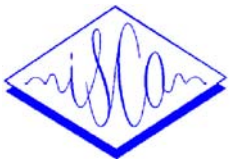

Speech Prosody 2002
Aix-en-Provence, France
April 11-13, 2002

Speech Technology, ToBI, and Making Sense of Prosody

Hansjörg Mixdorff

Faculty of Computer Science, Berlin University of Applied Sciences, Germany
mixdorff@tfh-berlin.de

Abstract

The current paper critically examines why prosodic knowledge has not yet found its way into commercial applications of speech technology. As a key issue of potential improvements to speech recognition and synthesis we identify the capability of understanding and expressing meaning by means of prosodic features of speech. We suggest that even a complete and 'correct' ToBI transcription will always remain some kind of intermediate and possibly incomplete stage of representation between the intended meaning of a message and the resulting speech signal. Examining the correspondences between a version of G-ToBI and the quantitative syllable-based integrated model developed by the author which uses the Fujisaki model for parametrizing *F0* contours we conclude that ToBI accent labels can be derived from Fujisaki parameters. Finally we show that perceived prominence which can be thought of as the result of perceptual integration of various prosodic cues with respect to the information structure of an utterance can be reliably predicted from accent command amplitudes and normalized syllabic durations.

1. Introduction

"Prosody is the key to meaning..."

The general dilemma with prosody research at present is that after all we still know very little about suprasegmental features of speech, and the little we know is split up into a large number of competing approaches and many different languages. Depending on the traditions followed by various researchers we find linguistics based approaches and more technology oriented ones. Although we have made progress with respect to the conviction of traditional grammarians who assumed that prosody was entirely predictable from syntax and constituency, there seems to be an unbridgeable gap between linguistics and speech processing communities.

This results in a large number of competing prosody 'schools', ToBI probably being the most famous. Only for German, intonation models applied in speech synthesis range in the dozens, and we have seen over the years several versions of German ToBI (see, for instance, [1][2]). As a consequence, research is split up not by the topics of studies, but by the underlying theoretical frameworks.

Unfortunately there is nothing like the forming of a standard in sight, nor a unified interface, not to be talking about a representation linking recognition and synthesis, or for comparing different languages.

In this talk we will first discuss some of the reasons why prosodic knowledge has found its way into relatively few applications of speech technology. We will then attempt to define requirements for prosodic models that will be 'usable'. We then will show the common grounds of G-ToBI, and the syllable-based Fujisaki model employed by the author, and explain why we feel that an entirely symbolic representation is likely to capture only part of the prosodic information. In the following we will revisit a study on syllable prominence showing that Fujisaki parameters are appropriate not only in terms of the production process of *F0*, but also with respect to perception. We conclude the paper with suggestions for future developments.

2. Requirements to Prosodic Models

We feel that the ultimate challenge of representing prosody and utilizing prosodic information in the speech signal is not how to script it, but to derive the meaning of an utterance from the speech signal. Even if we 'correctly' annotate prosodic features, the next step will be to integrate the transcription along with the segmental information into an equivalent of meaning. Though meaning can also be thought of something that is categorical, say, for instance, the difference between an echo question and an exhortation, the borderlines are somewhat fuzzy.

Ideally speaking a model of prosody applicable in speech technology should be bi-directional, that is, use the same or at least similar intermediate representations for speech synthesis as well as recognition, in order to be fully transparent and compatible with respect to data base sharing, for instance.

In speech synthesis we wish to produce naturally sounding speech that conveys an intended meaning from a set of symbolic and phonologically motivated units. In recognition we aim at detecting the same phonological units from the speech signals. Information units can be subdivided into the following categories:

Linguistic: lexical stress, sentence modality (question vs. non question), focus structure, segmentation.
Para-linguistic: speaker's attitude, intention, dialect, sociolect
Non-linguistic: health condition, emotional state, etc.

Hence current approaches to describing prosody aim at establishing a mapping between phonologically motivated entities of information and their phonetic realization manifesting in prosodic features, such as the *F0* contour. This mapping is usually performed by some sort of parametrization using superpositional [3] or shape-based (see [4][5], for instance) formulations which yield timing and amplitude information for intonational events. The distinction between prosodic function and form, however, cannot be strictly made, and even recent definitions of G-ToBI loosen the claim of being strictly phonology based [6].

3. Prosody and Speech Technology

Some general reasons why the speech industry has not yet widely considered the use of prosodic features in commercial applications might be the following:

- the research community does not yet provide a consistent prosodic framework neither for a single language, nor for the multi-lingual description of prosody
- prosody is very much speaker-dependent, even ideosyncratic (*F0* range, speech rate, etc.), so the issue of normalization is crucial with problematic implications for speaker-independent recognition, for instance
- we know relatively little about the way in which different types of information (linguistic, para-linguistic, non-linguistic) are coded into bundles of prosodic features - not just a single feature such as *F0*

Still, the situation for speech synthesis is clearly different than that of speech recognition: Whereas any modern TTS system incorporates some kind of standard 'read speech style' prosody model, there are much fewer applications in speech recognition.

