Resynthesis of Prosodic Information Using the Cepstrum Vocoder

Hussein Hussein, Guntram Strecha and Rüdiger Hoffmann

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

Abstract

The naturalness of synthetic speech depends on automatic extraction of prosodic features and prosody modeling. To improve the naturalness of the synthesized speech, we want to apply the concept of Analysis-by-Synthesis of prosodic information. Therefore, the accents and phrases of the speech signal were extracted using the quantitative Fujisaki model in a recognition model. In a generative model we resynthesized the speech signal using a cepstrum vocoder. The excitation signal of the vocoder are the pitch marks (PM), which were calculated from multiple levels of the accent and phrase marking algorithm. A preference test was performed to confirm the performance of the proposed method. For every speech signal four signals were resynthesized according to the calculated PM. Evaluators compared the resynthesized signals with one another. Results show that the quality of the resynthesized signal after prosodic marking is better.

Index Terms: analysis-by-synthesis, prosodic marking, Fujisaki model

1. Introduction

The naturalness of synthesis systems strongly depends on the automatic extraction of prosodic features and modelling of prosodic parameters. The prosodic features include accent, phrase break, pause, pitch level, sentence mood, speech rate, segment duration and voice quality. Prosodic marking is a segmentation and annotation of the speech database.

The first complete system that successfully uses prosody is the Verbmobil system, which is a speech-to-speech translation project in the domain of appointment scheduling dialogues [1].

The Unified Approach for speech Synthesis and speech Recognition (UASR), which was developed in our laboratory for education as well as for research, is a speech dialogue system with the synthesis and recognition components using unified databases. Prosodic features can be extracted from the speech signal, which can be used to improve the recognition result during the recognition process and to reconstruct the original speech signal by adding the prosodic parameters again during the synthesis process [2]. Therefore, a multidimensional analysis approach, which extracts all the prosodic features, is very important. The most important prosodic feature on the linguistic level is the marking of accents and phrases.

To improve the naturalness of synthesized speech, we implemented the concept of Analysis-by-Synthesis of prosodic information using a cepstrum vocoder [2], which is a part of the UASR system. For this purpose, we developed an automatic algorithm for accent and phrase marking [3] in the recognition model. The algorithm, which is based on the analysis of the F_0 contour, uses the quantitative Fujisaki model [4][5]. In a generative model we reconstructed the speech signal with the cepstrum vocoder. We generated the excitation signal of the vocoder from the pitch marks (PM), which were calculated from different levels of the accent and phrase marking algorithm.

The Fujisaki model is a quantitative intonation model, which is used especially in speech synthesis for intonation analysis and intonation generation. The Fujisaki model is described by the following equations:

$$\ln F_0(t) = \ln F_b + \sum_{i=1}^{I} A_{pi} G_p(t - T_{0i}) + \sum_{j=1}^{J} A_{aj}[G_a(t - T_{1j}) - G_a(t - T_{2j})] \quad (1)$$

$$G_p(t) = \begin{cases} \alpha_i^2 t \exp(-\alpha_i t) & t \geq 0 \\ 0 & t < 0 \end{cases} \quad (2)$$

$$G_a(t) = \begin{cases} \min[1 - (1 + \beta_j t)\exp(-\beta_j t), \gamma] & t \geq 0 \\ 0 & t < 0 \end{cases} \quad (3)$$

The Fujisaki parameters are: F_b: baseline value of fundamental 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 $\gamma = 0.9$).

Section 2 presents the detailed concept of the analysis and resynthesis method. Section 3 provides an overview of the database as well as the procedure we used for evaluating the experiment. Section 4 presents the experiment's results.

