Form versus Function – Prosodic Annotation and Modeling go Hand in Hand

Hansjörg Mixdorff

Department of Computer Science and Media, Beuth University of Applied Science

mixdorff@beuth-hochschule.de

Abstract

This paper argues that prosodic annotation and modeling should be combined for facilitating analyses of prosodic functions that invariably require perceptual judgments. It compares perceptual prosodic annotations of prominent syllables and phrase boundaries with labels yielded by the combination of linguistic information from a TTS-front end, model-based prosodic features, as well as a model of perceived syllabic prominence from an earlier study.

As can be expected this annotation of prosodic landmarks yields better results on reading style speech than on spontaneous speech data. Of the perceptual annotations, on average 89% of perceptually prominent syllables were identified correctly, as well as a similar percentage of prosodic boundaries. Hence a basic annotation of prosodic features is yielded which can later on be enhanced by additional information for which perceptual judgments are indispensable.

Index Terms: Prosodic annotation, prosodic modeling, Fujisaki model, perceptual prominence

1. Introduction

Prosodic annotation is usually faced with many conflicting requirements and so far there is no coding scheme which provides a "one fits all". ToBI [1] has been rather successful as a quasi-standard, but is still caught up somewhere half-way between phonological relevance and phonetic realization.

As we move from marking pure linguistic functions towards connotations of affect, for instance, prosodic coding schemes are often crafted with a specific purpose or application in mind. Furthermore, different from segmental annotations, prosodic functions are often not realized in a single place, i.e. a syllable, but may affect the utterance as a whole. Increasingly we develop prosodic annotations with a computational model in mind, that is, we define categories, for which the model will predict their prosodic realizations. The paradox that we are confronted with is that form and function often appear entangled in these realizations, and the hope that a certain tonal configuration is unanimously connected with a certain function seems futile (see, for instance, Daniel Hirst's discussion in [2]).

Another question is why so much of prosodic annotation is still performed by hand, making it incredibly time-consuming. Over the years we have seen automatic algorithms developing that are capable of directly identifying in the speech signal prominent syllables as well as phrase boundaries, two types of prosodic landmarks that are the basis for many analyses that we are interested in (see, for instance, [3]). Still we rely on human annotators to painstakingly perform these tasks, not necessarily with a better reliability than automatic methods.

The approach developed by Isačenko and Schädlich [4] and Stock and Zacharias [5] describes a given *F0* contour as a sequence of communicatively motivated tone switches, major transitions of the *F0* contour aligned with accented syllables. With respect to the form-function relationship they distinguished between three main classes of tone switches: *falls* associated in declarative utterances signaling finality, *rises* to a mid-level, signaling continuation and *rises* to a high level, for establishing contact. Although one can argue that these distinctions are rather coarse, the perception experiments that their theory was based on showed a rather good agreement in the judgment on linguistic distinctions.

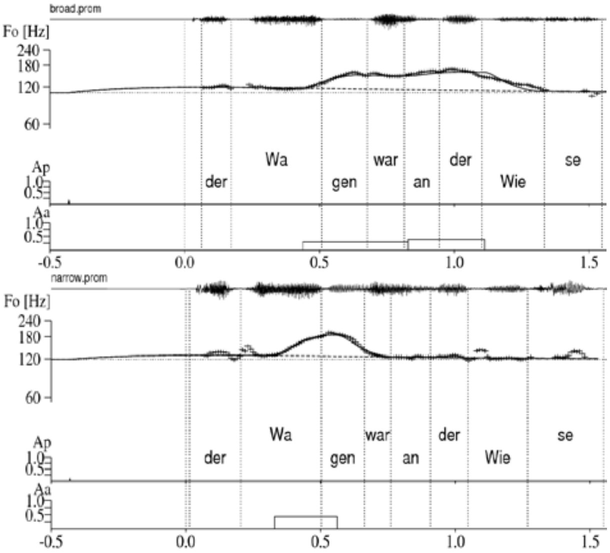

Figure 1: *Two examples of the German sentence "Der Wagen war an der Wiese."-"The car was at the meadow" with broad focus (top) and narrow focus on the word "Wagen" (bottom). The focus distinction has several effects on the F0 contour of the utterance (see text) and the resulting Fujisaki model parameters. The panel displays from the top to the bottom: the speech waveform, the F0 contour (extracted and modeled), as well as the underlying phrase and tone commands.*

