Automated Quantitative Analysis of F_0 Contours of Utterances from a German ToBI-Labeled Speech Database

Hansjörg Mixdorff [1] *and Hiroya Fujisaki* [2]

[1]Technical University Dresden [2]Science University of Tokyo
[1] Institute of Technical Acoustics
Mommsenstr. 13, 01062 Dresden, Germany
E-mail: mixdorff@teles.de

ABSTRACT

The present paper proposes a method for automating the analysis of F_0 contours using the Fujisaki-model on ToBI-labeled speech data. ToBI-labels are used to preselect the number of necessary phrase and accent commands and align the onsets and offsets of these commands with the segments of the utterance. Local optimization is then performed with special regard to 'reliable' portions of the F_0 contour, for instance, the syllable nuclei. Analysis results are used for formulating quantitative F_0 control rules for speech synthesis.

1. INTRODUCTION

In German, as in many other languages, the fundamental frequency contour (henceforth the F_0 contour) is an important acoustic correlate of accent and intonation. An accurate, quantitative modeling of F_0 contours which is based on a linguistically motivated description of intonation, is indispensable for high intelligibility and naturalness of text-to-speech (TTS) synthesis. Although there exist a number of models for describing German intonation we concentrate on two approaches with relevance to the present paper.

Isačenko and Schädlich [1] postulated that a given F_0 contour can be simplified to a sequence of 'tone switches' between two constant levels of F_0 at accented syllables while preserving the linguistic content of the utterance and proved their hypothesis in perceptual experiments. Tone switches can occur early or late in an accented syllable whereas utterance-medial syntactic boundaries are characterized be 'pitch interrupters'.

The tone sequence approach applied to German by Féry [2] and Uhmann [3] originates from the work of Pierrehumbert [4] who describes F_0 contours as a sequence of high (H) and low (L) tones. These are associated with accented syllables and prosodic boundaries. The actual F_0 range between 'high' and 'low' tones is only locally defined and subject to rule-determined reductions. An utterance is built of a number of parallel hierarchical layers, 'tiers', which consist of sequences of phonological segments. The 'tone tier' corresponds to the phonological elements which make up the F_0 contour. It consists of 'accent tones' assigned to accented syllables and 'boundary tones' describing the course of the F_0 contour at prosodic boundaries. 'Break indices' are used to denote the strength of prosodic boundaries with higher level boundaries being assigned higher indices.

Both approaches present phonologically motivated descriptions of F_0 contours, the former based on perception, the latter mainly based on visual inspection of the F_0 contour. They are, however, not sufficient for synthesizing natural-sounding F_0 contours for speech synthesis since they do not permit a quantitative description of intonational events, i.e. the interval spanned by the tone switch, or an assessment of the actual 'height' of high and low tones, nor the exact timing of intonational events in relationship to the segmental string.

2. METHOD OF ANALYSIS AND SPEECH MATERIAL

In an earlier work [5], the present authors addressed the problem of how prosodic rules for TTS can be derived from the analysis of natural F_0 contours.

2.1 Model of the F_0 contour

The quantitative Fujisaki-model [6] has been shown to be capable of producing close approximations to a given contour, expressed as a time function of the logarithm of the fundamental frequency, from two kinds of linguistically meaningful input commands: phrase commands (impulses) and accent commands (stepwise functions), which are characterized by the following model parameters: Ap: phrase command magnitude; T_0: phrase command onset time; Aa: accent command amplitude; T_1: accent command onset time; T_2: accent command offset time.

The phrase components produced by the phrase commands account for the declination characteristics of each of the prosodic phrases, and thus constitute the global shape of the F_0 contour. The accent components produced by the accent commands determine the local shapes of the F_0 contour, and are closely related to the word accents. These components are superposed on a baseline, represented by $\log Fb$ (where Fb indicates the base frequency), to form an F_0 contour in the $\log F_0$ domain (i.e. $\log F_0(t)$).

Since the underlying commands are sometimes difficult to infer from a given F_0 contour, modeling must be performed on the basis of a number of constraints. in order to ensure a linguistically meaningful analysis. In our earlier study [5] the initial numbers of phrase and accent commands used for the analysis of a particular contour were pre-selected considering the numbers of accented syllables and phrase boundaries as well as the linguistic content of the utterance. This procedure was rather time-consuming and hence the amount of speech data analyzed comparatively small. Rule-generation for TTS, however, requires analysis of larger speech corpora.

The present paper describes an approach for automating the analysis of F_0 contours of a ToBI-labeled speech database [7]. The database contains spontaneous speech from a scheduling task.