2. Proposed Method

This section describes the proposed algorithm for accent and phrase marking and resynthesis of the speech signal using the cepstrum vocoder. The block diagram of the proposed method is shown in Figure 1. The algorithm contains the following components:

2.1. Extraction of Pitch Marks

The hybrid algorithm for pitch marking [6], 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

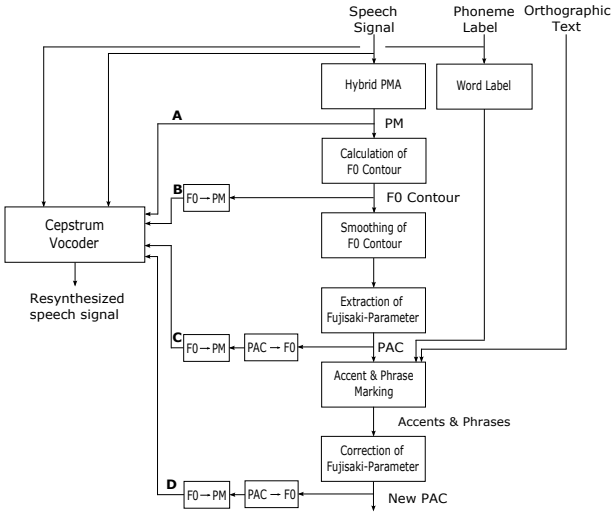

Figure 1: *Framework of the accent and phrase marking algorithm and resynthesis of the speech signal from the extracted prosodic features using a cepstrum vocoder*

peaks of the speech signal and on the selection of more accurate pitch marks that yield the highest confidence score.

2.2. Calculation and Smoothing of F0 contour

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

$$F_{0i} = \frac{F_S}{\text{Length of PM}_j \text{ in samples}} \quad (4)$$

in which F_S is the sampling rate of the speech signal ($F_S F_0$=100 Hz for F_S=16kHz) [7].

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.

2.3. Extraction of Fujisaki Parameter

The Fujisaki parameters were automatically extracted using a continuous wavelet transform (CWT) [9]. The F_b is subtracted from the logarithmic F_0 contour. The result is the residual contour $F_0 rest1(t)$. A CWT using a Mexican hat wavelet is performed to the $F_0 rest1(t)$. The accent commands are detected and optimized. A new F_0 contour is generated from accent commands and subtracted from the $F_0 rest1(t)$. The resulting contour is $F_0 rest2(t)$. Again the Mexican hat wavelet is applied to the $F_0 rest2(t)$ for detecting the 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 were saved in PAC files.

2.4. Accent and Phrase Marking

Two levels of accents and phrases (major and minor) [10] were automatically marked using the Fujisaki model [3]. Accent and phrase marking is based on the word level. The word boundaries were calculated from the available phoneme boundaries. The marking of accent and phrase depended on the analysis of Fujisaki parameters.

2.4.1. Automatic Marking of Prosodic Accents

The perception of prominence depends on the amplitude of the accent command [11]. Some accent commands extend over many words. The word that showed maximum duration within an accent command was marked as accented word. The duration of an accent command in an accented word must be greater than 60% of word duration. Very short or very long accent commands, very small amplitudes of accent commands, and accent comands which were locate in a pause were deleted. The success rate of accent marking (minor and major) is 75.28%.

2.4.2. Automatic Marking of Prosodic Phrases

The correlation between amplitude of phrase command Ap_i and pause duration at the onset of phrase command is high [5]. Therefore, the phrase commands were assigned to the nearest pauses. Phrase commands with a very small amplitudes were deleted. The success rate of phrase marking (minor and major) is 60.95%.

2.5. Modification of Fujisaki-Parameter

The Fujisaki parameters were recalculated after the accent and phrase marking. The false accent and phrase commands were deleted as described in (2.4). In case an accent command was comprised of many words, the new parameters $T1_j\grave{}$ and $T2_j\grave{}$ were modified to the beginning and end time of the accented word. The new parameters $Aa_j\grave{}$ and $\beta_j\grave{}$ contain the same old values of accent command. The new parameters of marked phrase commands were not changed. A new number of accent and phrase commands ($\mathcal{J}$ and $\mathcal{I}$) were detected.

