ToBI Accent Type Recognition

Arman Maghbouleh

Department of Linguistics
Stanford University
Stanford, CA 94305, U.S.A.

ABSTRACT

This paper describes work in progress for recognizing a subset of ToBI intonation labels (H*, L+H*, L*, !H*, L+!H*, no accent). Initially, duration characteristics are used to classify syllables as accented or not. The accented syllables are then subclassified based on fundamental frequency, F0, values.

Potential F0 intonation gestures are schematized by connected line segments within a window around a given syllable. The schematizations are found using spline-basis linear regression. The regression weights on F0 points are varied in order to discount segmental effects and F0 detection errors. Parameters based on the line segments are then used to perform the subclassification.

This paper presents new results in recognizing L*, L+H*, and L+!H* accents. In addition, the models presented here perform comparably (80% overall, and 74% accent type recognition) to models which do not distinguish bitonal accents.

1. INTRODUCTION

Intonation recognition is crucial for any speech understanding system. For example, the human client in a flight reservation dialogue can intone the words

flight thirty five from Chicago

in different ways and expect different actions from a, human or machine, travel agent:

- The client may say the words in a declarative manner as a way of confirming what he/she has heard. In this case, the travel agent does not need to reply if the client's understanding is correct.
- The client may say the words in a way as to question the first number: *Flight **thirty** five from Chicago?* In this case, the travel agent is required to reply about the flight number, but not the destination
- The client may also utter the exact same words but mean it a correction, as if to say: *I wanted flight **THIRTY** five not forty five*. In this case, the agent is required to revisit the previous task with a new flight number.

Using ToBI [10], a standard for describing intonation, the first syllable in *thirty*, in each of the above cases, would be marked with H*, L*, and L+H* respectively. Complete ToBI labeling of the utterance would include potential accent markings on other syllables, an indication of the phrasing of the utterance (i.e. break indices), and an indication of tonal movement at edge of phrases (i.e., phrase and boundary tones).[1]

Besides the already mentioned, H*, L*, and L+H* accent types, ToBI recognizes L*+H, H+!H*, !H*, L+!H*, and various markings for indicating uncertainty on behalf of the transcriber. The *H* in each label implies that there is a high F0 value, peak, or rise associated with the syllable in question. Similarly the L implies a low F0 value, valley, or fall. The L+H* and L*+H accents, then, are associated with a fall followed by a rise. The difference between the two is that for L+H*, the peak is aligned with the syllable, whereas in L*+H, the valley is so aligned. The *!* diacritic, called downstep, marks a variant of H labels with lowered peaks.

1.1. Material

The WBUR Radio News Corpus [5] was used in training and testing of the models developed in this study. This corpus consists of recordings of broadcast radio news stories. Thirty-two of the stories (9111 words) that were read by a professional female announcer were used here. The recordings were made in the studio during broadcast and were later digitized, transcribed, segmented, and hand-labeled with ToBI labels. Accent counts for this corpus are shown in table 1.

	NoAcc	**H***	**!H***	**L+H***	**L+!H***
count	9845	2873	800	540	238
% of total	66%	19%	5%	4%	2%
	?[2]	**L***	**H+!H***	**L*+H**	**L*+!H**
count	447	187	65	12	3
% of total	3%	1%	<1%	<1%	<1%

Table 1: Count of syllable accents in corpus.

This paper is concerned with the recognition of H*, L*, L+H*, !H*, and L+!H* because these are the most prevalent accents in the present corpus.[3]

2. OVERVIEW

Ross & Ostendorf [9] divides intonation recognition efforts into those that

- model complete F0 contours, versus those that
- use a transformation of local F0 patterns and other cues given an utterance segmentation

---

1 See [6] for interpretations of ToBI labels.

2 Here ? stands for labels signifying transcriber uncertainty.

3 In this study L*+H and L*+!H were combined with L*. The ? accents and H+!H* were combined with !H*.

The present work falls into the second category. Assuming that syllable and phone segmentation is available, each syllable is classified as being accented or not using duration information [4]. If accented, F0 information is used to determine the accent type. The rest of this paper describes the accent type classification procedure.

The current division between duration and F0, and the one-syllable-at-a-time approach are mere conveniences used during development of a full ToBI recognition system. These aspects of the current model will be revisited in future work.