3.1. Speech Synthesis

Listening to some of nowadays' best commercial TTS systems we might be led to think that synthesis is a solved problem. Current state-of-the-art systems sound very natural, but still not as if the computer knew what it was talking about. We have to admit though that for current applications of TTS there exist technologically feasible solutions. It seems embarrassing for the prosody community that most parametric approaches based on loads of prosodic knowledge sound very poor compared with corpus based synthesis. As a consequence knowledge has moved from the parametric rule systems to the database labels.

Based on the observation that synthesis from text as a highly impoverished representation will exhibit a stereotypical prosody, there have been proposals for mark-up languages for enhancing the symbolic prosodic input to a synthesizer. This kind of mark-up can especially be justified when the synthesizer is running in a limited domain, concept-to-speech mode, where we find highly recurrent prosodic patterns (weather reports, stock reports, etc.).

Still we might not have yet come to exploiting sufficiently the state-of-the-art syntactic and semantic parsing techniques provided by Natural Language Processing (NLP) in the prosodic pre-processing of our TTS systems, and there is still room for improvement, even if we deal with text as the input.

3.2. Speech Recognition

Current commercial recognizers work relatively well because they calculate the probability of a phone in the context of a word in the context of a phrase and maximally normalize the speech signal. Therefore, from an engineering point of view, even if we wish to enhance a phone-based recognizer with prosodic information, we run into unsolvable trouble as due the recognizer architecture it is hard, if not impossible, to get at the segmental alignment information.

On the other hand we have to keep in mind that prosody is not everything in ASR. In the presence of other information which facilitates top-down processing (syntactic structure, morphemic markers) we might not even need it. Yes/no question in Japanese or Finnish, for instance, are marked by question particles [7]. Therefore there is no need for a rising *F0* at the end of utterances of yes/no questions. Besides, prosody is not an issue in nowadays applications of ASR because

- we hope to deal with a cooperative user
- currently dialog models are mostly machine-guided
- most recognizer are expected to work speaker-independently. Although we have robust acoustic models, we lack this kind of model on the prosody side and we do not want recognizers to be confused by erroneous prosody recognition.

What can we gain by exploiting prosodic cues ? A few benefits may be the following:

- detection of minor (non-pause) boundaries
- dialog act classification (declarative, interrogative, unfinished, ...) where morphemic markers are absent
- detection of focus structure (important and less important words)
- emotion detection.

The advantages are obvious: A prosody-aware recognizer can facilitate a more flexible dialog with topic shifts induced by the user. It can also detect which items in the discourse are most important to the user.

4. Comparing G-ToBI with a Syllable-based quantitative Model of German Prosody

This section aims at pointing out the common grounds of two conceptually different approaches to modeling prosody and how their representations could be mapped onto one another. The approaches compared are a version of G-ToBI developed at IMS Stuttgart [1], and the syllable-based integrated model IGM recently proposed by the author [8]. During the development of IGM a larger speech data base was analyzed in order to determine the statistically relevant input features. This work provided the opportunity for comparison with the Stuttgart version of German ToBI [9]. The corpus is part of a German corpus compiled by the Institute of Natural Language Processing, University of Stuttgart and consists of 48 minutes of news reports read by a male speaker [10], of a total of 13151 syllables.

The corpus contains boundary labels on the phone, syllable and word levels and linguistic annotations such as part-of-speech and ToBI labels following the Stuttgart system which will be discussed in the following.

4.1. The Stuttgart G-ToBI System

For the sake of brevity we only point out some of the properties of the Stuttgart G-ToBI system. The basic accent types are L*H (a rise from a low accent syllable) and H*L (a fall from a high accent tones), augmented by HH*L (early high peak) and L*HL (rise-fall / "late peak"), boundary tones assigned are H% and L%. The system of break indices follows the conventions of Pierrehumbert as to 0: clitic, 1:

word, 2: disjuncture without tonal cue, 3: intermediate, and 4: intonation phrase boundary [11].