In order to quantify the interval and timing of the tone switches with respect to the underlying syllables, the author adopted the Fujisaki model [5] which reproduces *F0* from three components: Base frequency *Fb*, phrase component and accent component. The Fujisaki model parameters are typically extracted from a natural observed F0 contour without applying linguistic knowledge. This means, that a set of labels - typically syllabic - is required before the Fujisaki model parameters can be aligned and interpreteted. In his PhD works the author compared examples of the same sentence in broad and narrow focus conditions. It became apparent that narrow focus not only boosted the focused items – that is, expanded the tone switch associated with the focus exponent- but also deleted or at least reduced the tone switches on competing content words. Furthermore, the accent command alignment changed (see Figure 1 for an example). Now how could one account for all these differences by a segmentally motivated annotion scheme? One would have to mark the (boosted) focus exponent “Wagen” as well as the (dimmed) constituent “Wiese” and probably the change in the alignment as well since the contour between the two words changes drastically. The declarative mode of the utterance is marked by a falling tone switch, which occurs in the word “Wagen” for narrow and in the word “Wiese” for the broad focus case. This shows that prosodic functions are not necessarily coded in a single place or at a single level of description, but may influence an utterance in various ways.

The author’s intention is to explore to what extent an automatic acoustic and linguistic mark-up combined with a quantitative *F0* model could be used as a first step for a prosodic annotation scheme which subsequently concentrates on the features relevant to changes in meaning for which ultimately auditory judgements are indispensable.

2. Predicting Perceived Syllabic Prominence

Yielding information on prominent syllables and phrase boundaries is a typical baseline of prosodic annotation. In an earlier study [8], the author and a co-worker investigated the relationship between perceived syllable prominence and the *F0* contour in terms of the parameters of the Fujisaki model. A subcorpus of the Bonn Prosodic Database [9] was parameterized using the model, and normalized log syllable durations were calculated. Analysis showed that, for accented syllables, prominences labeled on a scale from 0-31 by three human labellers strongly correlated with the amplitude *Aa* of accent commands underlying the *F0* movements in these syllables, whereas comparable *F0* movements in unaccented syllables had only little effect on prominence. The influence of *Aa* versus syllable duration on prominence was hence greater in higher prominence classes.

The fact that the prominence-lending *F0* movement does not necessarily take place inside the accented syllable indicated that the prominence judgment is partly guided by linguistic considerations. Building on the results of this earlier study we revisited the corpus and calculated a regression model of perceived prominence taking into account the factors lexical stress, *F0* interval as expressed by *Aa*, syllable duration z-score and vowel type (being either schwa or non-schwa, as well as open/closed). The predictions of this model correlate with the averaged human judgments at r=.79 (Pearson’s r, p < .01) which is only slightly lower than the inter-labeller correlation of r=.80.

3. Speech Material and Method of Analysis

In the context of a recent prosodic study (presented at this conference) comparing German and Brazilian Portuguese [10] a corpus of read and spontaneous utterances was segmented using forced alignment on the syllabic level. The data was then prosodically annotated on a perceptual level with respect to prominent syllables, as well as phrase boundaries and the sentence mode signaled at these locations.

Two female and seven male speakers of Computer Science read a 1,500-word text about the pastries “Pastéis de Belém” [11]. Later, subjects retold the story (spontaneous narrative). *F0* values were extracted using the standard method in *Praat* [12] at a step size of 10 ms and inspected for errors. The *F0* tracks were subsequently decomposed using the standard automatic method [14] and if necessary corrected using the *FujiParaEditor*. A TTS front-end was applied to the texts of the utterances predicting stressed and unstressed syllables [15], as well as phrase boundaries and their underlying sentence mode. This information was then combined with the Fujisaki model-based representation of *F0* and syllabic durations to calculate a measure of prominence as well as sentence boundary strength and type using the regression model developed on the Bonn Prosodic Database.