3. ISSUES

3.1. Unreliable Signal

The most commonly cited problem when dealing with fundamental frequency data is that individual F0 values are unreliable indicators of intonational gestures: F0 values are missing during unvoiced sections, consonantal effects on F0 mask intonational gestures, and F0 detection errors are common. Figure 1 illustrates these issues. The dots are the actual F0 values reported in the corpus. The lines interpolate between missing values. The rises and falls for the three peaks/valleys marked in the figure are due to segmental effects of a stop consonant and not intonational intent.

Attempts to isolate intonationally meaningful F0 variation [1,3,11] usually smooth the F0 contour in various ways. The present author's attempts at overall smoothing were not successful in that, successful smoothing of segmental effects also smoothed over the fine distinctions between H* and L+H* accents. The present solution, to be described in section 4.1, has been to limit schematizations to allow falls, and rises only within reliable portions of the F0 contour.

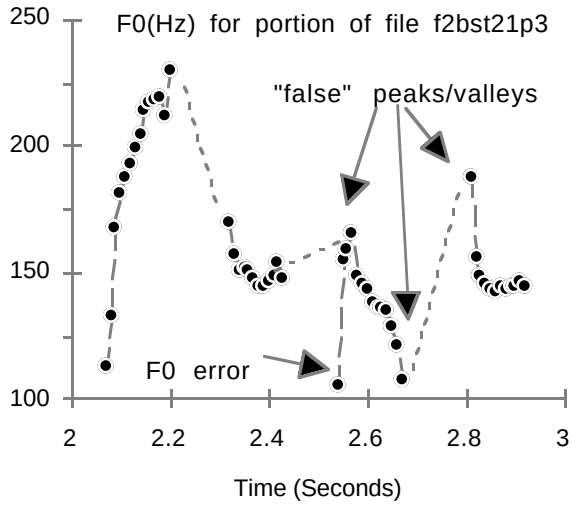

Figure 1: Segmental effects on F0.

3.2. Alignment

A second issue is that F0 gestures (e.g., rises and falls) must be interpreted with respect to the syllable to which they are assigned. This is not always an easy task. For example, a fall-rise-fall-rise structure on four syllables can be L*, H*, L*, H*; or L+H*, _, _, L+H*; or …?[4]

These alignments are critically important in recognition of ToBI labels. The approach taken in this study is to tie F0 gestures only to those syllables which have been determined, based on duration values, to potentially hold an accent.

3.3. Context

To the human eye, the F0 contour for a H* accent in isolation looks different than one which is immediately followed by a boundary tone. The approach taken in this study is to, as much as possible, identify context independent gestures for each label. This initial study, therefore, provides no explicit mechanism for handling context-dependent shapes.

4. ACCENT TYPE MODEL

4.1. Schematization

An exploratory study confirmed the presence of downward F0 gestures before syllables bearing L+H* or L+!H*. These two accents are the most tonally complex of the set. It was therefore hypothesized that a schematization which allowed three gestures was as complex a schematization as needed for capturing intonational F0 movements associated with an accent on a syllable.

A third degree polynomial, or a first degree spline with two knots (i.e., three connected lines) are the simplest ways of schematizing a contour with three gestures.[5] Both schematizations were carried out by means of regression and both were found to be adequate. Splines were ultimately chosen because their coefficients are more easily interpretable.

In order to facilitate alignment of tone movements with syllables, the first knot of the splines, that is, the end of the first line segment, was restricted to be before the syllable, and the second knot was restricted to be within the syllable.[6] The exact knot locations for each syllable were automatically selected to minimize regression error.

---

[4] There are many other possibilities, especially when one takes phrase accents, not discussed in this paper, into account.

[5] The adequacy of straight-line schematizations is investigated in [2].

[6] The peak of a high accent may be after the syllable itself. However, it turns out that restricting the second knot to be within the syllable does not preclude the recognition of accent types in these cases.

In order to ameliorate the negative effects of unreliable F0 points, the regression weights were reduced for F0 points near consonantal edges, for those too different from their neighbors, and for those below or above the speaker's F0 range. The result of schematizations for two accents are shown in figures 2 & 3.

