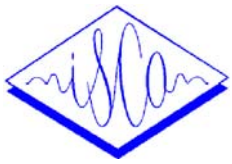

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

LINEAR PREDICTION CODING OF INDIVIDUAL PITCH ACCENT SHAPES

Joachim J. Mersdorf, Kai U. Schmidt, Stefanie Köster

Institute of Communication Acoustics, Ruhr-University 44780 Bochum, Germany

Tel. +49 234 700 2470, Fax. +49 234 709 4165

E-mail: mersdorf@ika.ruhr-uni-bochum.de

Keywords: *Intonation Model, Speaker Transformation, Speech Synthesis and Dialogue Systems*

ABSTRACT

The LPC-Intonation model presented here is dedicated to integrate speaker-dependent features into F0 modeling and analysis. It has been developed for an additional prosodic speaker-transformation based on the command-response approach for intonation features [1,2,3]. The basic principles of this integrated method for automatic analysis, coding, prediction and generation of F0 contours have already been presented in [7,9].
In this paper we are focusing on individual aspects of the resulting LPC filter parameters. Therefore we analyzed F0 contours of the 16 speakers in the German Phondat II Corpus. We found some individual differences of global parameters in speakers' accent shapes that can be also modeled in synthesis. It can be shown that the resulting parameters are necessarily independent of linguistic information. Moreover, there are significant differences for each speaker. When transforming one speaker into another, a set of individual pitch tones for accents ('peaks' and 'valleys') and the resynthesis filter must be changed.

10.21437/Eurospeech.1999-140

1. INTRODUCTION

In the near future we will need appropriate techniques for integrating of non- and paralinguistic features into intonation models. One of the main goals is to realize a greater degree of naturalness and a better performance in dialogue systems and speech synthesis. Important delexicalized information is found in intonation contours [5,6]. Acoustic correlates of F0 contours demonstrate [8] that there is also important information inside the intonation, which is independent of the linguistic and phonetic contents.
For this reason the model's approach is based on the following results of perceptual and statistical F0 contour analysis:

- First of all, we found that the absolute pitch and pitch intervals of accents are speaker-dependent and perceptually relevant in many cases [6]. Frequent tones and tone intervals can be found in the contours, which are vary significantly from speaker to speaker. They can be analyzed independently of the linguistic context and can be separated by means of statistical analysis from the contours. Thus, we aim to quantize the absolute pitch height of all the relevant, local extreme values of the original F0 contour [6].

- The connecting contour between the remaining local extreme values, corresponding to the accent shapes, can be generated by a linear source-filter model [1,3,4,7]. This results in a command-response structure with variable filter parameters. Therefore we developed some rules to automatically computate filter coefficients using linear prediction analysis of F0 contours [7].

- In the presented approach we ignore a declination or phrase component [2] within the model, because we assume the declination to be a consequence of the decrease of accent heights (see chapter 3).

In respect to the command-response approach [3], we also intend to optimize the model by minimizing the excitation energy. This should result in simplified and linguistically unified commands, independent of the speaker. This will allow detection and conversion of speaker-individual features in intonation contours.
Further, we are interested in a more detailed analysis of the influence and relationship between the command signal and the system's response. We assume that there is more than merely a physiological influence on the filter parameters.

2. MODEL OVERVIEW

The LPC-Intonation model [7,9] itself consists of 5 stages: F0-postprocessing, interpolation, analysis, approximation of command excitation and (re)synthesis.

2.1 Postprocessing of PDA

When applying the linear prediction analysis to intonation contours, a careful and very conscientious postprocessing of the pitch determination (PDA) results is required. All the outliers and (sub-)harmonics have to be detected and must be removed, otherwise we could estimate a predictor which is mainly determined by broad, but illegal (e.g. harmonic) accents. Further, longer voiceless parts in contours must be avoided, otherwise a filter that corresponds to the voiced/voiceless ratio of the speech material could be erroneously predicted. Moreover, it is possible that the filter is influenced by the interpolation type in large voiceless parts between the phrases. Additionally the resulting F0 curve must be smoothed and debugged by median and adaptive moving average filtering, especially at the edges of the unvoiced/voiced parts. Finally, the contour must be downsampled to fs = 160 Hz. The long-term spectrum of the final F0 signals has a low-pass structure with frequency components no higher than 5-10 Hz.

2.2 Interpolation

When preparing the computation of the LPC anlaysis, we recommend a continuous, deriveable representation of the F0 contour function. Such a representation can be generated by interpolating the F0 in unvoiced parts. Here we suggest a cubical spline interpolation assuming a *"virtual F0"* in unvoiced segments. The interpolation is motivated by the assumption that a continuous speaker's intonational gesture is only interrupted by temporary switching into unvoiced excitation [6,7,8].