4.2. A Syllable-Based Integrated Model (IGM)

Recent work by Mixdorff was dedicated to an integrated model of German prosody [8] (henceforth IGM) anchoring prosodic features such as *F0*, duration and intensity to the syllable as a basic unit of rhythm. In the framework of IGM, following the works by Isacenko & Schädlich [12] and Stock & Zacharias [13], a given *F0* contour is described as a sequence of linguistically motivated tone switches, major transitions of the *F0* contour connected to accented syllables, or by so-called *boundary tones* before prosodic boundaries. Tone switches can be thought of the phonetic realization of phonologically distinct intonational elements, so-called 'intonemes'. In the original formulation by Stock, depending on their communicative function, three classes of intonemes are distinguished, namely the N↑intoneme ('non-terminal intoneme' at phrase-medial accents, rising tone switch), I↓ intoneme ('information intoneme' at declarative-final accents, falling tone switch), and the C↑ intoneme ('contact intoneme' associated with question-final accents, rising tone switch). Hence intonemes in the original sense mainly distinguish sentence modality, although there exists a variant of the I↓ intoneme, I(E)↓ which denotes emphatic accentuation and occurs in contrastive environments, for instance. Intonemes – except for I(E)↓ which is governed by the context of an utterance - are predictable by applying a set of phonological rules to a string of text as to word accentability and accent group forming.

In order to quantify the interval and timing of the tone switches with respect to the syllabic grid, IGM employs the well-known quantitative Fujisaki formula [2] for parametrizing the natural *F0* contours [14]. The Fujisaki model has been shown to be capable of producing close approximations to a given contour from two kinds of input commands: phrase commands (impulses) and accent commands (stepwise functions). Different from other formulations (see, for instance, [4]) the *F0* contour can be decomposed into two tiers, the accent tier and the phrase tier. The interval of a tone switch, for instance, readily relates to the accent command amplitude *Aa* assigned to it, and as will be shown in the following, BI4 boundaries are reliably linked to phrase commands.

An additional attraction of the Fujisaki model is the physiological interpretation which it offers for connecting *F0* movements with the activity of intrinsic larynx muscles[15].

4.3. Fujisaki Parametrization, ToBI and Intonemes

From the signal processing point of view, the Fujisaki model facilitates a smoothing and interpolation of voiceless portions of the raw *F0* contour, providing a value of *F0* for every point in time. On a corpus of news readings we have shown that the estimation of Fujisaki parameters can be automated to a high degree [15] though possibly error-prone thresholds concerning the minimum duration of accent commands, the minimum distance between succeeding accent and phrase commands etc. must be applied. As has been shown, the original *F0* contour can be reproduced very faithfully from estimated Fujisaki parameters. It is obvious that this statement cannot be made for the G-ToBI representation, as it supplies symbolic markers for nuclear accents and their alignment with the accented syllable, phrase accents, boundary tones and boundary indices, and hence lacks quantitative information. We can, however, attempt to draw a parallel between tone labels and the intoneme classes, as both denote phonologically distinct categories and are explicitly linked to the segmental tier, namely accented syllables. In particular, as will be shown, accent labels L*H coincide with N↑intonemes and H*L with I↓ intonemes.

4.4. Comparison of Fujisaki parameters and G-ToBI labels

Figure 1 displays an example of analysis, showing from top to bottom: the speech waveform, the extracted and model-generated *F0* contours, the duration contour in terms of the syllabic z-score drawn as horizontal lines of the length of the respective syllable, the ToBI tier, the text of the utterance, and the underlying phrase and accent commands.

Accent Assignment. The corpus contains a total number of 13151 syllables. Of the 2498 syllables labeled as accented 96.1% were found to be linked to accent commands, 177 syllables were marked with H% boundary tones receiving a separate accent command which not linked to a preceding accent (see, for instance, the accent command assigned to the word 'abgegolten' in Figure 1). Accents followed by a BI3 or BI4 boundary are found to be significantly stronger (with a mean accent command amplitude *Aa* of 0.38) than non-boundary accents with a mean *Aa* of 0.26.

Non-downstepped' accents (98.0% of all accent labels) exhibit a mean accent command amplitude of 0.28 against 0.21 for accents labeled as down-stepped. Furthermore, accents marked as uncertain ('?', 1.9 % of all accent labels) exhibit significantly lower *Aa* than those labeled with certainty (0.21 against 0.28). This indicates that it is the assessment of weaker accents that usually poses problems to the labeler. The standard accent types 'H*L','L*H','HH*L' and 'L*HL' which account for 84% of the accent labels can be reliably identified by the alignment of the accent command with respect to the accented syllable, expressed as $T1_{dist}=(T1-t_{on})$; and $T2_{dist}=(T2-t_{off})$ where *T1* denotes the accent command onset time, *T2* the accent command offset time; t_{on} the syllable onset time and t_{off} the accented syllable's offset time. For type 'H*L', mean $T1_{dist}$ and $T2_{dist}$ are -60 ms and -37 ms, and for type 'L*H' 132 ms and 168 ms, respectively. In a similar manner, the HH*L ('early high peak') (-215 ms/-172 ms) and L*HL accent types (rise-fall / "late peak") (27 ms/-68 ms), can be associated with the timing of the underlying accent command.

