The Prosodizer - Automatic Prosodic Annotations of Speech Synthesis Databases

Norbert Braunschweiler

Speech Technology Group, Cambridge Research Laboratory
Toshiba Research Europe Ltd., 1 Guildhall Street, Cambridge, CB2 3NH, UK
norbert.braunschweiler@crl.toshiba.co.uk

Abstract

Prosodic annotations are used for locating and characterizing prominent parts in utterances as well as identifying and describing boundaries of coherent stretches of speech. In speech synthesis prosodic annotations can be used to improve the unit selection process and subsequently yield more natural sounding synthesis. A method for automatic prosodic annotations of speech is described in this paper. This method is implemented in a computer program called *Prosodizer* that integrates acoustic features of F0 and RMS as well as syntactic and segmental information like POS tags and syllable boundaries. Design and preliminary performance results are described.

1. Introduction

Prosodic annotations can be used in many ways in speech synthesis, for example to enable accurate unit selection from a synthesis corpus. With accurate prosodic mark-up one can gain better synthesis quality and especially more natural sounding synthesis. Since manual prosodic labeling is time consuming and costly, automatic methods have been proposed and implemented (e.g. [1], [2], [3]). Although the results reported are showing promising directions, the automatic handling of the vast acoustic variability given by F0 and energy parameters is still a challenge for producing a speaker and/or language independent automatic labeling tool.

The method proposed in this paper is based on earlier work ([4], [5]) on automatic prosodic labeling that used solely acoustic parameters and is speaker and language independent. In order to improve the recognition accuracy of this approach and to adapt it specifically for labeling speech synthesis corpora, syntactic and segmental information is integrated into the detection process. Syntactic information includes POS tags and syntactic roles (e.g. subject, object). Segmental information encompasses the position of stressed syllables, phone type and phone boundaries, syllable boundaries and word boundaries. These features provide important cues especially for detecting intonational phrase boundaries. Segmental information enables one to measure preboundary lengthening and subsequently improve the recognition of boundary tones. Knowing the position of stressed syllables enables better tonal alignment because pitch accents are labeled by definition in stressed syllables. The ToBI labeling scheme [6] was used in this study because almost all of the available manually labeled reference data uses these labeling instructions.

This paper presents the concept of the Prosodizer and reports first evaluation results based on a German speech corpus and an American English speech corpus as well as on data from the Boston Radio Speech Corpus.

2. Design and implementation

2.1. Previous work

The previous architecture uses a frame-based feature vector, where frames are calculated in 10 ms steps. F0 and RMS values are produced by the ESPS/waves get_f0 (version 1.14) program. A feature vector consisting of 74 acoustic features is created for each frame considering an analysis window of $\pm$400 ms. The feature vector includes 32 F0-features (e.g. duration and extent of rising/falling parts), 26 RMS-features (e.g. number of smaller RMS-values before the current value), and 16 features recording the duration and continuity of voicing (e.g. number of continuously voiced frames before/after the current frame). The motivation for the feature set is based on statistical analyses of manually labeled data and also integrates expert knowledge. The feature vector is evaluated in a scoring module where each individual tone is assigned a score according to its predefined feature constellations. The feature constellations have been established by analyzing features of pitch accents and boundary tones in several manually annotated corpora including different speakers and several languages (see [5] for more detail). Finally the calculated score is used in addition to further sequence restrictions to produce the output: a label file including position and type of pitch accents and boundary tones. The set of prosodic labels used are ToBI-labels [6] and their language specific implementations (e.g. [7]).

However, in order to improve labeling accuracy and tonal alignment and to adapt the tool specifically for annotating speech synthesis corpora, the new architecture is designed to integrate syntactic and segmental information. The main purpose of the Prosodizer is to provide an automatic prosodic mark-up of a given speech synthesis corpus in order to speed up voice building. For these kind of corpora text and phonetic transcriptions are known and segmental information as well as syntactic information are usually automatically annotated and often manually corrected. Therefore this additional information is easily available and accurate. Other domains where the Prosodizer could be applied include all domains where speech data is present and optionally has additional information starting from a phoneme labeling up to a full mark-up including syllable boundaries, word boundaries as well as POS tags and syntactic role information.