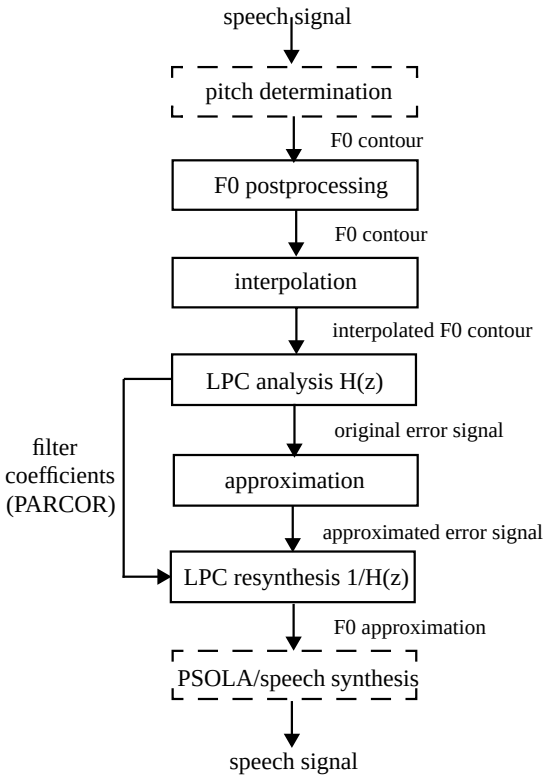

Fig. 1: The Process of LPC-Intonation

2.3 Analysis

The analysis consists of a 4th order LPC of the interpolated contour over the whole phrase or sentence. For the whole speech material a single set of individual filter coefficients can be built by computing the arithmetical mean value for each partial correlation coefficient of all the phrases and sentences.

2.4 Appoximation

One output of the analysis is the prediction error or residual signal. It can be approximated to an appropriate excitation signal for the resynthesis of a contour. But we also experimented with excitation signals which are generated as a command signal in respect to the original contour without any information on the prediction error. Thus, in the next step the command signal is generated by designing appropriate rectangles.

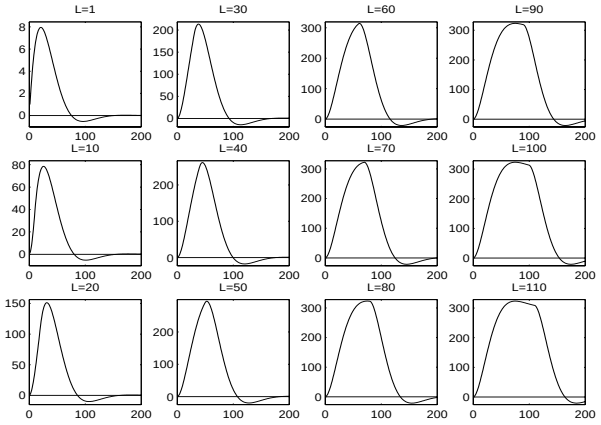

Fig. 2: Example of resulting accent shapes for one speaker (fs=160Hz)

Fig.2 points out different types of accent shapes that can be produced by a single set of filter coefficients. The different accent shapes are the result of convolution with rectangles of different time durations (see Fig.2 L = 1-110 samples). For instance, this method can produce 'early' and 'late' accent peaks. In the resynthesis of F0 contours certain accent shapes can be selected, shifted and scaled, so that they fulfil the conditions given as follows:

- The extension of the rectangle must be scaled, so that the peak of the accent shape is shifted at the same position as the corresponding local exteme value of the original F0 contour (see Fig.3).
- The rectangles' altitude must be scaled so that the peak altitude is the same as the corresponding local exteme value of the original F0 contour. The decay influence of possible previous pulse responses must be taken into account.

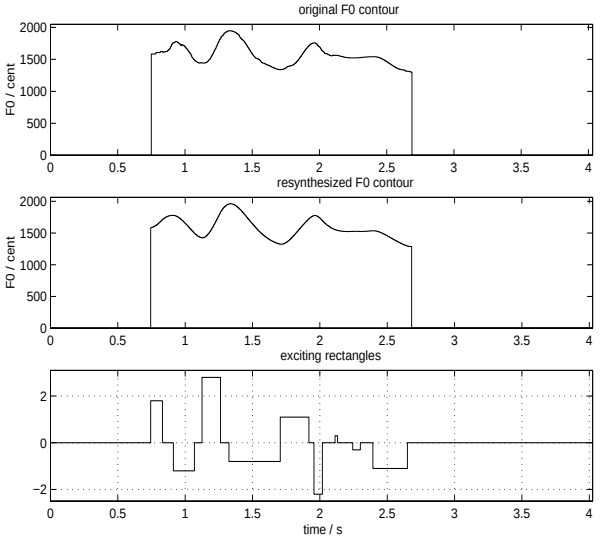

Fig. 3: Example of a resulting contour

The iterative process causes an accent command signal in subsequent, non-overlapping, positive and negative

scaled rectangles. The selection, shifting and scaling is realized by designing appropriate rectangles. The generation of the command signal work stepwise from left to right.

