ProsAlign - The Automatic Prosodic Aligner

Norbert Braunschweiler*

* Institute of Natural Language Processing (IMS)
Azenbergstr.12, 70174 Stuttgart, Germany
norbert.braunschweiler@ims.uni-stuttgart.de

ABSTRACT

A computer program (ProsAlign) is presented that detects prosodic events in a given speech file. The prosodic events are pitch accents and boundary tones according to an underlying phonological model of intonational structure (ToBI). Therefore the program formulates an explicit way of how to map an acoustic representation of prosody up to an abstract, discrete representation of its underlying prosodic structure as represented by phonological tones. Both bottom-up and top-down processing steps are integrated. Concept and performance of the program is reported in this paper.

1 INTRODUCTION

The process of information extraction from acoustic speech signals involves not only the recognition of segmental features, phonemes, syllables or words and subsequent processing, but also the recognition of prosodic events, including the position of accented words, the type of pitch movement associated with them, the general course of pitch and also the grouping of information units, phrases or words. These prosodic events are important conveyors of the information structure of utterances. In this paper an approach is presented that formulates an explicit way of how to map from continuous acoustic parameters to discrete and abstract phonological entities. The method is implemented in a computer program called ProsAlign (Automatic Prosodic Aligner) and uses a linguistic theory of the underlying structure of prosody in speech (ToBI model) [1].

ProsAlign is intended to serve as a tool for the linguist who wants to work with prosodically labeled speech material. Such material may be helpful for basic research purposes as well as for improving speech syntheses techniques (especially for unit selection approaches) or speech recognition performance.

This paper is structured as follows: first, the concept and development of a new approach to the automatic detection of prosodic cues is described; second, the implementation of the program is explained; and third, its evaluation is presented.

2 THE CONCEPT OF PROSALIGN

The concept of the program integrates both bottom-up and top-down processing in order to account for the vast variability in acoustic data. Bottom-up processing is intended to reflect the ability to discriminate differences in the acoustic input parameters represented by fundamental frequency (F_0), and root mean squared amplitude (RMS), and voicing. F_0 represents the frequency of oscillations of the vocal folds and is the most important acoustic correlate of perceived pitch. RMS amplitude is a rough estimate of perceived loudness. Top-down processing is inspired by the structuring influence of phonological knowledge upon acoustic features. Figure 1 depicts the taskflow in ProsAlign.

The technique maps acoustic feature bundles to prosodic labels by inspecting the course of F_0 and RMS combined with a subsequent phonological filter. The system works without any pre-segmentation of the speech waveform and does not use HMM techniques. The program steps linearly through the F_0 track and applies relative comparisons only within a defined analysis window. Furthermore, the program is designed to process speech coming from different speakers, i.e. it should be speaker independent as well as language independent and fairly robust against different recording conditions.

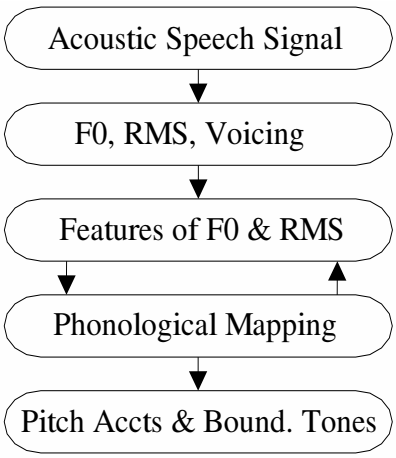

Figure 1: *Outline of the model underlying ProsAlign.*

ISBN 1-876346-48-5

3 THE IMPLEMENTATION OF PROSALIGN

3.1 Acoustic features

Since pitch accents are usually described only qualitatively in phonological textbooks (see e.g. the definition of pitch accent in [2], p.45-46), an analysis of acoustic features of pitch accents and boundary tones was conducted. To form initially an idea about types and number of acoustic features needed for an automatic detection of tones, a manually labeled corpus was taken as a starting point. First experiments were carried out on German speech material, therefore, the German adaptation of the ToBI model, GToBI and its accompanying [3] training corpus was used. This corpus was chosen because it has the advantages of providing a reasonable number of examples for each tone postulated in the underlying phonological model as well as the acoustic material along with the prosodic label files. The inventory of pitch accents and boundary tones in GToBI is as follows: 6 pitch accents (H*, L+H*, H+!H*, L*, L*+H, H+L*), 2 phrase accents (L-, H-), and 2 boundary tones (L%, H%).