2.6. Generation of F0 Contour from Fujisaki-Parameter

The F_0 contour was generated from Fujisaki parameters (PAC→F0). The F_0 values were calculated from F_b, phrase components and accent components as described in equation (1). Phrase control mechanism $G_p(t)$ and accent control mechanism $G_a(t)$ were computed in (2) and (3) respectively. Individual α_i and β_j values were used. The generated F_0 values are a continuous contour. The minimum value of the generated F_0 contour is F_b. The unvoiced segments, which were detected by the measured F_0 contour, were adapted to the generated F_0 contour. Figure 2 shows the speech signal, measured and smoothed F_0 contour, and resynthesized F_0 contour from Fujisaki parameters and from modified Fujisaki parameters.

2.7. Conversion of F0 Contour to PM

The F_0 contour was converted to a sequence of pitch marks (F0→PM). Therefore, the reciprocal of F_0 values N_P in the voiced frames was computed (in samples) as follows:

$$N_P = \frac{F_S}{F_{0i}} \quad (5)$$

Multiple pitch marks were added in the voiced and unvoiced frames. The distance between two consecutive pitch marks in the voiced frames is equivalent to N_P. The distance between PMs in the unvoiced frames was calculated as 50% of the frame step ($F_S/(2F_S F_0)$). The number of pitch marks in one frame is defined according to the sum of distances between PMs must be less than frame length. The distance for the first pitch mark in a frame is computed from the last pitch mark in the previous frame. The excitation signal for PM is 1 in voiced segments and 0 in unvoiced segments.

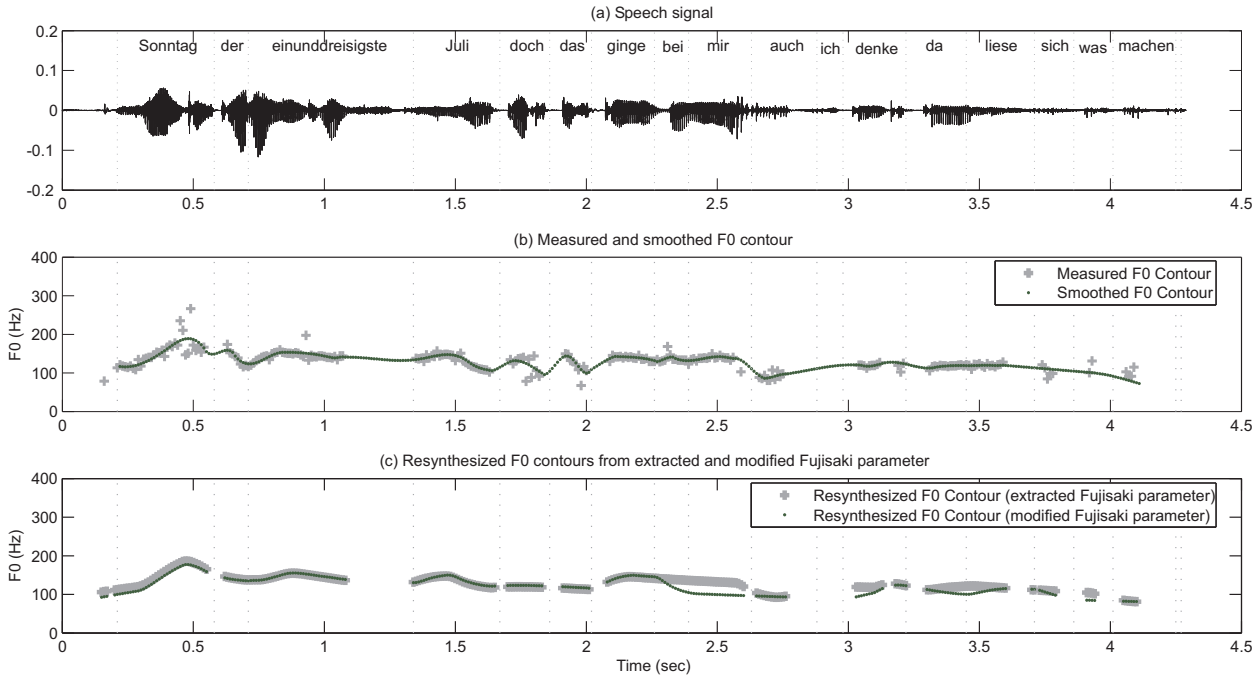

Figure 2: *Speech signal, measured and smoothed F0 contours, and resynthesized F0 contours from Fujisaki parameter and modified Fujisaki parameter. Sentence: "Sonntag, der einunddreißigste Juli. Doch, das ginge bei mir auch. Ich denke, da ließe sich was machen." ("Sunday, the thirty-one of July. However, it would be good. I think we can do something.")*