2.2. The new architecture

The new architecture of the Prosodizer is outlined in Figure 1. Before the Prosodizer can be applied it is necessary to calculate mean phone durations and standard deviations for each of the phonemes occurring in the corpus. This step is necessary because the measure for normalized phone durations (see [8]) subtracts a phones' mean duration from the current duration and divides the result by its standard deviation. In addition a

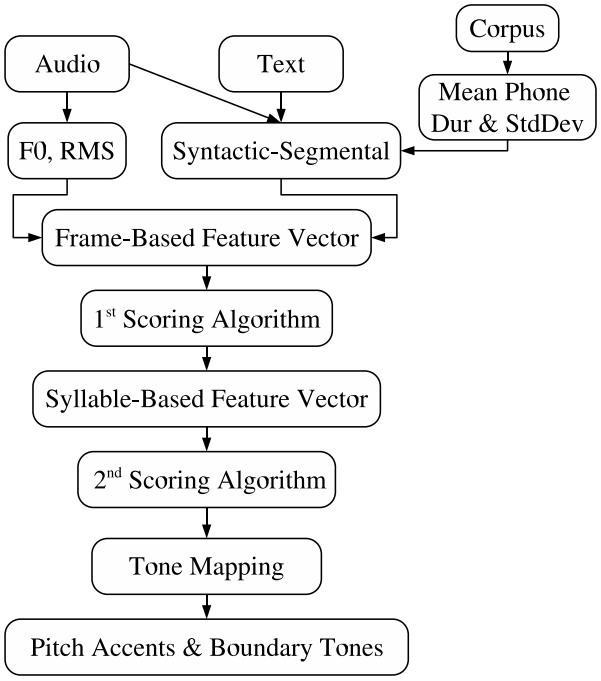

Figure 1: *Outline of the model underlying the Prosodizer.*

scale factor compensating for speaking rate is included in this measure which is calculated on all the phones in a sentence (cf. [8]). The calculation of mean phone durations and standard deviations is done offline based on the phone segmentations. In addition it is also necessary to know the syllable boundaries in order to measure rhyme duration which has been shown to be the most important unit for measuring preboundary lengthening [8]. Syllable boundaries were produced by a rule based algorithm that syllabifies the phonetic transcriptions.

The Prosodizer starts with the same frame-based feature extraction module as described in the previous section. Then the feature vector is combined with syntactic-segmental information and evaluated in the first scoring module. Within this scoring module, feature vectors are scored according to predefined feature constellations for each tone. The feature constellations were established by statistical analyses of manually labeled corpora and take the form of hand-written rules. The first scoring module provides an initial evaluation of the feature vector in terms of how well the feature vector corresponds with predefined feature constellations. The motivation for this first step is based on the fine-grained resolution of these features especially regarding the representation of F0-movements (e.g. detecting rises and falls). For instance, the features estimate rising and falling parts in the F0 contour by comparing successive F0 values and allowing a certain number of outlying F0-values, hence trying to achieve a similar ability to humans who are able to separate short-term deviations from general trends in F0-movements. Methods to compensate for micro-prosodic influences are already built into this approach. These include putting less weight on F0-values at the beginning or at the end of voiced parts, measuring absolute differences between adjacent F0-values and checking the continuity of voicing.

However, in order to do comparisons on a syllable-by-syllable basis the many frame-based feature vectors are reduced to one vector per syllable with the following procedure: from all the feature vectors within the syllable the one with the highest scored pitch accent is selected. For word-final syllables the highest scored boundary tone is added because only those can bear a boundary tone. A subset of the original features is transferred into the new feature vector. Basically features which have already been scored are removed. Additional features that perform comparisons with neighboring syllables are calculated (e.g. comparing mean F0 in the current syllable with the mean F0 in the next/previous syllable). Some of these features overlap with features reported by [8].