A considerable number of syllables (N=444) exhibit accent commands but not any accent label. Figure 1 shows such an instance where in the utterance "Zudem sollen Überstunden nur noch in Freizeit abgegolten und die Lohnnebenkosten gesenkt werden." -*"Furthermore, overtime will be compensated by time off in lieu only, and additional costs of wages are to be reduced."*, an accent command was assigned to the word 'nur', but not a tone label. Closer analysis shows that labels are mainly missing when accents are relatively weak or in the case of secondary accents of longer compound words.

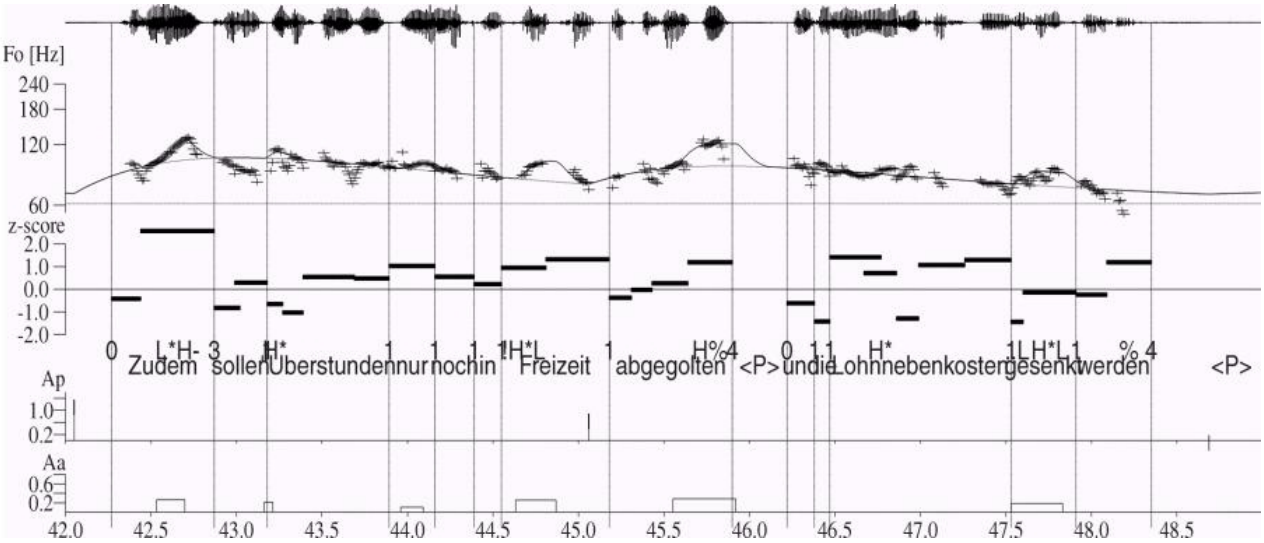

Figure 1: An example of analysis from the data base. In the utterance "Zudem sollen Überstunden nur noch in Freizeit abgegolten und die Lohnnebenkosten gesenkt werden." "Furthermore, overtime will be compensated by time off in lieu only, and additional costs of wages are to be reduced." the third accent command marks a minor accent on 'nur'/'only' which was not assigned a ToBI-label.

Phrase Boundaries. About 54.8% of break index 3- and 96.2% of break index (BI) 4-labeled-boundaries are aligned with the onset of a phrase command, with a mean phrase command magnitude *Ap* of 0.67 and 1.32, respectively.

It must be stated, however, that the assignment of BIs by the labeler was sometimes inconsistent as boundaries with quite different prosodic cues and syntactic depths were assigned the same BI. Prosodic cues observed for boundaries include declination line resets - as triggered by phrase commands -, pauses, boundary tones and pre-boundary lengthening, the latter sometimes being the only cue at BI3 prosodic boundaries. As can be seen in Figure 1, the BI 3 boundary after 'Zudem' is mainly signaled by a durational cue (z-score=2.8 on the syllable 'dem'), whereas the BI4 boundaries after 'abgegolten' und 'werden' exhibit durational cues, as well as pauses. The sentence-medial boundary is also preceded by a phrase command adjusting the declination line and a high boundary tone connected to an accent command.

If we distinguish inter-sentence from intra-sentence boundaries we find that all inter-sentence-boundaries are aligned with the onset of a phrase command. 68% of all intra-sentence boundaries exhibit a phrase command, with the figure rising to 71% for 'comma-boundaries'. The mean phrase command magnitude for intra-sentence boundaries, inter-sentence-boundaries and paragraph onsets amounts to 0.8, 1.68 , and 2.28 respectively, which shows that *Ap* is a good indicator for boundary strength.

As we have shown on the corpus of news readings a rather consistent mapping between ToBI accent labels and accent commands can be achieved if we relate accent command timing to the syllabic boundaries. This seems logical since ToBI labels are assigned by the labeler in a similar fashion. As far as boundaries are concerned, only BI4 boundaries reliably coincide with phrase commands whereas lower level boundaries mainly use durational cues which can, however, be derived from the duration contour. A complete description of boundaries should not only refer to something like a perceived disjuncture, but also to the prosodic means employed (pause, reset of declination line as indicated by a phrase command, durational).

