0 rest1(t)$ contour. The resulting contour is $F_0 rest2(t)$. Again, the CWT is applied to the $F_0 rest2(t)$ 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,c).

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 <aa> (emphatic) and Minor accent <a> (normal).

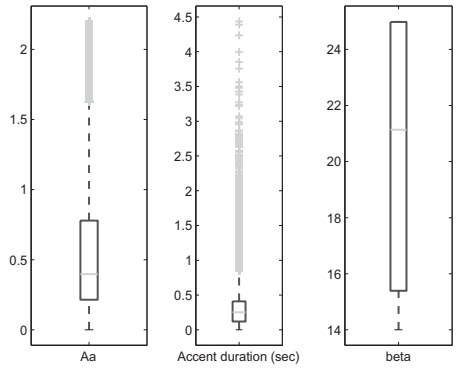

Figure 3: Boxplot of accent command amplitude, accent duration and accent parameter (β)

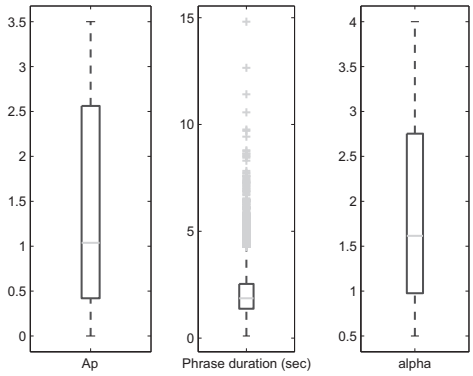

Figure 4: Boxplot of phrase command amplitude, phrase duration and phrase parameter (α)

The following steps were taken in accent marking:

- Several accent commands in a word were combined into an accent [12].
- The accent command which was extended to more than one word was divided between the words. Long words contain more syllables than short words. Therefore, the probability of an emphasised syllable in long words is higher than in short words. As a result, the word which contained the largest duration in an accent command was marked as accented. The new parameters $T1_j$` and $T2_j$` were modified to the start and end time of the accented word.
- The accent duration in an accented word was changed from 20% to 80% of the word duration. The best results were obtained by an accent duration of more than (60%) of the word duration and a word length greater than ($0.2sec.$).
- Accent commands which were fully located in the pauses were deleted.
- Long accent commands ($> 1sec.$), extremely small accent amplitudes ($Aa < 0.2$) and very long accent commands ($> 1.5sec.$) were ignored. These can be interpreted as additional phrase commands [13].
- Accent commands with very large amplitudes ($Aa > 0.8$) and large beta variables ($\beta > 15$) were marked as major accents.