A program was designed in order to acquire quantitative acoustic criteria of each tone. As a starting point, and because there where no obvious answers to these questions, it was decided to analyze the manually labeled pitch accents and boundary tones in the GToBI corpus with respect to the following criteria:

(i) duration of increasing or decreasing parts of F_0 and RMS before and after tone location (F0in/deB/A and RMSin/deB/A)

(ii) size of F_0 and RMS increase or decrease before and after tone location (AF0in/deB/A and ARMin/deB/A)

(iii) duration of voiced or voiceless parts before or after tone location (Voic/VoilB/A)

The resulting 20 parameters are illustrated in figure 2. Idealized F_0 and RMS tracks are used to show the value of each parameter. The parameters are presented below the chart. Point t_0 represents the position of pitch accents or boundary tones and individual parameter values are always calculated relative to this starting point as depicted.

3.2 RESULTS

Each manually labeled tone was analyzed with respect to the above-mentioned parameters within an interval of ± 400 ms around its labeled position. The decision for this window was based on a first inspection of pitch accents and boundary tones and seemed to be a reasonable analysis frame for covering enough contextual material for the selection of acoustic features.

Since the parameter values are given every 10 ms, always the number of values (frames) starting from point t_0 are calculated that represents the location of a pitch

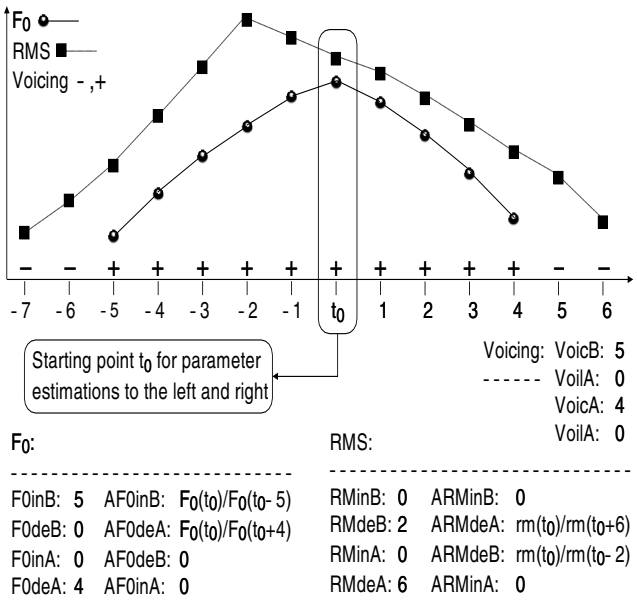

Figure 2: *Illustration of 20 acoustic parameters chosen for the analysis of pitch accents and boundary tones. F_0 and RMS tracks are idealized. Point t_0 represents the location of pitch accents and boundary tones and all parameter values are calculated relative to it.*

accent or boundary tone. A segment of the results is illustrated in table 1.

The results of the parameter analysis showed that the chosen criteria only partly captured the acoustic differences between individual pitch accents and boundary tones. In particular, estimations of increases and decreases in the course of F_0 were unsatisfactory.

When estimating F_0 movements for automatic detection purposes it is very important to provide means for reliable estimations. Otherwise the basic question whether or not a given movement is perceptually important cannot be reliably answered. At this point it is essential to take microprosodic effects into account as well as knowledge about possible errors in the automatically generated F_0 track. Microprosodic influences result from perturbations of the glottal signal at transitions between voiced and unvoiced segments and vice versa. F_0 values are wrong as a result of erroneous estimations about possible oscillation periods in the waveform often as consequence of laryngealizations or poor signal quality. Pitch doubling or halving errors are typical examples [4].

As a consequence of these drawbacks additional parameters were introduced. Altogether 74 acoustic parameters are used in ProsAlign, and estimations of faulty as well as microprosodically affected F_0 values form an integrated part of the algorithm. One of the parameters that allowed a much better estimation of increases and decreases in F_0 was a criterion that allows a limited number of outliers from a general increase or decrease. Visual control of the algorithms performance confirmed its efficiency. For more information on these parameters see [5].

ISBN 1-876346-48-5

Tone	H* (51)		L+H* (25)		L* (11)	
	Md	SD	Md	SD	Md	SD
F_0inB	3	(4)	8	(4)	1	(3)
F_0deB	0	(1)	0	(0)	0	(3)
F_0deA	0	(4)	0	(5)	0	(1)
F_0inA	1	(3)	1	(2)	9	(8)

Table 1: *Median values (Md) and standard deviations (SD) for selected acoustic features of H*, L+H*, and L* tones in the GToBI training material.*

3.3 IMPLEMENTATION

Input to the program consists of a structured file with values for F_0, voicing and RMS extracted in 10 ms steps from the input speech file. Currently the output of the ESPS/waves F_0 tracker *get_f0* is used but other algorithms may be used as well. The values of the 3 parameters are then processed by the acoustic feature detector, which extracts 74 acoustic features.