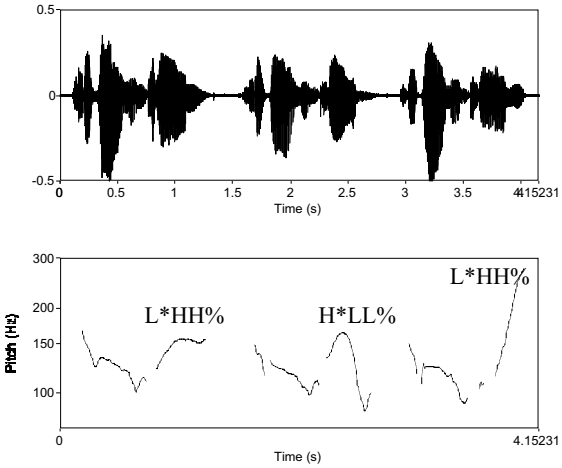

*Figure 2: Utterances of the sentence "Sie wollen ihn sehen"- "They wish to see him", uttered with non-terminal, declarative and interrogative intonation (from the left to the right) with associated ToBI labels. As can be seen, although different in shape non-terminal and interrogative utterance can be labeled using a sequence of L*HH% of the Stuttgart system.*

4.5. A Recent Consensus Version of G-ToBI

Accent labels assigned on the news corpus fall mainly into two classes, namely non-terminal (L*H) and declarative (H*L). These correspond to the N↑intoneme and I↓ intoneme in the tone switch approach. There is, however, no boundary tone label in the Stuttgart System differentiating between the different height of *F0* offset in non-terminal and interrogative

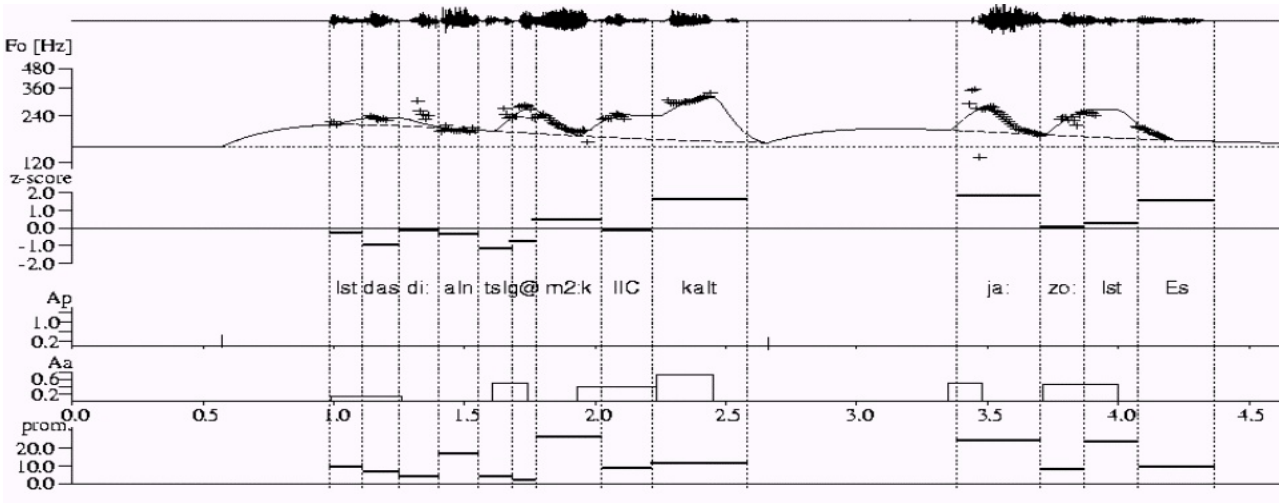

Figure 3: Example of analysis from the database. Utterance: "Ist das die einzige Möglichkeit? - Ja, so ist es."-"Is this the only possibility? - Yes, it is." From top to bottom: speech waveform, extracted and model-generated F0 contours, duration contour (syllabic z-score), SAMPA transcription, underlying phrase and accent commands and median perceived prominence.

intonation as represented by the C↑intoneme (see Figure 2). This suggests that even for this distinction we might need to use a quantitative parameter like the accent command amplitude *Aa* of the Fujisaki model.

In a more recent formulation of G-ToBI by Grice, Baumann and Benzmüller which is meant to present a consensus integrating earlier variants of G-ToBI, this problem is 'fixed' by adding a high ^H% boundary tone marker for very high offsets [6] (for quick reference, see http://www.coli.uni-sb.de/phonetik/projects/Tobi/gtobi.html).

