

# Automatically extracted F0 features as acoustic correlates of prosodic boundaries

Mattias Heldner<sup>1</sup>, Jens Edlund<sup>1</sup> and Tomas Björkenstam<sup>2</sup>

<sup>1</sup>Centre for Speech Technology (CTT), KTH, Stockholm

<sup>2</sup>Language Engineering Program (STP), Uppsala University, Uppsala

## Abstract

*This work presents preliminary results of an investigation of various automatically extracted F0 features as acoustic correlates of prosodic boundaries. The F0 features were primarily intended to capture phenomena such as boundary tones, F0 resets across boundaries and position in the speaker's F0 range. While there were no correspondences between boundary tones and boundaries, the reset and range features appeared to separate boundaries from no boundaries fairly well.*

## Introduction

We know from literature that the acoustic signaling of prosodic boundaries is a complex one, and that a number of acoustic correlates have been proposed. Apart from the silent pauses, the most important ones are generally held to be speaking rate phenomena such as final lengthening, and intonation phenomena such as boundary tones, fundamental frequency resets across boundaries, and position in the F0 range. These correlates are also the ones most often controlled in text-to-speech systems. But apparently other correlates, such as intensity variation, voice quality and reduction phenomena may play a role too.

Within the GROG project (see Carlson et al., 2002), we have previously investigated features reflecting silent pauses and final lengthening as correlates of prosodic boundaries in Swedish (Heldner & Megyesi, 2003). The present study continues the work towards a general model of the structuring of speech in terms of prosodic boundaries by examining various automatically extracted F0 features intended to capture boundary tones, F0 resets across boundaries, and relative position in the speakers F0 range as acoustic correlates of prosodic boundaries.

## Method

### Speech material and annotation

The speech material used in this study is a rendering of a children's book by a male professional reader at a publisher of talking books. In terms of speaking style it could be classified as a read-aloud monologue. The duration of the recording is 17 minutes and it contains about 2700 words. It was collected and annotated within the GROG project.

This material was manually annotated for perceived boundaries by three experienced transcribers. Each word was marked as being followed by either a weak or a strong boundary, or as not followed by a boundary at all. The inter-rater reliability of this task was fairly high. The pair-wise agreement was 97%, and the corresponding Kappa 88%. See Poesio & Vieira (1998) for agreement and Kappa methods. To further increase the quality of the annotations, the majority votes of the three transcribers were used. The majority votes resulted in 282 cases of strong boundaries, 169 weak boundaries, 2232 no boundaries, and 2 cases of disagreement among the raters.

### Segmentation

The first step in the feature extraction process was to segment the speech material using NALIGN, an automatic alignment algorithm developed at CTT (Sjölander, 2003). NALIGN uses a Hidden Markov Model based method to generate word and phoneme level transcriptions from a verbatim orthographic transcription and a sound file. The phoneme tier is supplemented with lexical prosodic information including primary and secondary stress, and word accent type. For an evaluation of the alignment precision at the word level, see Sjölander & Heldner (2004).

The aligner output was subsequently used to identify three sub-word regions from where the features were to be extracted: the word initial onset plus vowel (WIO) including all con-

sonants in the word initial syllable onset plus a vowel; the stressed syllable rhyme (SSR) containing the stressed vowel plus one consonant (if there was one); and the word final rhyme (WFR) containing the vowel in the word final syllable and any following consonants until the word boundary. The WFR is thus the mirror image of the WIO.

### **F0 preprocessing**

There were also a number of preprocessing steps prior to the F0 feature extraction. First, the fundamental frequency was extracted using the `get_f0` function from the Entropic Signal Processing System (ESPS) as implemented in the freely available software WaveSurfer (Sjölander & Beskow, 2000), see also <http://www.speech.kth.se/wavesurfer>. Subsequently, F0 values were interpolated across voiceless segments skipping the first voiced frame after the voiceless segment. This was done to reduce the influence of consonants raising or lowering F0 at voice onsets. The interpolated F0 output was then transformed into the distance in semitones relative to a fixed value of 100 Hz. This transformation was motivated by the following step; an online estimation of speaker F0 range. The F0 range was bounded by a topline and a baseline defined as the cumulative mean  $\pm 1$  standard deviation (also calculated cumulatively). The F0 range was online in the sense that it was calculated using left context only. The semitone scale was used to ensure that +1 standard deviation was the same musical (and perceptual) interval as -1 standard deviation.

### **F0 features**

A wide range of F0 features was extracted from the preprocessed F0 data for each extraction region (i.e. WIO, SSR and WFR). These features (listed below) included raw features, as well as boundary tone, reset and range features that were calculated from the raw features. Several alternative range and reset features were extracted.