2.5 Synthesis

The final convolution with the filter response from the analysis results in an F0 contour that is very close and perceptually equal to the original (Fig. 3). The approximated and simplified command signal represents the excitation signal of the LPC synthesis. Filter-parameters are taken from the analysis process. Finally, the new F0 contour is applied to the PSOLA-algorithm manipulating natural or synthetical speech.

3. EXPERIMENTAL ANALYSIS

3.1 Experimental Design

The following results are based on the analysis of the German Phondat II Corpus. The corpus contains 200 utterances from train inquiries. Each of them were read aloud by 16 speakers (3200 utterances all in all). 8 speakers are prosodically labeled. From a lexical, syntactical and semantical point of view we can assume that each speaker produced the same material. Thus, overall differences should be the result of the delexicalized speaker's individuality. The 16 speakers were analyzed by the LPC-Intonation model and the resulting filter coefficents were statistically examined. For each speaker the mean values of the resulting 4th order filter coefficients from all the 200 utterances were computed. They were sorted by the 16 speakers and various tests of significance were carried out.

3.2 Experimental Results

For all the coefficients and their different representations (PARCOR, FIR) we found significant differences in the mean values of the coefficients for most of the speakers ($p<0.01$ Dunnet-C3).

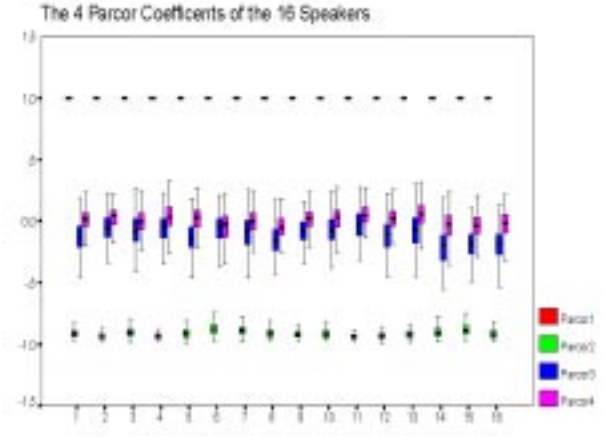

Fig. 4: Individual Filters: Overview by boxplots

Figure No.4 gives a graphical overview using boxplots to illustrate filter differences. Nevertheless, the representation of accent shape individuality using PARCOR-coefficients is difficult to interpret.
For this reason the pulse responses based on the mean filter parameters were printed. They allow a better interpretation of the general features of the individual accent shapes.

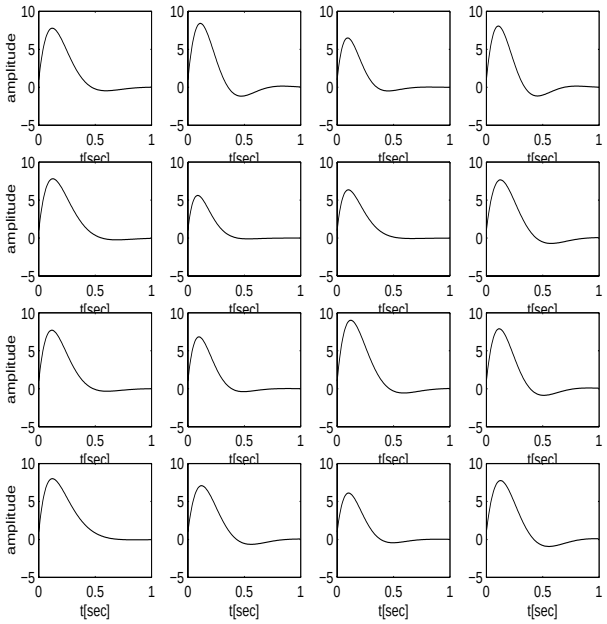

Fig. 5: overall pulse responses for each speaker

If figure no.5 is examined, it can be seen that there are speakers with and without strong 'valleys' after accents. The first result is that there are some speakers with a distinct step-down after intonation accents.