Implicitly this offends the original rationale of ToBI using only two levels of tonal representation, namely L and H, but it helps account for an observed phenomenon. Although Grice, Baumann and Benzmüller advertise their G-ToBI system as being phonologically motivated, the claim is loosened. Particularly interesting is the authors' attempt to account for eleven different combinations of nuclear accent, phrase accent and boundary tone which appear to be phonologically relevant with respect to earlier and often impressionistic studies on German intonation. Contour types corresponding to strictly linguistic functions (declarative, interrogative, unfinished) are listed together with those ascribed paralinguistic functions ('indignation', 'self-evident assertion', 'polite offer') in an undifferentiated manner. This criticism was already made by Isacenko & Schädlich [12] when they reviewed the impressionistic literature on intonation of their time. In some cases the same contour type is attributed completely different functions ('indignation' vs. 'answering phone', for instance). The notion of 'phrase accents' describing the portion of the *F0* contour between the nuclear accent and the boundary tone appears especially problematic, since German is a free accent language, and as a consequence, there will be no need for a phrase accent if the ultimate or penultimate bears the nuclear accent. The large number of possible tone labels (H*, L*, L+H*, L*+H, H+L*, H+!H*) suggests that the system is rather designed for offering a close phonetic transcription of observed F0 contours than a phonological description as claimed.

One major drawback of ToBI systems in general is that they disregard the fact that *F0* contours are produced by a bio-mechanical system with an inherent latency. A particular contour therefore also results from the proximity or distance of intonational events in time which influences the *F0* contour associated with an accented syllable. Especially with respect to speech recognition a possibly overspecified ToBI system as the one discussed above appears problematic.

Just in order to illustrate that we might further run into trouble using purely symbolic representations when we attempt to model para-linguistic functions, *Figure 4* shows two versions of the sentence 'Du hast ja gelogen'-'You have

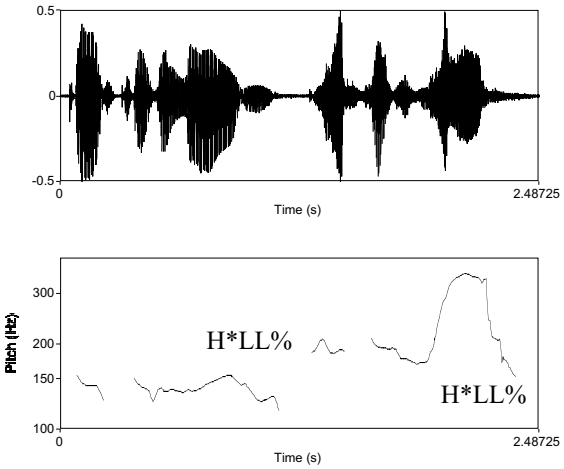

*Figure 4: Utterances of the sentence "Du hast ja gelogen" - "You have been lying", uttered with connotations of disappointment (left) and anger (right) with associated ToBI labels. As can be seen, despite the very different shapes in principle both utterances can be labeled with H*LL% (Stuttgart system).*

been lying' uttered by the author with connotations of disappointment (left) and anger (right). Although the pitch tracks look very different, they can be symbolized by the same sequence of ToBI labels. More moderate goals for

prosody recognizer than the recognition of para-linguistic information could be seen in the following:

- the detection of accented syllables by matching prosodic parameters, segmental timing and word information as to lexical accent syllable
- the classification of accent types with respect to sentence mode (declarative, interrogative, unfinished) by utilizing the timing information of tone switches
- the quantification of accents with respect to their neighbors in a stretch of speech.

As we will show in the following, the accent command amplitude *Aa* is a reliable correlate of syllabic prominence. Furthermore, not all excursions of the *F0* contours are equally relevant.

5. Prominence and Prosodic Features

The perceived prominence of syllables can be regarded as a gradual parameter suited for describing the emphasis assigned to linguistic units in relation to their environment and with respect to the meaning of an utterance. In a recent study [17], the relationship between the perceived prominence of a syllable and two important prosodic features assigned to the syllable was examined. These features are (1) the interval of a major *F0* transition connected to the syllable[1], as expressed by the accent command amplitude *Aa*, (2) normalized log syllable durations.

5.1. Prominence of syllables

The notion of prominence followed is based on [18]. Three labelers had to judge the degree of prominence on the syllable level relative to the surrounding syllables on a scale from 0 to 31. Between subjects, the labeled prominences correlate strongly (*rho* > 0.8; [19]). Earlier investigations showed that the relation between prominence ratings and syllable duration, as well as *F0* peaks, described by parameters of a maximum based description of *F0* contours [5], are linear. However, prominence is also related to linguistic features (i.e. word class, position in a phrase, and focus). Thus perceived prominence can be regarded as a gradual parameter integrating linguistic features and acoustic parameters.