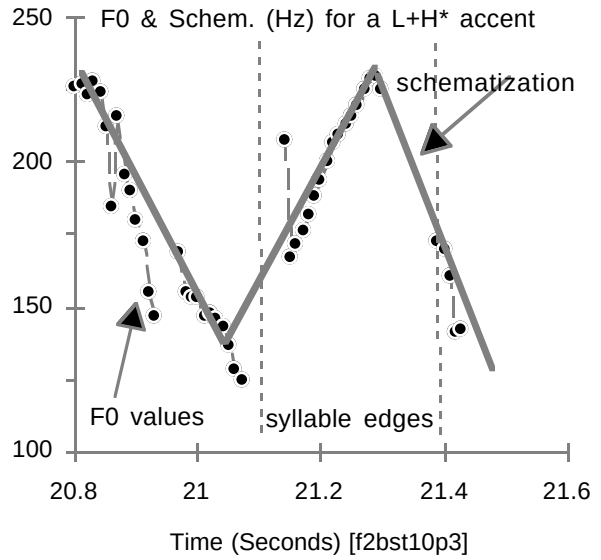

Figure 2: F0 values and three-line schematization of a prototypical L+H* accented syllable.

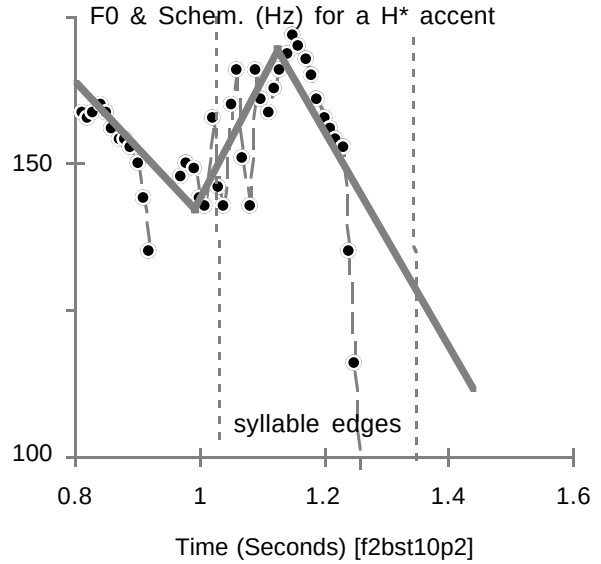

Figure 3: F0 values and schematization of a H* accent.

4.2. Parametrization & Recognition

L+H* accents turn out to be consistently shaped like figure 2: a falling gesture followed by a rising one, with slope difference between the two segments being larger than 50 Hz per 100 milliseconds. L* accents consistently harbor low F0 values (less than 150 Hz). Unfortunately H* accents seem to take on a myriad of shapes. Also, H* accents are ubiquitous, hence, error rates can be high even though the other accent types are distinguishable.

Some H* accents superficially resemble L+H* accents (e.g., figure 3) but can be distinguished from L+H* accents because they lack sufficient evidence for the fall prior to the beginning of the syllable. One can observe this phenomena by comparing the number of F0 points which fall along the first line segment in figures 2 and 3. Having located L+H* and L* accents, what is left is labeled as H*.

Downstepped high accents are at the moment recognized by comparing peak F0 values of the schematization to the peak value in the prior syllable. If the peak in syllable in question is lower, it gets labeled as downstepped. The current model was heuristically configured and not optimized, however, its results can form a touchstone against which future work can be judged.

5. RESULTS

The focus of this study was accent type recognition, therefore to test the model one hundred syllables already known to be accented, twenty of each type under investigation, were chosen at random from the corpus. The confusion matrix for their automatic classification results is shown in tables 2a and 2b.

Recog-nized	Hand-labeled				
	H*	!H*	L+H*	L+!H*	L*
H*	14	10	8	1	1
!H*	0	1	0	1	2
L+H*	3	2	9	7	1
L+!H*	1	6	0	11	1
L*	2	1	3	0	15

Summary	Model Accuracy	Baseline
Overall	50%	20%

Tables 2a,b: Confusion matrix for accent type classification.

Prior published results on ToBI recognition [5,7,8,9] have clumped the bitonal accents with their monotonal counterparts. To allow comparison to previous work, the model was also tested on 3366 syllables, not pre-screened to be already accented, in the corpus. The results along with results of [8] are reproduced in tables 3 and 4. In these tables *high* refers to collection of H* and L+H*. Similarly *downstep* refers to !H* and L+!H*. In table 3, the uncertain accents, see table 1, were combined with their most prevalent prediction, downstep.

The summaries include overall accuracy (i.e., percentage of syllables correctly classified), percentage of syllables correctly labeled as being accented or not, and percentage of syllables correctly classified with accent type, given that those syllables were already correctly identified as being accented.

