RC23086 (W0401-130) January 28, 2004
Computer Science

IBM Research Report

ToBI Prosodic Analysis of a Professional Speaker of American English

John F. Pitrelli
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

ToBI Prosodic Analysis of a Professional Speaker of American English

John F. Pitrelli

IBM T.J. Watson Research Center, Yorktown Heights, New York, U.S.A.
pitrelli@us.ibm.com

Abstract

We analyze the distribution of ToBI labels in a corpus collected from a professional speaker for use in concatenative speech synthesis. Our goals include using such statistics to aid automatic ToBI labeling of such a corpus, analogously to how a language model aids speech recognition. We find that the professional speaker produces a rich variety of prosodic events. ToBI labels occur with skewed frequencies, with a trigram model for occurrences of 34 ToBI labels yielding a perplexity of 3.23, indicating that such statistics will likely aid recognition of those prosodic categories. We relate ToBI label occurrence to sentence type and word frequency, determining patterns which confirm that text information would also useful to such a recognizer.

1. Introduction

While speech technologies remain far from perfect, they are reaching a level of performance such that research focus can expand from merely trying to decrease word error rate for recognition and trying to increase word-level intelligibility for synthesis. Recognition research has grown to include pursuits such as automatic detection of user frustration; synthesis research now pursues proper **expression**, *e.g.* for various types of questions, contrastive emphasis, conveying good or bad news, etc. [3] These newer directions imply an increased role for prosody in speech technologies. Because both recognition and synthesis are reliant on statistical techniques and large quantities of speech data, it is appropriate to begin prosodic analyses of large speech corpora with an eye to these technologies.

The present work is motivated primarily by American English expressive concatenative speech synthesis. While one approach is to collect corpora directly representing a variety of such expressions, the approach motivating the present study is to associate each expression with a linguistic representation of prosody, and then in turn to associate elements of that representation with acoustic correlates, as described previously [3]. Thus, we are confronted with several tasks: (1) choosing an appropriate prosodic representation, (2) collecting a prosodically-rich corpus, (3) labeling it using the chosen representation, and (4) learning the associations between these labels and both acoustics and expressions.

We chose American English Tones and Break Indices (AmE-ToBI) for the prosodic representation, as it appears to represent a reasonable consensus of many researchers of English prosody, and it exhibits acceptable stability as evidenced by consistency among transcribers [2] [5]. AmE-ToBI analyzes intonation in terms of a hierarchy of intonational phrases, each of which ends in a boundary tone such as L% or H% and contains one or more intermediate phrases, each of which in turn ends in a phrase accent such as L- or H- and contains one or more pitch accents, such as H*, L*, or L*+H. In addition to this tone tier information, AmE-ToBI also provides a break index tier, representing the degree of disjuncture between adjacent words. Examples are 4 for a full intonational phrase break, 3 for an intermediate phrase break, and 1 for most phrase-internal word boundaries. Silverman *et al.* [4] and the ToBI official web site [7] provide full descriptions of AmE-ToBI.

We model the relationship between ToBI labels and expressions using rules, and the relationship between labels and acoustics using decision trees, as described previously [3]. Training the latter ultimately requires a prosodically-rich corpus, to provide sufficient examples for each label's model. For now, we use a concatenative synthesis corpus consisting of a professional speaker reading a script chosen for phonetic coverage as well as real-world applicability. Thus, one goal of the present work is to determine an inventory of prosodic units and sequences, to play a role in the design of future scripts and perhaps directions to the speaker, in order to produce a corpus prosodically rich enough to enable developing complete models relating ToBI both to acoustics and to a full range of expressions.

The labeling task is key here; manual prosodic labeling is expensive, requiring substantial time by someone with significant training. Accordingly, we ultimately seek automatic prosodic labeling of the corpus. As with phonetic labeling, this is a recognition task. However, while recognition technology is routinely used for the phonetic labeling task, the prosodic labeling task is complicated by two factors: (1) the "correct" prosodic labels for an utterance are unknown *a priori*, while the phonetic recognition problem is nearly one of alignment of a known sequence obtained from a dictionary, and (2) the relative newness of AmE-ToBI means that large-scale corpus collection oriented toward complete coverage of prosodic contexts is just beginning, and so automatic recognition of prosodic labels such as ToBI is in its infancy compared to phonetic recognition; current work focuses mainly on merely detecting rather than classifying accents [1]. In the interim, therefore, a large expense for manual ToBI labeling is a cost of our approach, in terms of time and specialized expertise [8], though someday we anticipate this situation improving, much as phonetic time-alignment is now largely automated. Work on automated pre-processing to speed hand-labeling has achieved a small benefit so far [6].

Thus, our interest in synthesis creates for us an interest in recognition of prosodic units. Recognition tasks in general, for units such as phones or words, are typically performed using an acoustic model, representing the acoustics of each unit, and a language model, representing the *a priori* likelihood of these units. Drawing analogy to prosodic units, the frequencies of occurrence of the various ToBI labels is to be expected to be highly skewed, suggesting that applying a "language model" of label-occurrence statistics to this process may yield substantial benefits in an eventual recognizer. One advantage of this recognition scenario for labeling a synthesis corpus is that the text is available, so such a model can employ text as an input feature.

Unfortunately, we are unaware of any study of a ToBI-

Table 1: *Percentages of word occurrences marked with each ToBI break index, excluding obligatory sentence-final* 4*'s.*

Break Index	Percent	Break Index	Percent
1	70.13%	1-	0.75%
4	20.91%	2-	0.41%
3	2.67%	0	0.24%
4-	1.81%	2p	0.01%
3-	1.70%	1p	0.003%
2	1.35%		

labeled corpus of more than a few hundred sentences collected from a professional speaker. Therefore, we have several reasons to study the characteristics of our own prosodically-labeled corpus for concatenative synthesis, beyond providing a window on the prosody of professional speech, which we expect to make particularly rich use of the prosodic inventory of the language. We anticipate using the data obtained in this study to help us design a future prosodically-rich corpus, and to lay the groundwork for the occurrence-statistics component of an automatic prosodic labeler to facilitate the practical development of large corpora for expressive concatenative speech synthesis.