Two labellers marked the perceptually prominent syllables. The analyses shown here concern excerpts of 150 to 200 words from each speaker and style.

A computer program was developed which associates the accent commands from the automatic estimation with lexically stressed syllables as predicted by the TTS front-end. In a first search each syllable which exhibited onsets or offsets of accent commands was labeled accordingly. Then it was checked whether the current syllable was predicted as lexically stressed and several alignment options evaluated. Depending on the situation found, the tone switch associated with the syllable was classified as either rising or falling. The search takes into account the current marked syllable and its immediate left and right neighbours.

4. Analysis Results

Figure 2 shows an example of Fujisaki model-based analysis. The panel displays from the top to the bottom: the speech waveform, the *F0* contour (extracted and modeled), as well as the underlying phrase and tone commands. The syllable boundaries are indicated by the dotted vertical lines. At the bottom the syllabic prominence values predicted by the regression model are indicated. The text and the English translation of the utterance are given in the caption.

Hence a basic annotation of prosodic features was yielded which can later on be augmented by additional information. As can be expected this approach yields better results on reading style speech than on the spontaneous speech data. For instance, for the reading task 90.4% of syllables were classified correctly as prominent/non-prominent when the perceptual measures from the regression model were split at a value of 13, that is, when all syllables with a score greater than 13 were classified as prominent. As expected, the proportion was slightly lower for spontaneous speech (86.8%). 64.7% of the phrase boundaries realized had been predicted by the TTS front-end in the reading style version whereas the proportion was only 44.1% for spontaneous speech. However, since a large number of phrase boundaries is signaled by speech pauses (duration > 100 ms), many can be recovered even when they do not occur where the TTS front-end predicts them (23.5% for reading style and 47.1% for spontaneous speech).

5. Discussion and Conclusions

The current paper argues that prosodically marked differences in meaning, i.e. functions, are typically complex and therefore annotation of these requires time-consuming human judgments. However, this process can be facilitated by (semi-)

automatic procedures for marking up the speech signal based on prosodic modeling, information on human prominence judgments, combined with a linguistic analysis of the underlying text. Taking linguistic information into account is important as prosodic judgments by humans are guided by their underlying language representation.

The approach presented yields information on prosodic land-marks (i.e. forms), such as prominent syllables and phrase boundaries. In the current study about 89% of these could be identified correctly, though the performance was poorer on spontaneous speech compared with read speech. In future work, we plan to enhance the approach by applying more robust statistical methods and integrating prosodic features in the process of forced alignment proper.

References

[1] Pierrehumbert, J., "The Phonology and Phonetics of English Intonation", PhD thesis, MIT, Cambridge, MA.E, 1980.

[2] Hirst, D.: "Form and function in the representation of speech prosody", Speech Communication 46(3-4): 334-347 (2005)

[3] 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. of IEEE Transactions on Speech and Audio Processing, 8(5):519–532, 2000.

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

[5] Stock E., Zacharias, C., "Deutsche Satzintonation", VEB Verlag Enzyklopädie, Leipzig, 1982.

[6] Fujisaki, H. and Hirose, K. "Analysis of voice fundamental frequency contours for declarative sentences of Japanese", J. of the Acoustical Society of Japan (E) 5(4), 233-241, 1984.

[7] Mixdorff, H., "Intonation Patterns of German - Model-based Quantitative Analysis and Synthesis of F_0-Contours", PhD thesis submitted to TU Dresden, 1998.

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

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

[10] Mixdorff, H. and Barbosa, P.A. "Alignment of Intonational Events in German and Brazilian Portuguese – a Quantitative Study. Proceedings of SpeechProsody 2012, Shanghai, China, 2012.