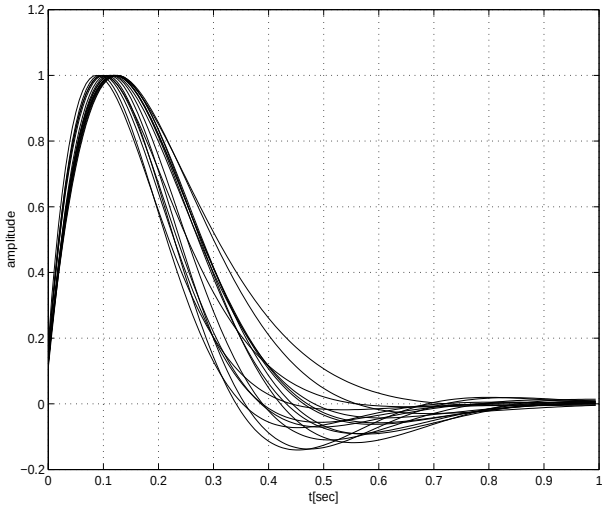

Fig. 6: differences in speaker-individual pulse responses

Figure 6 presents some more individual differences between the speakers. Differences in the steepness of decay after the accent maximum and in the position of zero-crossing can be found. Only a few of them represent the

case of critical damped pulse responses in comparison to the Fujisaki model [1].

4. DISCUSSION

At present only one single 'overall' filter is used for each individual speaker. According to the LPC approch this filter minimizes the speaker's command 'energy' for the analyzed F0 material of 200 utterances.
We recommend that the analysis should be extended to smaller prosodic units, so that we can obtain more than this global individual accent shape information.
In this approach the shifting of accents is not treated as speaker-individual. Also speaker-individual linguistic and phonetic differences are not taken into account.
The aim is a prosodic speaker and speaking-style transformation in speech synthesis and dialogue systems independent of lexical, phonetic and linguistic information. Therefore, perceptually relevant, melodic parameters are focused on.

5. CONCLUSION

We found out that there are speakers with and without distinct valleys after accent peaks. The results can enliven and stimulate the discussion, as to whether there is a declination or a distinct step-down.
Beyond this, a flexible intonation model grow up based on the command-response approach [3].
Together with the linguistic and phonetic analysis, the model can be used for generating F0 in TTS-Systems as usual, but there are some other advantages. In contrast to other approaches, the contour generation requires no analysis by synthesis.
Furthermore, the model allows automatic parametrization and coding of large intonational units, which is becoming increasingly important in phrase concatenative speech synthesis systems.
Accent detection can be supported where the sign of the approximated rectangle height changes from a high positive to a high negative value. Thus, the model can support automatic labeling.
Furthermore, the presented results can help to identify and discriminate speakers in multi-party dialogue scenarios.

REFERENCES

[1] Fujisaki, H. (1983), *"Dynamic Characteristics of Voice Fundamental Frequency in Speech and Singing"*, in: Mac Neilage (Ed.): The Production of Speech, Springer New York, pp. 39-55.

[2] Fujisaki H.(1997), *"Prosody, Models and Spontaneuos Speech"*, in: Campbell N. et al, *"Computing Prosody"* Springer New York 1997

[3] Ohno S. , Hara Y., Fujisaki H., (1999) *"Influences of various factors upon parameters of the command-response model for fundamental frequency contour generation"* ASA'99/Forum Acousticum '99, in: acta acustica/ACUSTICA Vol.85 Suppl. 1, S. Hirzel Verlag, p.377

[3] Möbius B. et al. (1993), *"Analysis and Synthesis of German F0 Contours by Means of Fujisaki's Model"*, in: Speech Communication 13, North-Holland 1993, pp. 53-61.

[4] Mixdorff. H. (1998), *"Intonation Patterns of German, Model-based Quantitative Analysis and Synthesis of F0 contours"*, Ph.D.Thesis TU Dresden 1998

[5] Higuchi N., Hirai T., Sagsaka Y. (1996), *"Effect of Speaking Style on parameters of fundamental Frequency Contour"*, in: J.P.H van Santen et al. (ed.) Progress in Speech Synthesis, Springer New York 1997

[6] Mersdorf J., Domhöver T., (1997) *"A Perceptual Study for Modelling Speaker-Dependent Intonation in TTS and Dialog Systems"*, in: Conference Proceedings of Eurospeech'97, Rhodes, pp. 867-870

[7] Mersdorf J., Rinscheid A., Brüggen M., Schmidt K.U., (1997) *"Coding of Large Intonational Units by Linear Prediction"*, ESCA Tutorial and Research Workshop on Intonation: Theory, Models and Applications 1997 Athens, pp. 235-238.

[8] Mersdorf J., (1998) *"The Sound Of Prosody: A Perceptual Approach"*, in Conference Proceedings of Euronoise'98, Designing for Silence, edited by H.Fastl, (DEGA-Verlag, Oldenburg 1998), pp.545-550

[9] Mersdorf J., Schmidt K.U., (1999) *"Analysis and Synthesis of F0 Contours by Linear Prediction Coding"* ASA'99/Forum Acousticum '99, in: acta acustica/ACUSTICA Vol.85 Suppl. 1, S. Hirzel Verlag, p.320