2.8. Cepstrum Vocoder

The excitation signal of the cepstrum vocoder is generated using the PMs. Therefore, four speech signals were resynthesized according to the PM, which were calculated at four different levels (A, B, C, D) (see figure 1). These levels are:

- A: original PM.
- B: calculated PM from measured F_0 contour.
- C: conversion of Fujisaki parameters to PM.
- D: conversion of modified Fujisaki parameters to PM.

The cepstrum vocoder system used for these tasks is part of a unified HMM-based speech recognition and synthesis system (UASR). In particular, we drove the vocoder to resynthesize existing speech signals as follows:

1. calculating the feature vector sequence from the natural speech signals,
2. generating the excitation signal using the PMs of A, B, C resp. D,
3. synthesizing the speech signal from feature vector sequence and excitation signal using the cepstrum synthesis filter.

3. Experiments

3.1. Speech Material

A subset of the multilingual Verbmobil database (only the spontaneous utterances of German) was used in the preference test. The speech signals of the Verbmobil database were provided with a sampling frequency of 16 kHz and a resolution of 16 bit. The speech signals used for the preference test are five uttered by three males and two females. For every speech signal we resynthesized four signals according to the pitch marks, which were calculated in four positions (A, B, C, D) during the accent and phrase marking algorithm. The total number of resynthesized signals which were used in the preference test is 20. The sentences ranged between five and nine words.

3.2. Evaluation Experiment

Pairs of resynthesized signals were compared. Each pair consists of two resynthesized signals from two different positions in the accent and phrase marking algorithm. This means, for every speech signal we got six pairs of resynthesized signals. For a total of five speech signals we got 30 pairs and each pair was repeated two times in a randomized order. 20 evaluators (12 of them are experts in speech technology) participated in the experiment. They were between 22 and 34 years of age. Evaluators listened to 60 pairs and decided which resynthesized signal had a more preferable quality.

4. Results

The difference in the results between experts and non-experts for female and male speakers in this experiment is not significant. Therefore, we combined the results of both categories. The results of the preference test are reported in Figures 3 and 4. Figure 3 shows the results for all possibilities of comparison. The resynthesized signal from modified Fujisaki parameters (D) have been preferred over the resynthesized signal from Fujisaki parameters (C) for about 3%. All Evaluators preferred the resynthesized signal after accent and phrase marking (C and D) over the resynthesized signal from original pitch marks and measured F_0 contour (A and B) for more than 50%. The result

of comparison of each resynthesized signal with the other signals is shown in figure 4. The resynthesized signal from Fujisaki parameters were also valued more than the other resynthesized signals (34.58%). The resynthesized signal from the measured F_0 contour yielded the lowest quality (10.75%).

The resynthesized F_0 contour from the Fujisaki parameters and from the modified Fujisaki parameters are smoothed (see figure 2, c). Therefore, the results after prosodic marking are better. The measured F_0 contour is not smoothed. Thus, the quality of the resynthesized speech signal from the original pitch mark and from the measured F_0 contour is not satisfying, but it doesn't mean that the quality is bad, because it was difficult for the evaluators to select the best signal.

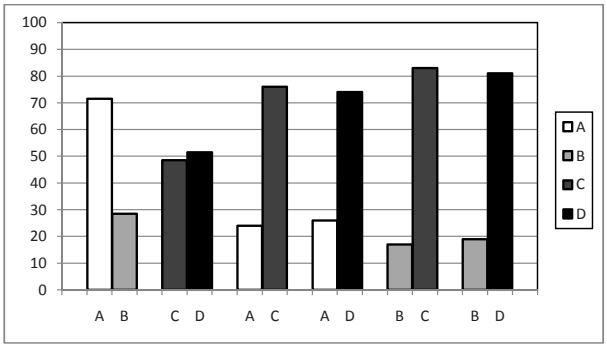

Figure 3: *Results of the preference test for each pair of resynthesized signals*