The resulting acoustic feature vectors are subsequently evaluated by a scoring system that assigns positive scores to feature values that have been previously identified as supporting the existence of a specific pitch accent or boundary tone, and negative scores for constellations that do not support them. The latter step integrates both expert knowledge as well as the results of the parameter analysis of manually labeled tones. Finally, the resulting score is evaluated by the phonological module that selects pitch accents and boundary tones based on the combination of sequence restrictions and scores (see taskflow in fig. 1). Therefore, a considerable amount of work within the algorithm is performed by phonological top-down processing. Hence both bottom-up and top-down processing are integrated in the ProsAlign program. The result is written to a label file that includes type and position of pitch accent or boundary tone and may be viewed time-aligned with the speech signal and F_0 track (see figure 3).

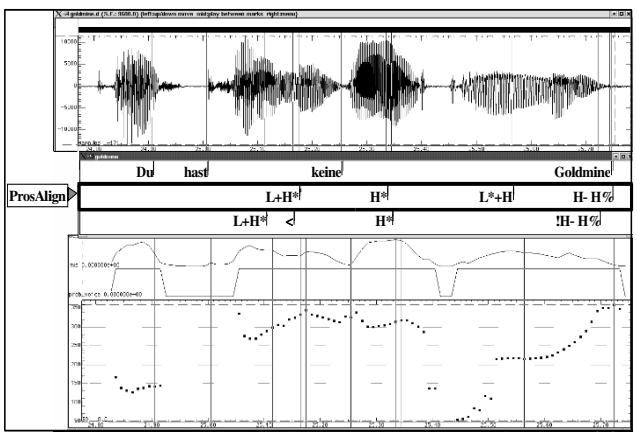

Figure 3: *Illustration of ProsAlign output. The manually produced label file is shown below the one produced by ProsAlign. Example "goldmine" taken from GToBI training material [3].*

4 EVALUATION

In order to get an estimation of the program's performance an evaluation was conducted. However, evaluating the output of an automatic prosodic aligner is a challenge. At first glance, the obvious evaluation method is to use a manually annotated corpus as a reference and to compare the automatically generated label files with the corresponding manually ones. However, there are a number of problematic aspects when doing so. One of the main issues here is that there are disagreements between different human labelers about the position and type of labels. In order to circumvent this problem it was decided to take the GToBI training corpus as basis. Although the corpus served as input for parameter estimations during the development of ProsAlign it should be possible to estimate the programs performance, especially the number of matches and insertions. Since the corpus is intended to serve as a training corpus, one can expect the examples to represent generally agreed cases of the individual pitch accents and boundary tones. However, it has to be mentioned that this corpus includes a diverse set of speech material consisting of utterances from different speakers, including several different speaking styles, unequal recording levels, background noises, and even cross-talk.

Table 2 shows a segment of the results of the evaluation. A total of 40 files was processed including 175 tone labels in the manually labeled set (column "man"), whereas the total number of automatically produced labels was 215 (column "aut"). The next 4 columns present the numbers of perfect matches ("perf"), partial matches ("part"), insertions ("insn"), and mismatches ("mism"), respectively. The last column shows the missing tones ("misg"). The last row shows the relative percentage of each column when taking the total number of automatically detected tones as 100% reference (except for the insertions, where the number of manually detected tones is the 100% reference).

	# tones		# tones from auto				
File	man	aut	perf	part	insn	mism	misg
august	3	5	2	0	3	0	0
blaue	2	2	2	0	0	0	0
dina4	6	6	2	2	2	0	2
...	...	...	...	...	...	...	...
Sum 40	175	215	92	27	84	12	44
% aut		100	43%	13%	39%	6%	25%

Table 2: *Segment of the evaluation results of ProsAlign for the GToBI corpus.*

When setting the number of automatically established labels as 100% reference level, the percentage of perfect matches is 43% and the number of partial matches is 13%. This means that 43 + 13 = 56% of the manually established labels are detected by the algorithm, although not all of them perfectly. The number of in-