#### **Raw features**

- The values at onset and offset positions
- The minimum and maximum values and the corresponding times
- The mean value
- The values of the online baseline at the onset and offset positions

#### **Boundary tone features**

- F0 shape (categorical): rise, fall, rise+fall, fall+rise and none
- The sizes and slopes of the F0 shapes

#### **Range features**

- The difference between the offset and baseline at the offset position
- The difference between the mean and the baseline at the offset position

#### **Reset features**

- The difference between the offset and onset in neighboring sub word regions (left to right)
- The difference between the minimum and maximum in neighboring sub word regions
- The difference between the mean values of neighboring sub word regions

## **Results**

Although all features were extracted from three sub word regions (i.e. WIO, SSR and WFR) the presentation of results will be limited to those from the word final rhymes, as the word endings seem the most relevant for examining prosodic boundaries.

### **Boundary tone features**

Figure 1 summarizes some of the boundary tone feature data by showing the distribution and average size of F0 shapes for the different boundary types. The slopes of the F0 shapes will not be commented on here. This analysis revealed that there was no simple one-to-one correspondence between presence of a boundary tone (in terms of falling, rising, falling+rising and rising+falling F0 shapes) and boundary type. For example, 'none', the shape for movements smaller than one semitone, was the most frequent shape in the no and strong boundary categories, and the second most frequent shape in weak boundaries. Similarly, the falls and rises were frequent at boundaries as well as at no boundaries. The combinations of falls and rises, too, occurred in all boundary types, although they were considerably less frequent. Thus, neither absence nor presence of a particular F0 shape in the word final rhymes seemed to indicate any special boundary category.

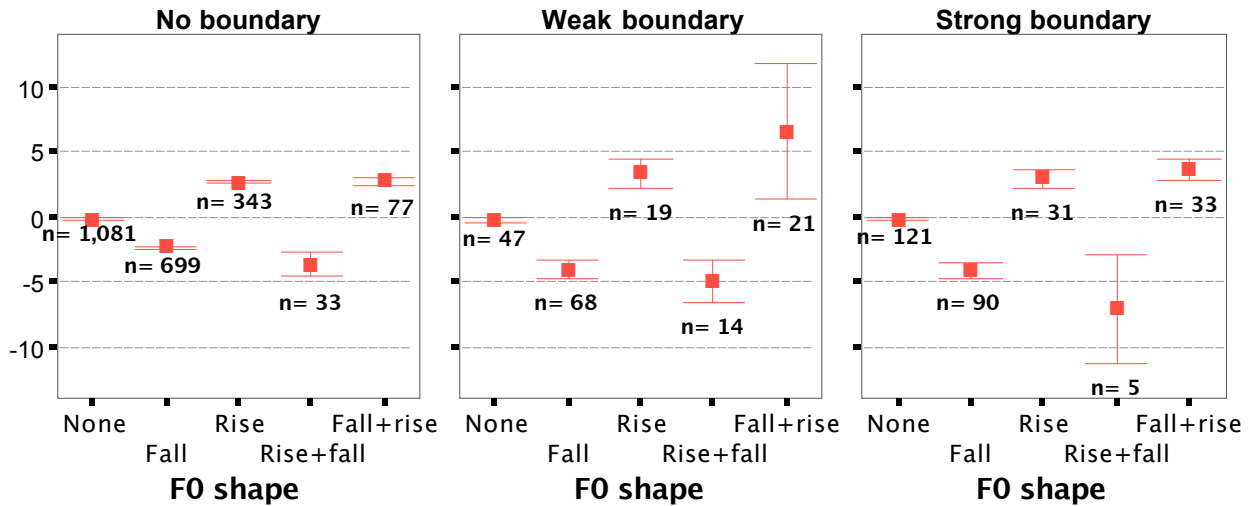


Figure 1. The distribution of F0 shapes for the different boundary types, as well as the average size of these shapes (in semitones) with error bars representing a 95% confidence interval. When the F0 shapes are composed of two movements, i.e. rise+fall or fall+rise, the size of the second movement is shown.

However, the size of at least one of the F0 shapes could possibly discriminate no boundaries from weak and strong boundaries. The average fall was almost twice as large in weak and strong boundaries as in no boundaries. A similar pattern was also found in the rise+fall shapes. There were no apparent mean differences between boundary types for the rises and the fall+rise shape.

### Range features

There were two alternative features intended to capture the position in the F0 range based on the online estimation of F0 range. Both features showed a trend for the distance to the baseline to decrease with increasing boundary strength. Figure 2 shows the feature measuring the distance between offset and baseline. Thus, a position close to the baseline could possibly indicate a boundary. Both measures appeared to separate the boundary categories equally well.

### Reset features

There were also three alternative features intended to capture the F0 reset across boundaries by looking at differences between various raw features from neighboring units; in this case the difference between the WFR and the following WIO. They all showed tendencies for a larger F0 reset the stronger the boundary. Figure 3 shows the results for the one of them that gave the best separation of boundary types, namely the difference between mean values.