Since the Fujisaki model is inherently **production**-based, one major issue in this study is to establish the relationship between the amplitude parameter *Aa* and the **perceived** prominence of a syllable. Furthermore the implicit claim underlying IGM that not all parts of the *F0* contour are 'equally important' is investigated. If the claim is tenable, linguistically motivated *F0* transitions, i.e. tone switches, should strongly contribute to the perceived prominence of a syllable, whereas the so-called 'pitch-interrupters' [12], *F0* transitions at non-accent syllables, should not.

5.2. Speech Material and Method of Analysis

The speech material was taken from the Bonn Prosodic Database (BPD, [20]) of read speech. The subcorpus chosen is composed of isolated sentences, question-answer pairs, and short stories of one female speaker, and contains a total of 3401 syllables. Every syllable is assigned information about its position and its number in higher-level units (i.e. position of syllable in a word or in a prosodic phrase), its nucleus, as well as the number of phones it consists of. The syllables are annotated with respect to word class and lexical word stress, as well as their prominence scaled from 0 to 31, as judged by three phoneticians. The prominence of a syllable is taken to be the median of the judgments.

Log syllable durations were computed from phone labels in the BPD and normalized to their syllable count and the property of the nuclear vowel, being either schwa or non-schwa, the most important intrinsic features as shown in [9].

5.3. Results of Analysis

Figure 1 shows an example of analysis from the database displaying the utterance "Ist das die einzige Möglichkeit? - Ja, so ist es."-*"Is this the only possibility? - Yes, it is."* The figure displays from top to bottom: the speech waveform, the extracted and model-generated *F0* contours, the duration contour in terms of the syllabic z-score drawn as horizontal lines of the length of the respective syllable, the SAMPA transcription of the utterance, the underlying phrase and accent commands and the median perceived prominence. It can be seen that syllables with the highest prominence are accented syllables connected to tone switches (rising on [Ist], [aIn] and [m2:k], falling on [ja:] and [Ist]). High prominence is assigned to an accented syllable, even if the *F0* movement starts late in the syllable as in [m2:k] or in the following syllable as in [aIN].

Pre-boundary syllables, such as [kaIt] and [Es] exhibit relatively long durations compared with accented or unaccented syllables. [ja:] is a case of a syllable that is both accented and in a pre-boundary location, showing high prominence, high *Aa* as well as long duration.

5.4. Perceived Prominence and Acoustic Parameters

Perceived prominences are evaluated in relation to the acoustic parameters *Aa* and normalized log syllable duration (*nsyldur*). The correlation over all syllables (rang correlation coefficient *rho*) is about 0.5 for *Aa* and about 0.4 for *nsyldur*. These relatively low values may be explained by other influences such as phrase-final lengthening and boundary tones. If we only include syllables with lexical word accent, the correlation between prominence values and *Aa* (*rho*= 0.6) as well as *nsyldur* (*rho* = 0.5) increases. Analysis shows that *Aa* is mostly related to higher prominence values (>15). In contrast to *Aa*, *nsyldur* correlates more strongly with lower prominence values (<16). Hence, weak perceived prominence gradings are associated with durational cues and strongly perceived prominence is mostly related to *F0* movements. The relationship between perceived prominence and the two acoustic parameters can be regarded as nearly linear.

5.5. Perceived Prominence and Tone Switches

We examined whether the linguistic notion of tone switches is reflected by prominence values. Prominence values and acoustic parameters of the linguistically motivated tone switches (I↓ intonemes and N↑ intonemes) are compared with the values of non-linguistic *F0* movements, i.e. rising and falling pitch interrupters. The comparison of falling

[1] i.e. a rise and/or fall during the syllable proper or in the preceding or following one.

pitch interrupters with I↓ intonemes shows that I↓ intonemes more strongly contribute to prominence than falling pitch interrupters. Comparable results are also found for the N↑ intoneme and rising pitch interrupters. Furthermore the results show that the average prominence value of N↑ intonemes is lower than those of the information intonemes.

Our results show that, for accented syllables, prominences strongly correlate with the amplitude *Aa* of accent commands underlying the *F0* movements in these syllables, whereas comparable *F0* movements in unaccented syllables have little effect on prominence. The correlation between perceived prominence and *Aa* is significantly stronger than with respect to the *F0*-maximum-based parameter used by Heuft [20] (*rho*=0.2-0.3) which can only be reliably determined for accents with clear *F0* peaks.

However we must bear in mind that the accent command amplitude parameter *Aa* of the production-based Fujisaki model is a very strong correlate of perceived prominence *wherever F0 movements can be motivated linguistically*. We may tentatively interpret this relationship as follows: While *Aa* - inter alia - reflects the 'relative importance' of accented constituent words in an utterance as intended by the speaker, prominence reflects the 'realized performance structure' of the utterance as perceived by the listener.