2.2 Linguistic Background

In our approach for German, the accent components of the Fujisaki-model are basically defined by major transitions of the F_0 contours (the locations of Isačenko's and Schädlich's 'tone switches') at accented syllables. Hence each tone switch is connected to either the onset or the offset of an accent command. In order to denote the linguistic function of each accent we apply the term 'intoneme' (after Stock [8]).

Analysis of natural speech data showed that additional tone switches occur at phrase boundaries which are not necessarily connected to accented syllables. In the tradition of Pierrehumbert [4], we call these 'boundary tones'.

Hence a given F_0 contour is described as a sequence of the following intonational events:

N(on-terminal)-intoneme Rising tone switch at utterance-non-final accents signaling continuation

I(nformation)-intoneme Falling tone switch at utterance-final accents signaling completion of a message.

C(ontact)-intoneme Rising tone switch at utterance-final accents in intonationally marked questions, signaling that the speaker wishes to establish contact.

Boundary tone B_{cat} Question-final rise in globally marked questions, following a C-intoneme

Boundary tone B_{nocat} Question-final rise in locally marked questions.

The phrase components are defined by the onsets of phrase commands which are typically linked to major syntactic boundaries and cause a readjustment of the declination line of the F_0 contour. The portion of the F_0 contour between consecutive phrase commands is called a 'prosodic phrase'.

3. GERMAN TOBI SYSTEM AND TONE SWITCHES

Table 1 gives an excerpt from the tone inventory as used by Reyelt [7] for labeling the VERBMOBIL-corpus and the correspondence between ToBI-labels and tone switches. '*' denotes a tone linked to a prominent accent syllable, '%' denotes a tone linked to a syllable left of a prosodic boundary. Word boundaries with break indices above B2 are labeled with a sequence of 'phrase tone' and 'boundary tone', the former being the tone on the pre-final syllable and the latter being the tone on the final syllable of the phrase. Ordinary inter-word boundaries are labeled 'B1', and 'B3' denotes boundaries of intonational phrases. Break indices 'B9' denote irregular phrase boundaries due to hesitation etc.

Some of the tone labels applied by Reyelt directly correspond to tone switches such as L+H* (early rising tone switch) and L*+H (late rising tone switch). Accent labels H* and H+!H* both imply a sequence of rising and falling tone switches.

4. AUTOMATED ANALYSIS

Although ToBI-labels represent a rather impressionistic, qualitative description of the F_0 contour, they can be used to pre-select a minimum number of phrase and accent commands and their approximate locations. This has been shown for Japanese by Hirai and Higuchi [9].

4.1 Modeling Procedure

In the current approach, the initial values of the onsets of accent commands are aligned with L-H label sequences and the initial values of the offsets with H-L label sequences. Accents with the labels H* or H+!H* are assigned an accent command each.

The onset of the utterance and boundaries with break indices above B2 are assigned a phrase command each.

The base frequency Fb is assumed to be almost constant over a set of utterances of the same speaker, and

Table 1. Labels used by Reyelt et al. (excerpt)

Function types	Label	Phonetic Realization	tone switches
Accent Tones	H*	standard 'peak' accent	rising/falling
	H+!H*	early peak, accent syllable perceptually low	early falling
	L + H*	low syllable followed high accented syllable	early rising
	L* + H	low accent syllable followed by F_0 rise	late rising
	L*	F_0 valley in the nucleus of accented syllable	early falling
Boundary Tones	L-L%	low terminal phrase boundary	none
	H-H%	high phrase boundary	rising
	L-	low phrase boundary (boundaries < B3, and B9)	falling
	H-	low phrase boundary (boundaries < B3, and B9)	none