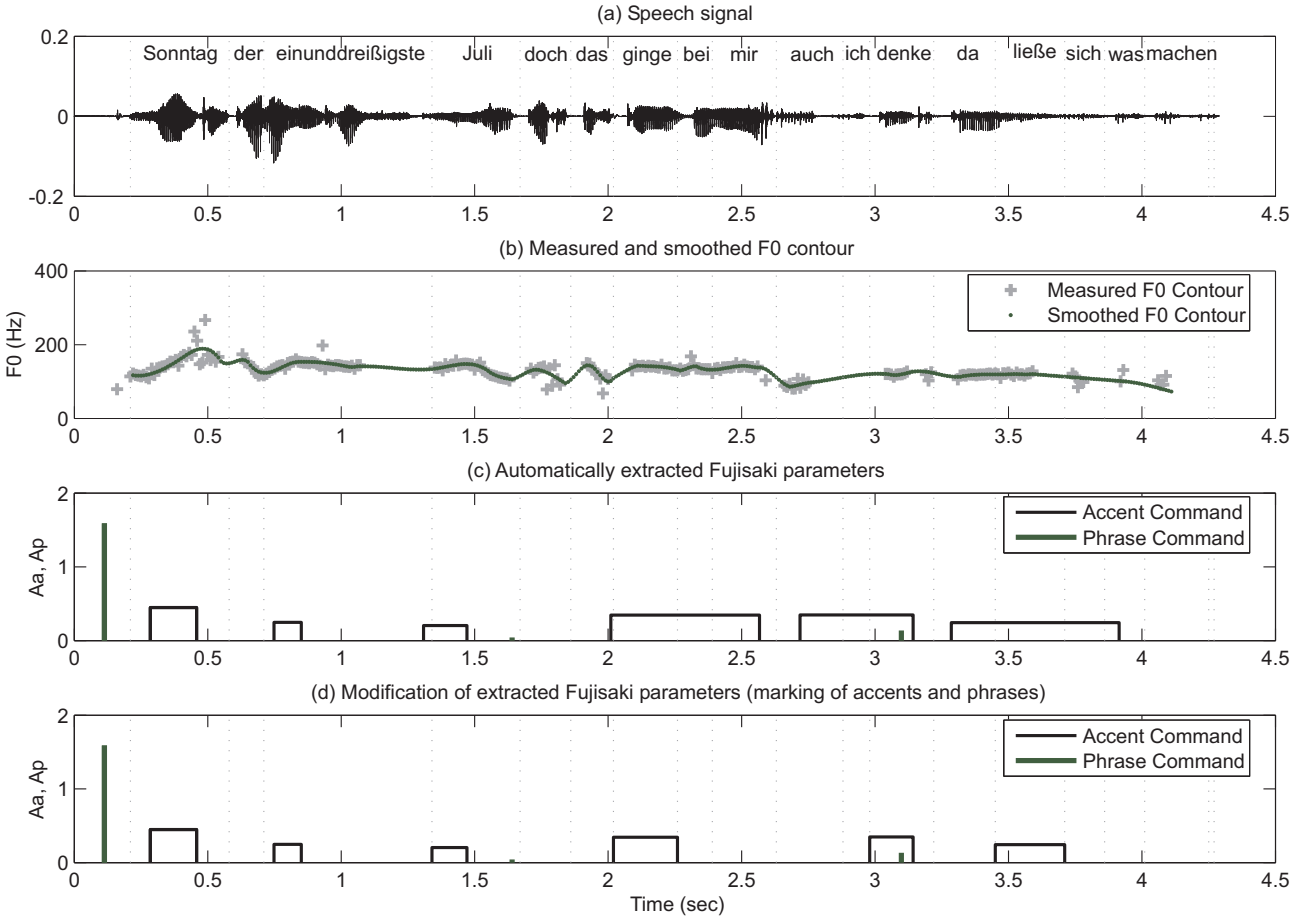

Figure 2: Speech signal, F0 contour, extracted and modified Fujisaki parameters. Sentence: "Sonntag, der einunddreißigste Juli. Doch, das ginge bei mir auch. Ich denke, da ließe sich was machen." ("Sunday, the thirty-first of July. Certainly, this would work for me as well. I think we can work something out.")

- Words that have an accent duration above ($0.4sec.$) were marked as major accented words.
- Short accent commands ($< 0.08sec.$) and large accent command amplitudes ($Aa >= 1$) were deleted.
- Accents with very small amplitudes ($Aa <= 0.01$) were ignored.

3.4.2. Automatic Extraction of Prosodic Phrases

Two types of phrase labels were distinguished: Major breaks <ss> (full intonational phrase) and Minor breaks <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 ($t >= 700ms$) 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: "<s> Sonntag <a> der einunddreißigste <a> Juli <a> <s> doch das ginge <a> bei mir auch ich denke <a> <s> da ließe <a> sich was machen".

4. Experiments

4.1. Evaluation Criteria

So far, texts that deal with prosodic marking are only considering the success rate, i.e. only the corrected prosodic marks, in order to assess the efficiency of an algorithm. There are, however, further important factors such as substitution, deletion and insertion that have to be taken into account to this end.

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:

$$SR_{\%} = \frac{|\{x|(x \in Test) \wedge (x \in Ref)\}|}{|Ref|}.100\% \quad (2)$$

where $|Ref|$ represents the set of all manually corrected prosodic marks (reference) and $|Test|$ 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:

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

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=27.46%) 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 authors optimised all parameters in the algorithm to achieve the best results for both success rate and accuracy. This means that we can receive better results for the success rate. In this case, however, the number of substituted, deleted and inserted prosodic marks is higher, whereas the accuracy is lower. The results are still not satisfying if compared to the state-of-the-art method for automatic marking of accents and phrases, which yields a SR of 82.6% for accents and of 88.3% for phrases in [2].