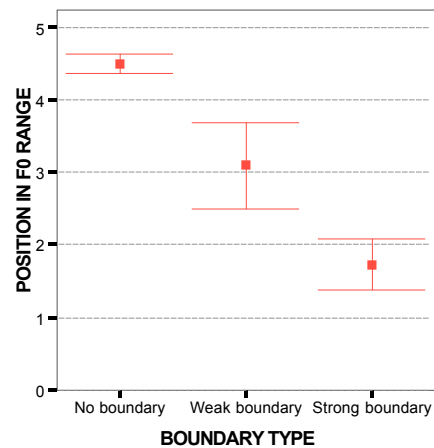


Figure 2. The average distance between offset and baseline (in semitones) for the different boundary types. Error bars represent a 95% confidence interval.

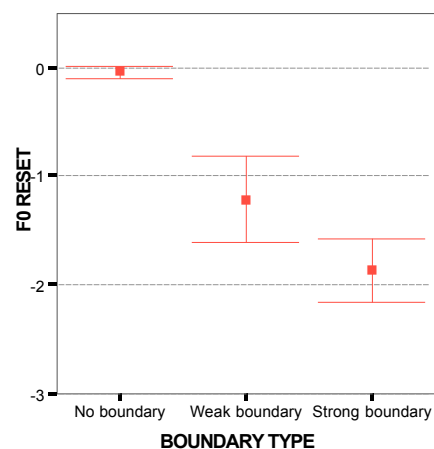


Figure 3. The average difference between the mean values of neighboring units (in semitones) for the different boundary types. Error bars represent a 95% confidence interval.

As can be seen in Figure 3 there was hardly any F0 reset at no boundaries and a reset of almost two semitones at strong boundaries. Thus, it seems that a (relatively) large F0 reset could indicate a boundary, and that this feature separate the different boundary types fairly well.

## Discussion

In this study, we have examined a number of automatically extracted F0 features intended to capture certain intonation phenomena, which in turn are held to be important for the signaling of prosodic boundaries. Thus, the F0 features are merely operationalizations of the intonation phenomena, and the results of the study are dependent on the precision and quality of these operationalizations. While a positive result may be taken as support for the claim that a phenomenon is important for the signaling of prosodic boundaries, a negative result might just as well indicate a poor implementation of features.

This said, we may observe that the simple categorization of F0 shapes into rising, falling, rising+falling, falling+rising, and none did not reveal any one-to-one correspondences between boundary tones and prosodic boundaries. However, whether a further subcategorization (for example into high vs. low, fast vs. slow or large vs. small movements) would result in a clearer picture remains to be investigated. The observed mean differences in size of falls between boundary types indicate that this might be a reasonable path to follow.

Furthermore, the features reflecting position in F0 range showed a trend in the expected direction, and separated the boundary categories fairly well. Since the range estimation is online and based on the cumulative mean and cumulative standard deviation, it takes about 30 to 40 seconds before the range stabilizes, which would contribute to variability.

The features intended to capture F0 reset across boundaries also demonstrated a trend in the expected direction, and separated the boundary categories fairly well.

Future work will most certainly include experiments with machine learning techniques to investigate to what extent the automatically extracted F0 features can be used to discriminate between boundary categories, by themselves and in combination with other features capturing for example silent pauses and final lengthening.

## Acknowledgements

The authors are grateful to the Swedish Library of Talking Books and Braille (TPB) for making the speech data available. This work was carried out within the GROG project (Gräns och gruppering – Strukturering av talet i olika kommunikativa situationer) supported by The Swedish Research Council (VR) 2002–2004. The research was carried out at CTT, the Centre for Speech Technology, supported by VINNOVA (The Swedish Agency for Innovation Systems), KTH and participating Swedish companies and organizations.

## References

- Carlson R., Granström B., Heldner M., House D., Megyesi B., Strangert E., and Swerts M. (2002) Boundaries and groupings - the structuring of speech in different communicative situations: a description of the GROG project. *THM-QPSR 44*, 65-68.
- Heldner M., and Megyesi B. (2003) Exploring the prosody-syntax interface in conversations. In *Proceedings ICPHS 2003* (pp. 2501-2504). Barcelona.
- Poesio M., and Vieira R. (1998) A corpus-based investigation of definite description use. *Computational Linguistics 24*(2), 183-216.
- Sjölander K. (2003) An HMM-based system for automatic segmentation and alignment of speech. In *Proceedings of Fonetik 2003* (pp. 93-96). Umeå.
- Sjölander K., and Beskow J. (2000) WaveSurfer - an open source speech tool. In *Proceedings ICSLP'2000*. Beijing, China.
- Sjölander K., and Heldner M. (2004) Word level precision of the NALIGN automatic segmentation algorithm. In *Proceedings of Fonetik 2004*. Stockholm.