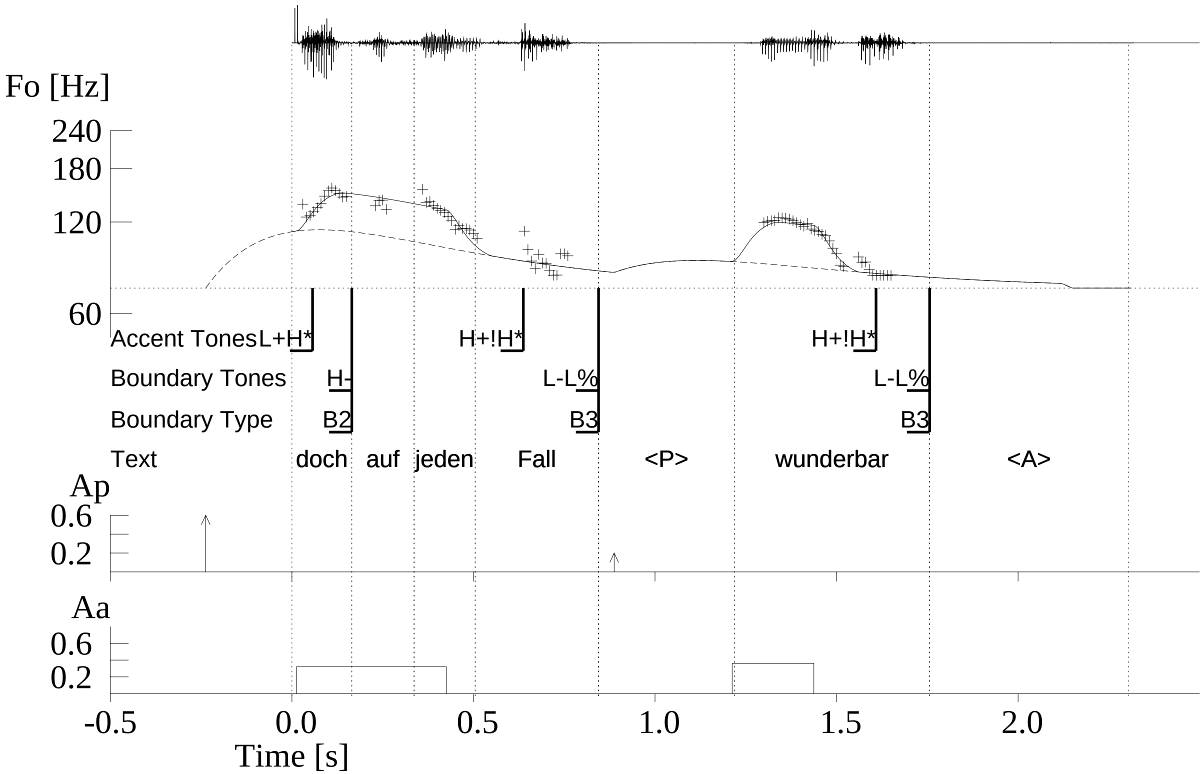

Fig. 1. An example of analysis of a tone-labeled utterance from the VERBMOBIL-corpus showing the correspondence between ToBI-labels and Fujisaki-model analysis: "Doch, auf jeden Fall, wunderbar."—"*Certainly, in any case, wonderful!*".

its initial value is set equal to the lowest value of F_0 found in the set of utterances by one speaker. These parameter values, however, are modified to minimize the least mean square error between the measured F_0 contour and the model-generated contour in the log F_0 domain. For this optimization, values in the F_0 contour are weighted according to their 'reliability', i.e., the value of the autocorrelation function at the delay value $1/F_0$ and the intensity found in the respective frame, in order to favor the vowel nuclei of accented syllables which are perceptually important.

4.2 Results

Figure 1 shows an example of analysis of the utterance "Doch, auf jeden Fall, wunderbar."—"*Certainly, in any case, wonderful!*" taken from the VERBMOBIL-corpus. From the top to the bottom it displays: The speech waveform, the extracted F_0 contour ('+'-signs), the model-generated contour (solid line), the tone labels (accent and boundary tones), boundary type (break indices), text of utterance, phrase commands, and accent commands. The dotted vertical lines denote word boundaries, whereas the solid vertical lines denote the alignment of ToBI-labels with the F_0 contour. It can be seen that the beginning of the utterance and B3-boundaries are assigned a phrase command each. The tone label L+H* at 'doch' coincides with the onset of an accent command, whereas labels H+!H* at 'Fall' and 'wunderbar' coincide with