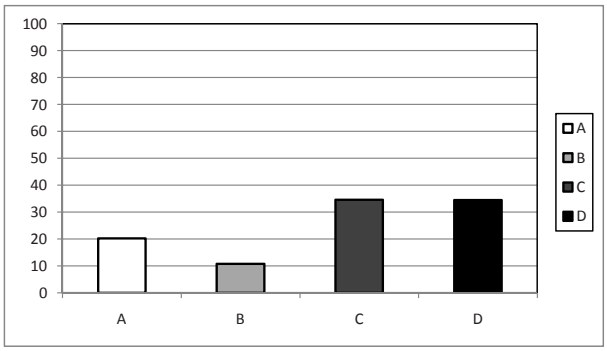

Figure 4: *Results of the preference test for all resynthesized signals*

5. Conclusion and future research

An algorithm for the analysis and resynthesis of prosodic features has been introduced. The algorithm marked the accents and phrases using Fujisaki model. The pitch marks are the excitation signal for the cepstrum vocoder. Therefore, PM from different levels were calculated in the algorithm and were used to resynthesize the speech signal. A preference test was performed to confirm the performance of the proposed algorithm. We used five speech signals from the Verbmobil database uttered by three males and two females. For every speech signal we resynthesized four signals according to the PM. 20 evaluators participated in the preference test to compare pairs of the resynthesized signals. Every evaluator listened to 120 resynthesized signals and determined which signal had a more preferable quality. Results showed that the quality of the resynthesized signal after accent and phrase marking is better than the resynthesized signal resulting from original PM and calculated PM from measured F_0.

A multidimensional approach for the extraction of all prosodic features from the speech signal is required in order to apply the complete concept of Analysis by Synthesis in future research. Therefore, we want to implement an algorithm for automatic marking of pause, pitch level, and sentence mood. The generation of the F_0 contour using all extracted prosodic features and adding it during the synthesis process is the aim of our work.

6. Acknowledgements

Many thanks to Prof. Hansjörg Mixdorff from BHT University of Applied Science in Berlin for his comments on this work and to Dr.-Ing. Matthias Wolff from our laboratory for the description of the cepstrum vocoder and HMM synthesis software.

7. References

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

[2] Eichner, M., Wolff, M., Hoffmann, R., "A Unified Approach for Speech Synthesis and Speech Recognition Using Stochastic Markov Graphs". Proc. ICSLP, Vol. 1, pp. 701-704, Beijing, October 2000.

[3] Hussein, H., "Evaluation and optimization of prosodic labeling algorithms". TU Dresden, diploma thesis (in German), Dresden, Germany, Januar 2007.

[4] Fujisaki, H. and Hirose, K., "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.

[5] Mixdorff, H., "Intonation Patterns of German Model-based Quantitative Analysis and Synthesis of F0 contours". TU Dresden, PhD thesis, Dresden, Germany, 1998.

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

[7] Jäckel, R. and Hussein, H., "Kontrastive Untersuchung zur Realisierung der Fokusakzente in gelesenen Äußerungen (am Beispiel der Sprachenpaarung L1 Russisch–L2 Deutsch)". Proc. 20. Konferenz Elektronische Sprachsignalverarbeitung ESSV (in German), Dresden, Germany, September 2009.

[8] Kruschke, H., "Advances in the parameter extraction of a command-response intonation model". Proc. of ISPACS, pp. 135-138, Nashville, Tennessee, USA, 2001.

[9] Kruschke, H. and Lenz, M., "Estimation of the parameters of the quantitative intonation model with continuous wavelet analysis". Proc. Interspeech, pp. 2881-2884, Geneva, Switzerland, September 2003.

[10] TC-STAR, `http://www.tc-star.org`

[11] Mixdorff, H. and Widera, C., "Perceived Prominence in Terms of a Linguistically Motivated Quantitative Intonation Model". Proc. Eurospeech, Vol. 1, pp. 403-406, Aalborg, Denmark, September 2001.