Table 1: Results of accent marking

Accent Level	Method	Ref	Test	Corr	Sub	Del	Ins	SR[%]	ACC[%]
Minor	Mixdorff	420	594	255	57	108	330	60.71	-17.86
Accent <a>	Kruschke	420	502	276	44	100	213	65.71	15.00
Major	Mixdorff	17	71	7	9	1	7	41.18	00.00
Accent <aa>	Kruschke	17	52	4	13	0	4	23.53	00.00
Minor and Major	Mixdorff	437	665	328	-	109	337	75.06	-02.06
Accent	Kruschke	437	554	337	-	100	217	77.11	27.46

Table 2: Results of phrase marking

Phrase Level	Method	Ref	Test	Corr	Sub	Del	Ins	SR[%]	ACC[%]
Minor	Mixdorff	209	231	120	7	82	108	57.42	5.74
Phrase <s>	Kruschke	209	236	135	3	71	98	64.59	17.70
Major	Mixdorff	13	14	7	3	3	0	53.85	53.85
Phrase <ss>	Kruschke	13	11	8	3	2	0	61.54	61.54
Minor and Major	Mixdorff	222	245	137	-	85	108	61.71	13.06
Phrase	Kruschke	222	247	149	-	73	98	67.12	22.97

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 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.

7. References

[1] C. W. Wightman and M. Ostendorf, *Automatic Labeling of Prosodic Patterns*, Proc. IEEE Transactions on Speech and Audio Processing, Vol. 2, No. 4, pp. 469-481, 1994.

[2] E. Nöth, A. Batliner, A. Kießling, R. Kompe and H. Niemann, *VERBMOBIL: The Use of Prosody in the Linguistic Components of a Speech Understanding System*, Proc. IEEE Transactions on Speech and Audio Processing, Vol. 8, No. 5, pp. 519-532, 2000.

[3] H. Fujisaki and K. Hirose, *Analysis of Voice Fundamental Frequency Contours for Declarative Sentences of Japanese*, In Journal of the Acoustical Society of Japan (E), 5(4), pp. 233-242, 1984.

[4] W. Wahlster, *Verbmobil: Foundations of Speech-to-Speech Translation*, Springer, 2000.

[5] H. Hussein, M. Wolff, O. Jokisch, F. Duckhorn, G. Strecha and R. Hoffmann, *A Hybrid Speech Signal-Based Algorithm for Pitch Marking Using Finite State Machines*, Proc. Interspeech, pp. 135-138, Brisbane, Australia, September 2008.

[6] D. Hirst and R. Espesser, *Automatic Modelling of Fundamental Frequency Using a Quadratic Spline Function*, Travaux de l' Institut de Phonétique d' Aix, Univ. de Provence, Vol. 15, pp. 75-85, 1993.

[7] H. Mixdorff, *A Novel Approach to the Fully Automatic Extraction of Fujisaki Model Parameters*, Proc. IEEE ICASSP, Vol. 3, pp. 1281-1284, Istanbul, 2000.

[8] H. Kruschke, *Advances in the Parameter Extraction of a Command-response Intonation Model*, Proc. IEEE ISPACS, pp. 135-138, Nashville, Tennessee, USA, 2001.

[9] H. Kruschke and M. Lenz, *Estimation of the Parameters of the Quantitative Intonation Model with Continuous Wavelet Analysis*, Proc. Interspeech, pp. 2881-2884, Geneva, Switzerland, 2003.

[10] A. Bonafonte, H. Höge, H. S. Tropf, A. Moreno, H. Heuvel, D. Sündermann, U. Ziegenhain, J. Perez and I. Kiss, *TC-STAR: TTS Baselines and Specifications*, Deliverable D8. Sept. 2004.

[11] A. Mueller, *Generation of Prosodic Markers for a Multilingual Speech Synthesis System*, TU Dresden, PhD thesis (in German), Dresden, Germany, 2003 (ISBN 3-935712-80-4).

[12] M. Hofmann, *Prosodic and Phonetic Segmentation and Annotation of German and English Speech Corpora*, TU Dresden, diploma thesis, Dresden, Germany, 2006.

[13] H. Mixdorff, *Fujisaki Parameter Extraction Environment (FujiParaEditor)*, `http://public.beuth-hochschule.de/~mixdorff/thesis/fujisaki.html`, 12/11/2009.

[14] O. Jokisch, S. Wittenberg, M. Cuevas, H. Hussein, G. Strecha, H. Ding and R. Hoffmann, *Towards an Automatic Process Chain for the Speech Corpora Annotation*, Proc. SPECOM 2007, pp. 869–875, Moscow, October 2007.