The syllable-based feature vector is then fed into a second scoring algorithm that distributes additional scores based on the new features available at that level. As for the previous scoring algorithm the feature scores are established by analyses of manually labeled corpora and additional hand-written rules. Finally the output of this second scoring round is processed in a tone mapping module where tones are selected, deleted or transformed according to further cues like sequence restrictions (e.g. whether a word-final syllable ending in an H-H% is allowed to have an L+H* pitch accent when the preceding syllable already includes an L-H* accent). The latter processing is intended to reduce rather improbable tone sequences (especially when two tones appear within less than 100 ms of each other) instead of actually applying a grammar of the tones in order to eliminate impossible events. The latter has not yet been applied because we wanted to estimate the selectivity of the existing feature set first.

Knowing the word boundaries enables one to restrict the placement of boundary tones to those locations and knowing the position of syllable nuclei allows better tonal alignment. The Prosodizer uses the following types of segmental information:

1. phone identity (e.g. long vowel vs. short vowel),
2. normalized phone duration,
3. normalized syllable duration,
4. normalized rhyme duration,
5. phone boundaries,
6. syllable boundaries,
7. word boundaries.

Additional acoustic features that are calculated based on the knowledge about segment and syllable boundaries are:

1. mean, maximum, and minimum F0 and RMS within a syllable
2. mean maximum, and minimum F0 and RMS within the nucleus, and
3. the shape of F0 within the syllable using a four-way classification scheme: fall, rise, fall-rise, and rise-fall, calculated on the basis of the initial, final and mean F0 (similar to the feature used by [8]).

Syntactic information is currently used only partly. There are simple rules for specific POS classes like function words that are less likely to be accented. Further syntactic features that are planned to be included into the Prosodizer are syntactic roles and rules that check the sequence of POS tags. These additional features are included with the intention to improve the recognition accuracy further (cf. [3] who showed that the inclusion of syntactic information improved recognition results).

Table 1: *ToBI labels used in the Prosodizer. Variants in brackets are not yet detected as separate tones by the Prosodizer but subsumed under the base tone.*

Pitch Accents	Bound. Tones
H* (^H*)	L-L%
!H*	H-L% (!H-L%)
L+H* (L+^H*, L+!H*)	L-H%
H+!H*	H-H%
L*	L-
L*+H	H- (!H-)
H+L* (German only)	

Table 2: *Overall recognition accuracy (%) of the Prosodizer for pitch accents (PA) and boundary tones (BT) in a German corpus of 22,504 syllables and in an American English corpus of 52,954 syllables.*

	German		American Engl.	
	PA (%)	**BT (%)**	**PA (%)**	**BT (%)**
Perfect	65	71	60	68
Partial	8	4	12	3
Insertion	13	7	16	9
Missing	11	14	9	14
Mismatch	3	4	3	6

3. Evaluation

The Prosodizer was evaluated on two speech synthesis corpora and on parts of the Boston Radio Speech Corpus (speaker F2B). The first synthesis corpus includes 2075 German sentences with 13,580 words and 22,504 syllables. The second corpus includes 2749 American English sentences with 32,903 words or 52,954 syllables. Both corpora were automatically segmented into phones and then manually corrected. Syllable boundaries were automatically inserted into the phonetic transcriptions. The corpora were manually labeled with ToBI labels by professional labelers. The set of ToBI labels in the corpora is listed in Table 1. The American English corpus did not include the H+L* accent. In the current system up-stepped variants of H* and L+H* are not differentiated from their base tone, that means L+^H* is merged to L+H* and ^H* becomes H*. Similarly the down-stepped boundary tones L+!H% and !H- are also merged with their base category.

In order to provide a comparison of Prosodizers performance with previously reported work in this domain the Prosodizer was also evaluated on the Boston Radio Speech Corpus. Material from speaker F2B was chosen because it was used as evaluation material in [1]. Since the Prosodizer is designed to include information about phone boundaries, syllable boundaries and POS-tags, only those recordings were chosen which included that information. The subset of the Boston Radio Speech Corpus from speaker F2B included 9082 words and 15005 syllables. Pitch accents are correctly detected in 81% and falsely detected in 12%. Boundary tones are correctly detected in 77% and falsely detected in 8%. These results are slightly lower than previously reported results with regard to pitch accents (e.g. 84% accuracy for pitch accent labeling on the Boston Radio Speech Corpus by [1] or 84.2% by [3]) but exceed the 71% detection accuracy for boundary tones reported by [1] although the false detection rate is higher in the Prosodizer (8% false detections vs. 3% false detection rate in [1]). The detection rate for boundary tones is lower than the one reported in [3] (93%, no notion of false detection rate). However, those approaches did not use the full range of ToBI labels but reduced the label set to a more simple four-class set.