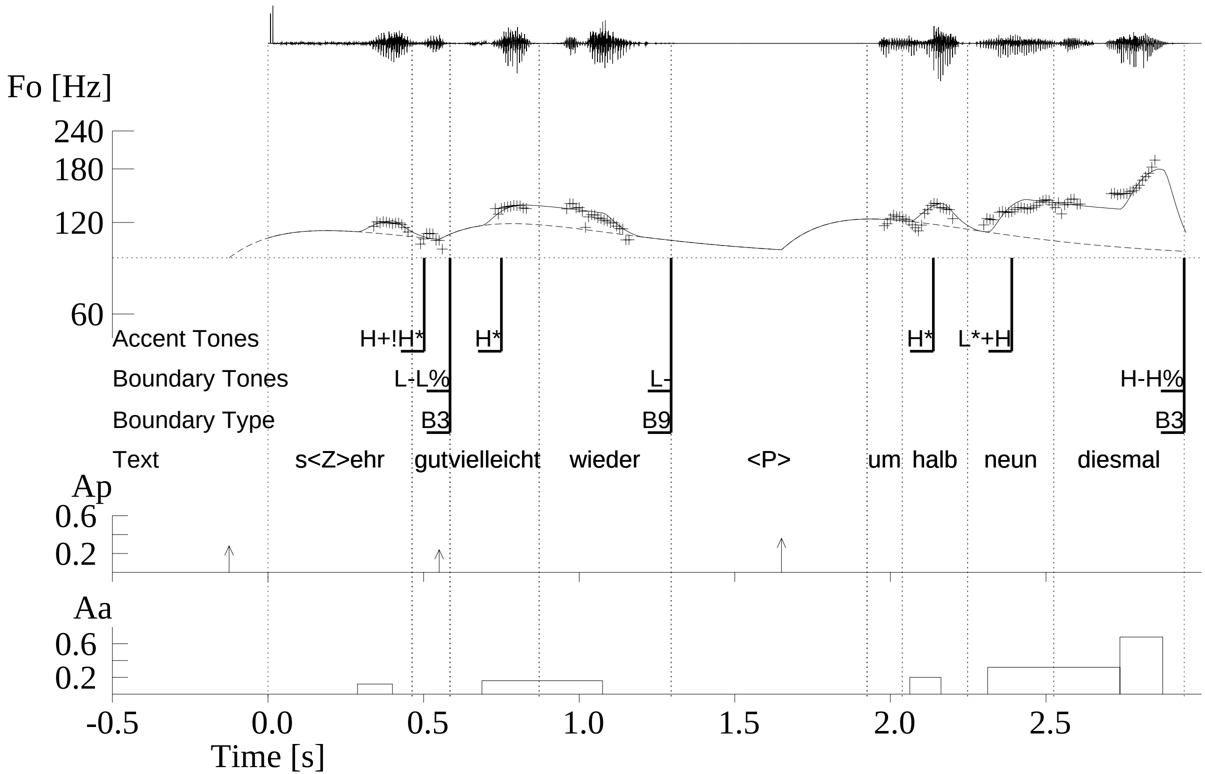

Fig. 2. An example of analysis of a tone-labeled utterance from the VERBMOBIL-corpus: "Sehr gut, vielleicht wieder...um halb neun diesmal ?"—"*Very good, maybe once again...at half past eight this time ?*"

accent command offsets. Whereas the former functionally represents an N-intoneme, the latter two correspond to the statement-final I-intoneme.

Figure 2 shows an example of analysis of a utterance of the question "Sehr gut, vielleicht wieder...um halb neun diesmal ?"—"*Very good, maybe once again...at half past eight this time ?*". The last accent in the utterance is labeled 'L*+H' and hence connected with the onset of an accent command. Functionally, this configuration corresponds to a C-intoneme which is followed by a concatenated boundary tone B_{cat} at the tail of the utterance (label 'H-H%').

5. CONCLUSION

The current study introduced a method for determining Fujisaki-model parameters from a ToBI-labeled database. Since ToBI-labels are commonly applied for labeling larger German speech corpora, these corpora can now be used for quantitative analysis of intonational events. Analysis results can be used for developing parametric rules for generating F_0 in speech synthesis. Furthermore, the Fujisaki-model has been shown to be also applicable to utterance of spontaneous speech. Work remains to be done for automating the statistical evaluation of analysis results and determining 'intonational profiles' for particular speakers.

REFERENCES

[1] A.V. Isačenko and H.J. Schädlich. *Untersuchungen über die deutsche Satzintonation.* Akademie-Verlag, Berlin, 1964.

[2] C. Féry. Rhythmische und tonale Struktur der Intonationsphrase, 1988.

[3] S. Uhmann. Akzenttöne, Grenztöne und Fokussilben. Zum Aufbau eines phonologischen Intonationssystems für das Deutsche, 1988.

[4] J.B. Pierrehumbert. *The phonology and phonetics of English intonation.* PhD thesis, MIT, 1980.

[5] H. Mixdorff and H. Fujisaki. A scheme for a model-based synthesis by rule of f0 contours of german utterances. In *Proceedings of Eurospeech '95, vol. 3*, pp. 1823–1826, Madrid, Spain, 1995.

[6] H. Fujisaki and K. Hirose. Analysis of voice fundamental frequency contours for declarative sentences of japanese. *Journal of the Acoustical Society of Japan (E)*, 5(4), pp. 233—241, 1984.

[7] M Reyelt and A. Batliner. Memo 33: Ein Inventar prosodischer Etiketten für Verbmobil, 1994.

[8] E. Stock and C. Zacharias. *Deutsche Satzintonation.* VEB Verlag Enzyklopädie, Leipzig, 1982.

[9] T. Hirai and N. Higuchi. Automatic extraction of fundamental frequency control rules using Japanese Tone and Break Indices (j-tobi) system. *Technical report of the Institute of Electronics, Information and Communication Engineers, vol. SP97* , pp. 27—32, 1997.