Recog-nized	Hand-labeled No Accent	high	downstep	L*
No Acc.	1919 (86%)	60 (08%)	22 (08%)	17 (24%)
high	105 (05%)	590 (78%)	120 (42%)	4 (06%)
down.	115 (05%)	76 (10%)	125 (43%)	7 (10%)
L*	105 (05%)	35 (05%)	22 (08%)	44 (61%)
Total	2244	761	289	72

Summary	Model Accuracy	Baseline
Overall	80%	67%
Acc Y/N	87%	67%
Acc Type	74%	69%

Tables 3a, 3b: Recognition results on a corpus segment.

Recog-nized	Hand-labeled No Accent	high	downstep	L*
No Acc.	2120 (91%)	52 (07%)	57 (25%)	52 (63%)
high	157 (07%)	644 (89%)	89 (39%)	14 (17%)
down.	50 (02%)	23 (03%)	80 (35%)	12 (15%)
L*	5 (00%)	5 (01%)	2 (01%)	4 (05%)
Total	2332	724	228	82

Summary	Model Accuracy	Baseline
Overall	85%	69%
Acc Y/N	89%	69%
Acc Type	83%	77%

Tables 4a, 4b: Recognition results from [9].

6. DISCUSSION

As mentioned before, there are no existing benchmarks for human or machine performance for the full set of accents in table 2a. However, the high recognition results of L* accents which is based on the extremely simple parameter of F0 value of schematized contour in mid-syllable is noteworthy in itself. Also noteworthy is the extremely poor recognition of downstepped H* accents which suggests that one may need to go beyond the isolated-syllable approach to recognize downstep. The present model is good at recognizing the bitonal nature of L+!H* accents (90% recognition) but 40% of L+H* accents are misclassified as the H* monotone.

The performance of this model in its abridged form, table 3b, is encouraging in that it achieves accent type recognition results similar to the much more complex model in [9] (both models perform at 5% above the baseline).

Pitrelli et al. [7] do report intertranscriber agreements of 68.3% overall,[7] 80.6% accent placement, and 64.1% accent type recognition. However, those summary results are not comparable to results here since the task in [7] was considerably more difficult with varied speaking styles and required agreement for more than two transcribers.

Ross [8] reports intertranscriber agreements of 73% overall, 90% accent placement, and 60% accent type recognition for the corpus used in this study. However, 65% of the accents in this corpus are H* or uncertain, so one can potentially achieve summary human performance on accent type recognition by labeling all accented syllables as H*.

The present work will extend to other ToBI labels, and will most likely include more context information in the models so that downstepped accents can be better discriminated.

[7] [7] reported results based on words, not syllables. Assuming one accent per word, the accent type recognition rates do not change. Also, at approximately 1.65 syllables per word, the overall accuracy rates based on words decrease by no more than 4%.

7. REFERENCES

1. Alessandro d', C. and P. Mertens (1995). "Automatic pitch contour stylization using a model of tonal perception." Computer Speech and Language **9**(3): 257-288.
2. Hart 't, J. (1991). "F0 stylization in speech: straight lines versus parabolas." Journal of the Acoustical Society of America **90**(Dec. '91): 3368-70.
3. Hirst, D. and R. Espesser (1993). Automatic modeling of fundamental frequency using a quadratic spline function, Travaux de l'Institut de Phonétique d'Aix. **15:** 71-85.
4. Maghbouleh, A. (1996). A logistic regression model for detecting prominences. International Conference on Spoken Language Processing, Philadelphia, PA.
5. Ostendorf, M., P. J. Price and S. Shattuck-Hufnagel (1995). The Boston University Radio News Corpus, Boston University Electrical Engineering.
6. Pierrehumbert, J. and J. Hirschberg (1990). The Meaning of Intonational Contours in the Interpretation of Discourse. Intentions in Communication. P. R. Cohen, J. Morgan and M. Pollack. Cambridge, Mass., MIT Press**:** 271-311.
7. Pitrelli, J. F., M. E. Beckman and J. Hirschberg (1994). Evaluation of Prosodic Transcription Labeling Reliability in the ToBI Framework. International Conference on Spoken Language Processing, Yokohama, Japan.
8. Ross, K. and M. Ostendorf (1994). A Dynamical System Model for Generating F0 for Synthesis. ESCA/IEEE Workshop on Speech Synthesis.
9. Ross, K. and M. Ostendorf (1995). A Dynamical System Model for Recognizing Intonation Patterns. Eurospeech.
10. Silverman, K., M. Beckman, et al. (1992). ToBI: A Standard for Labeling English Prosody. International Conference on Spoken Language Processing.
11. Taylor, P. A. (1994). "The Rise/Fall/connection Model of Intonation." Speech Communication **15**: 169-186.