For the German corpus overall accuracy is 76% on pitch accent presence/absence prediction which is higher than the baseline of 56% (the percentage of unaccented syllables out of all syllables). Overall boundary tone detection rate is 78% which is above the baseline of 59% (the percentage of words without boundaries out of all words).

For the American English corpus overall accuracy is 75% on pitch accent presence/absence prediction which is above the baseline of 58% (the percentage of unaccented syllables out of all syllables). Overall boundary tone detection rate is 76% which is above the baseline of 61% (the percentage of words without boundaries out of all words).

These results are lower than previously reported results and are also slightly lower with regard to the pitch accent detection accuracy observed for the Boston Radio Speech Corpus. This is a result of different levels of accuracy in the mark-up and availability of additional information. For instance, syllable boundaries are more accurate in the Boston Radio Speech Corpus than the automatically generated syllable boundaries for the German and American English Corpora. Furthermore, the current implementation is still incomplete with regard to weight settings and lacks a number of rules regarding the syntactic features as well as tone mapping rules. Therefore further improvements in recognition accuracy can be expected.

The results are represented in more detail in Table 2. The recognition results are represented as the percentage of *perfect* matches, e.g. when the Prosodizer detected an H* in the same syllable as in the manually produced reference; the percentage of *partial* matches, e.g. when the Prosodizer detected an L+H* in the same syllable where the reference material has an H*; the percentage of *insertions*, i.e. when the Prosodizer detected a tone in a syllable where the reference labeling did not have a tone; the percentage of *missing* tones, i.e. when the Prosodizer did not detect a tone in a syllable where the reference material had a tone; and the number of *mismatches*, i.e. when the Prosodizer detected an L* in the same syllable where the human labeler placed a high pitch accent. The evaluation for pitch accents is syllable-based whereas the boundary tones are evaluated on a word-by-word basis because boundary tones can only appear at the end of words.

Looking at the results for the German corpus: a rate of 8% partial matches for pitch accents indicates that the Prosodizer does not differentiate well between individual variants of high or low tones (e.g. H* tones in the manual annotation are often labeled as L+H* by the Prosodizer). 13% insertions and 11% missing tones for pitch accents indicate that the accent-individual feature constellations are not yet sufficient. However, when checking the actual results on a sentence-by-sentence basis the overall impression of Prosodizer's recognition accuracy is good. Many of the missing tones are !H* or low accents. Many down-stepped accents are labeled in areas of falling F0 during the syllable, where there are fewer distinctive acoustic cues available, compared to, for example, an L+H* accent. Low accents can be particularly ambiguous, since there is often no strong F0 movement associated with them. The number

of mismatches is small but still indicates that there are different concepts underlying the manual labeling and the automatic labeling.

Regarding the boundary tone detection it is important to mention that the intermediate phrase boundaries (L-, H-) are particularly difficult to detect. This is associated with the perceived boundary strength. Major intonational phrase boundary tones are usually associated with characteristic F0 movements and strong preboundary lengthening effects or additional cues like following pauses (although this does not necessarily trigger a major phrase boundary). However, intermediate phrase boundary tones often do not have characteristic F0 movements nor are there always strong preboundary lengthening effects. The large number of insertion errors for H- and L- tones further supports this view. Insufficient feature constellations and/or unoptimized weight settings are also reasons for the poor detection rate in this category. Another error source is the syllabification, which is not always optimal. For example, it could happen that a coda consonant gets erroneously associated to the onset of the following syllable, which disrupts the measurement of normalized syllable duration because a segment is included which does not belong to this syllable.

The recognition rates for the American English corpus are very close to the ones achieved for the German corpus. However, the overall accuracy is 2% smaller for pitch accents and 1% smaller for boundary tones. There is also an increased rate of insertions, 16% vs 13% for pitch accents. The latter could be a result of missing rules regarding the influences of certain POS tags. This involves a certain amount of language specific adaptations, for example to specify the function words in a language.