[11] Barbosa, P., Mixdorff, H. and Madureira, S., "Applying the quantitative target approximation model (qTA) to German and Brazilian Portuguese", in Proceedings of Interspeech 2011, Florence, Italy.

[12] Boersma, Paul. "Praat, a system for doing phonetics by computer". *Glot International* **5:9/10**, 341-345, 2001.

[13] Mixdorff, H., "A novel approach to the fully automatic extraction of Fujisaki model parameters", in Proc. of ICASSP, Istanbul, 3:1281–1284, 2000.

[14] Mixdorff, H. (1/10/2009). *FujiParaEditor*, http://public.beuth-hochschule.de/~mixdorff/thesis/fujisaki.html

[15] Hilbert, A., and Mixdorff, H.,"Weiterentwicklung eines Sprachsynthesesystems", in G. Görlitz [Ed.], Nachhaltige Forschung in Wachstumsbereichen Band I, Logos Verlag, Berlin, 35-42, 2011.

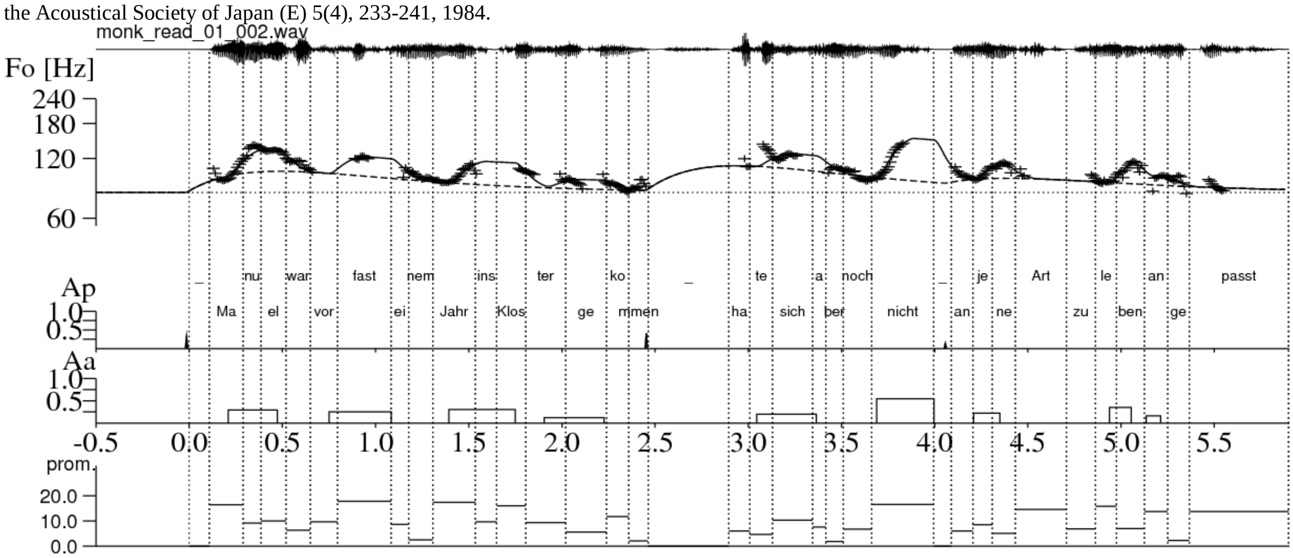

Figure 2: *Example of analysis. Male speaker 1, reading style (perceptually prominent syllables set in bold face), "**Ma**nuel war vor **fast** einem **Jahr** ins **Klo**ster gekommen, hatte sich aber noch **nicht** an jene **Art** zu **le**ben **an**gepasst"-"Manuel had come to the monastery almost one year ago, but had not yet adapted to that way of living." The predicted syllabic prominence values on a scale from 0 to 31 are reflected by the horizontal lines the length of the syllables on the bottom tier (prom.)*