6. Discussion and Conclusions

In this article we showed the parallels between a ToBI representation and a syllable-based quantitative description of prosody and showed the potential of deriving a ToBI-style representation from syllable-related Fujisaki parameters which can be useful for a purely linguistic annotation of large data bases with a reduced set of ToBI labels.

We argue that a recent G-ToBI development with a largely extended set of labels is more likely to account for phonetic variation in the *F0* contour than to be phonologically justified. As an alternative we suggest to make use of the quantitative information present in prosodic features. This could be done on the basis of any quantitative formulation yielding timing and amplitude information of intonational events, but obviously a physiologically motivated approach is much preferred as it gives certain properties (latency, *F0* declination, ...) for free.

After all, meaning is not created by a ToBI accent here and there, but is the result of a bundle of prosodic features spread across an utterance teaming up with syntax and wording. We cannot even be sure that it can really be captured purely symbolically. If perceived prominence is a cue to highlighted information, then capturing prominence with a qunatitative model might get us closer to...making sense.

7. References

[1] Mayer, J., 1995. Transcription of German Intonation: The Stuttgart System. Technischer Bericht, Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung.

[2] Grice, M.; Benzmüller, 1995. Transcription of German Intonation using ToBI-Tones – The Saarbrücken System. Saarbrücken: Phonus 1. 33-51.

[3] Fujisaki, H.; Hirose, K., 1984. Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *Journal of the Acoustical Society of Japan (E)*, 5(4), 233-241.

[4] Taylor, P., 1998. The Tilt intonation model. *Proc. ICSLP 98*, vol. 4, 1383-1386.

[5] Heuft, B.; Portele, T.; Höfer, F.; Krämer, J.; Meyer, H.; Rauth, M.; Sonntag, G., 1995. Parametric description of *F0* contours in a prosodic database. Proceedings of the ICPhS'95, Stockholm KTH, 378-381.

[6] Grice, Martine, Stefan Baumann & Ralf Benzmüller (to appear). German Intonation in Autosegmental-Metrical Phonology. In Jun, Sun-Ah (ed.) Prosodic Typology, Oxford University Press.

[7] Mixdorff, H.; Vainio, M.; Werner, S.; Järvikivi, J., 2002. The Manifestation of Linguistic Information in Prosodic Features of Finnish. To be presented at *Speech Prosody 2002*, Aix, France.

[8] Mixdorff, H.; Jokisch, O., 2001. Building An Integrated Prosodic Model of German. In *Proceedings of Eurospeech 2001*, vol. 2, Aalborg, Denmark, 947-950.

[9] Mixdorff, H.; Fujisaki, H., 2000. A quantitative description of German prosody offering symbolic labels as a by-product. *Proc.ICSLP 2000,* vol. 2, Beijing, China, 98-101.

[10] Rapp, S. Automatisierte Erstellung von Korpora für die Prosodieforschung. PhD thesis Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung. 1998.

[11] Pierrehumbert, J., 1980. The phonology and phonetics of English intonation. Ph.D thesis. MIT.

[12] A.V. Isacenko; Schädlich, H.J., 1964. *Untersuchungen über die deutsche Satzintonation*. Berlin: Akademie-Verlag,.

[13] E. Stock; Zacharias, C., 1982. *Deutsche Satzintonation*. Leipzig: VEB Verlag Enzyklopädie.

[14] Mixdorff, H., 1998. *Intonation Patterns of German - Model-based. Quantitative Analysis and Synthesis of F0-Contours.* PdD thesis TU Dresden, (http://www.tfh-berlin.de/~mixdorff/thesis.htm).

[15] Fujisaki, H., 1998. A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour. In Fujimura, O. (ed.) *Vocal Physiology: Voice Production, Mechanisms and Functions*. New York: Raven Press, 347-355.

[16] Mixdorff, H., 2000. A novel approach to the fully automatic extraction of Fujisaki model parameters. *Proceedings ICASSP 2000*, vol. 3, Istanbul, Turkey, 1281-1284.

[17] Mixdorff, H.; Widera, C., 2001. Perceived Prominence in Terms of a Linguistically Motivated Quantitative Intonation Model. In *Proceedings of Eurospeech 2001*, vol. 1, Aalborg, Denmark, 403-406.

[18] Fant, G.; Kruckenberg, A., 1989. Preliminaries to the study of Swedish prose reading and reading style. *Speech Transmission Laboratory – Quarterly Progress and Status Report, KTH Sockholm*, 2:1-83.

[19] B. Heuft; Portele, T., 1996. Synthesizing prosody: A prominence-based approach. *Proceedings ICSLP'96*, Philadelphia, 1361-1364.

[20] Heuft, B., 1999. Eine prominenzbasierte Methode zur Prosodieanalyse und –synthese. W. Hess and W. Lenders (eds.): Computer Studies in Language and Speech, Vol. 2. Frankfurt am Main.: Peter Lang.