ISBN 1-876346-48-5

sertions is 39% from the 215 automatically set labels, indicating that the algorithm labels much more intonational events than the human labelers do. However, this number as well as the number of missing tones (total number: 44 = 25% of all manually labeled tones) has to be evaluated with respect to the general problem of correspondency between different human labelers. It is an everyday experience in prosodic research that there are often large differences between human labelers when labeling one and the same speech material, especially for tonal identity. Transcribers often "agree very well on whether or not a word is prominent or whether or not a phrase boundary follows it, they often do not agree on the identity of the specific tone involved. Manually labeled speech corpora may not be sufficiently consistent for successful training or modeling for recognition or TTS systems." ([6], p. 4)

The number of mismatches that the program produces is very small. Only 12 labels out of 215 (6%) are mismatches. A closer look reveals that the main source of these mismatches results from boundary tone mismatches (8 out of 47) and not from pitch accent mismatches (4 out of 168). This indicates that the intonation phrase final F_0 movements are more often misleading probably as a result of final laryngealizations. Therefore, the procedure chosen to detect such cases still appears to be incomplete.

The results show that there is not a significant number of mismatches between human and automatic labels. ProsAlign covers at least 56% of the manually established labels. Important is the observation that when we look at the automatically produced labels and find that the labels are most often set at reasonable positions and very seldom in positions that do not make sense.

5 CONCLUSION

The evaluation of the ProsAlign algorithm showed its performance characteristics: high pitch accents are detected fairly well, whereas low pitch accents and especially high boundary tones are detected less reliably. L-L% boundary tones are detected fairly well, although there is still room for improvement.

Despite the fact that ProsAlign's ability to detect pitch accents and boundary tones does not reach the performance of human labelers it is nevertheless useable and a valuable tool for speech processing. The visual inspection of the automatically produced labels shows generally good accuracy. Automatically produced pitch accent labeles are most often at positions that are interpretable as possible positions of pitch accents, the same holds for positions of boundary tone labels. Whenever there is a need for more refined results the algorithm may be used in a semi-automatic way by applying manual corrections after the program has been processed the data.

However, a number of mismatches in the boundary tone class indicates that there are still problems with respect to the detection of erroneous F_0 values at intonation phrase final positions. This point is certainly in need of improvement. Problems with the decision between low or high boundary tones could be circumvented by using just '%' in ambiguous cases and using the concrete labels only in more obvious cases. Nevertheless, the overall performance qualifies the program as a usable automatic prosodic labeler for linguistic work as well as an interesting research tool for studying acoustic features of (phonological) pitch accents. Moreover, the method models in detail the way from acoustic speech signals up to perception. Poor detection results for certain tone categories also indicate that there could be arguments for a reduction of the number of pitch accents and boundary tones in the underlying phonological model (e.g. H+L*).

Initial experiments on American English ToBI corpus indicate similar results as for the GToBI corpus. The American English ToBI system does not include the H+L* and therefore the set of pitch accents was adapted in ProsAlign. The same procedure should be able to handle other languages, i.e. adaptation of the underlying phonological tones but no adaptation of the acoustic feature detector. Whether the last hypotheses holds has to be shown on further speech material from other languages. An indication for language or even speaker specific acoustic cues comes from bad recognition results for low pitch accents and a number of missing boundary tones in the American English corpus.

In conclusion ProsAlign showed encouraging results for very diverse speech material consisting of utterances from different languages, different speakers (male and female), and a number of other changing conditions. The robustness of ProsAlign against this changing conditions shows the power of its underlying method. Important aspects of the work include integration and evaluation of linguistic theory and quantitative acoustic modeling.

REFERENCES

[1] Beckman, M. E. and Ayers, G. M., "Guidelines for ToBI Labelling", ftp://www.ling.ohio-state.edu/pub/TOBI-/DOCS/labelling_guide_v3.ASCII, Ohio State University, 1997

[2] Ladd, R. D., *Intonational Phonology* Cambridge, UK: Cambridge University Press, 1996

[3] Grice, M. and Benzmüller, R., "Trainingsmaterialien zur Etikettierung deutscher Intonation mit GToBI and accompanying speech materials", *Phonus 3*, Institute of Phonetics, University of Saarland", 1997, pp. 9-34

[4] Reetz, H., *Pitch Perception in Speech: A Time Domain Approach*, Dordrecht: Foris, 1996

[5] Braunschweiler, N., *Automatic Detection of Prosodic Cues*, PhD thesis, 2003

[6] Syrdal, A. K. and McGory, J., "Inter-Transcriber Reliability of ToBI Prosodic Labeling", ICSLP-2000, Bejing, China, 2000

ISBN 1-876346-48-5