These results reveal that the Prosodizer can be applied to another speaker and another language with almost similar recognition rates. Therefore the Prosodizer is capable of handling unseen data with similar recognition accuracy. Whether a similar recognition accuracy can be achieved for a wider range of languages needs still to be tested. The language-specific modifications include (1) adapting the phone set; (2) adapting the POS set; (3) adapting the set of syntactic roles and in the case where there are additional ToBI-tones these would need to be included as well.

4. Conclusions

A method that automatically annotates a speech corpus with ToBI-labels has been presented. The method is implemented in a computer program and integrates acoustic features of F0 and RMS with syntactic-segmental information. This information is then evaluated in two scoring modules and the result is subsequently used in a tone mapping module that selects, deletes or transforms tones based on sequence restrictions. The architecture is rule based and first recognition results on a German corpus, an American English corpus and parts of the Boston Radio Speech Corpus show promising recognition accuracy. However, a relatively large number of insertion errors and the number of missing tones indicate that the selection criteria are not yet sufficient and the weight settings are non-optimal.

Directions for future research include improved feature bundles and optimized weight settings, which could be generated using automatic methods. Furthermore speaker specific parameters in the F0 and RMS domains could be incorporated and used during the selection process. These speaker-specific acoustic parameters could be automatically calculated from the speech corpus and then handed over to the Prosodizer during run-time.

Another issue in the evaluation of the Prosodizer is how important accuracy actually is when comparing its output with manually produced labels. A previous study by [9] has shown that a unit selection system built with automatically produced prosodic mark-up resulted in significantly higher opinion scores than one produced with manual mark-up. Although the manually produced labels are annotated by a professional labeler there are still subjective considerations involved. Therefore the recognition results already achieved might be sufficient in order to produce a reliable prosodic mark-up for speech synthesis corpora. In addition the Prosodizer has the advantage of producing prosodic annotations quickly and reliably and therefore enables fast voice building.

5. References

[1] Wightman, C. W.; Ostendorf, M., 1994. Automatic Labeling of Prosodic Patterns. *IEEE Transactions on Speech and Audio Processing*. 2(4), 469-481.

[2] Ostendorf, M.; Ross, K., 1997. A Multi-Level Model for Recognition of Intonation Labels. In *Computing Prosody*, Y. Sagisaka, N. Campbell and N. Higuchi (Eds.), 291-308, New York: Springer, 291-308.

[3] Chen, K.; Hasegawa-Johnson, M.; Cohen, A., 2004. An Automatic Prosody Labeling System Using ANN-Based Syntactic-Prosodic Model and GMM-Based Acoustic-Prosodic Model. *ICASSP 2004*, Montreal, Canada, 509-512.

[4] Braunschweiler, N. 2003. ProsAlign - The Automatic Prosodic Aligner. *ICPhS 2003*, Barcelona, Spain, 3093-3096.

[5] Braunschweiler, N. 2005. Automatic Detection of Prosodic Cues. PhD thesis. www.ub.uni-konstanz-.de/kops/volltexte/2005/1500/pdf/ProsAlign4.0.pdf, University of Konstanz, Germany.

[6] Beckman, M.E.; Elam, G. 1997. Guidelines for ToBI Labelling, version 3.0, Ohio State University Research Foundation.

[7] Grice, M.; Baumann, S.; Benzmüller, R. 2005. German Intonation in Autosegmental-metrical Phonology. In Jun, Sun-Ah., Prosodic Typology: The Phonology of Intonation and Phrasing. Oxford: OUP, 55-83.

[8] Wightman, C. W.; Ostendorf, M.; Shattuck-Hufnagel, S. 1992. Segmental Durations in the Vicinity of Prosodic Phrase Boundaries. *JASA*. 91(3), 1707-1717.

[9] Wightman, C. W.; Syrdal, A. K.; Stemmer, G.; Conkie, A.; Beutnagel, M. 2000. Perceptually based Automatic Prosody Labeling and Prosodically Enriched Unit Selection Improve Concatenative Text-to-Speech Synthesis. *ICSLP*, Bejing, China, vol.2, 71-74.