2. Corpus

A female professional speaker recorded a training script designed for concatenative speech synthesis. The speaker is in her late 30's, raised in many states mainly on both coasts of the U.S., mostly in the northeast, with professional speaking experience in theater, advertising and educational audio publishing.

As mentioned above, the script is designed for phonetic coverage and real-world applicability; it provides roughly 10 hours of speech. Recordings were made in a studio, and sampled at 22 kHz. An excerpt of 2880 sentence utterances was manually transcribed by one labeler using full AmE-ToBI, not counting utterances for which the ToBI grammaticality checker indicated that there was an inconsistency in the transcript. However, the manual transcription omitted HiF0, which was estimated by an automatic algorithm later, and so is not included in this analysis. Also omitted from the present analysis are $<$ and $>$, and ToBI's miscellaneous tier. The corpus provides 42,250 word occurrences, for an average of 14.7 per sentence.

3. Results

3.1. Overall statistics

Table 1 shows the percentage of words marked with each AmE-ToBI break index, excluding sentence-final obligatory 4's, for a total of 39,370 words. Most striking is that full intonational phrases average under five words, as evidence by more than 20% of words being marked 4 even with sentence-final 4's excluded. This finding is consistent with our informal observation that the professional speaker's utterances are prosodically rich; compared to typical talker's speech, she and other professional speakers produce more complex intonational and phrasing patterns, some of which clearly aid the listener in perceiving the structure of the sentence. Also noteworthy is the relative lack of 3's, indicating that most intermediate phrase boundaries coincide with full intonational phrase boundaries.

Table 2: *Percentages of word occurrences marked with each ToBI tonal element. "None" indicates no ToBI tone marking on a word. Percentages add to more than 100 because some words have multiple marks, such as* H* *and* L-L% *for a typical declarative-sentence-final accented word.*

Tone	Percent	Tone	Percent
none	36.05%	L*+!H	1.20%
H*	16.37%	*?	0.72%
L-L%	16.07%	!H-L%	0.53%
L+H*	14.57%	H-H%	0.40%
!H*	11.04%	!H-	0.40%
L-H%	7.76%	H+!H*	0.37%
L*	4.60%	%H	0.35%
L*+H	2.86%	!H-H%	0.05%
L+!H*	2.42%	X*?	0.03%
H-L%	1.92%	-X?	0.002%
L-	1.74%	%r	0.002%
H-	1.74%		

Similarly, we analyzed the frequency of events in the tone tier. Table 2 shows the percentage of the 42,250 words which were marked with each AmE-ToBI tonal element. As expected, we see considerable skew in the distribution of frequencies of occurrence of various tonal elements, suggesting that occurrence statistics should assist in recognizing them. Not surprisingly, "none", variants of H*, and L-L% are the most common, as to be expected respectively for the most typical unaccented word, accented word, and final word in phrase for declarative sentences, which comprise 2711 (94%) of the 2880 sentences in the corpus. H* and its variants appear on 45% of words, compared to 9% for versions of L*. Interestingly, the professional speaker produces almost as many occurrences of L+H* as H*, presumably to be perceived more strongly as an accent, part of the richer prosody perceived in her speech, though the pattern does not hold for the downstepped counterparts L+!H* and !H*.

Phrase accents and boundary tones likewise reflect the predominance of declarative sentences in the corpus. Twenty-six percent of words $(16.07 + 7.76 + 1.74)$, or 84% of intermediate phrases $(26\ /\ (26 + 1.92 + 1.74 + 0.53 + 0.4 + 0.4 + 0.05))$ are marked with L-, while 5% of words / 16% of intermediate phrases have H- or !H-. Most of these occur with boundary tones due to the scarcity of intermediate phrases which are not also full intonational phrases. Nineteen percent of words / 69% of full intonational phrases are marked with L%, while 8% of words / 31% of intonational phrases have H%. Thus we see that in order to obtain a prosodically-rich corpus, it is necessary to analyze the sentences which provided instances of H- and H%, and to design future scripts to include enough sentences of these types. However, the current script has reasonable quantities of these labels to work with for now.

3.2. Perplexity

Perplexity serves as a measure of the branching factor or uncertainty of the next label given the history of labels seen so far; it represents one facet of the difficulty of a recognition task. For example, in the absence of occurrence statistics for break indices, we might assign break-index recognition a perplexity of 11, because the 11 observed labels might be presumed equally likely. Table 3 shows how bigram and trigram models derived from 90% of this corpus reduce perplexity on the held-out 10%. We observe that a bigram model, in which we seek statistics for what label to expect simply based on the one preceding label, provides a substantial reduction in the perplexity, regardless of

Table 3: *Perplexity for recognizing ToBI labels based on no occurrence statistics (therefore presuming all occur equally), bigram statistics, and trigram statistics, considering the break-index and tone tiers separately and jointly.*

	No statistics	Bigram	Trigram
Break-index tier	11	2.69	2.62
Tone tier	23	3.89	3.67
Both tiers jointly	34	3.54	3.23

Table 4: *Percentages of word occurrences marked with each ToBI tonal element, for yes-no questions, 90 sentences / 1629 words.*

Tone	%	Tone	%
none	51.5	L*+!H	1.6
H*	13.8	L+!H*	1.2
L+H*	10.1	*?	1.0
L-L%	9.4	L-	0.8
!H*	8.2	!H-L%	0.7
L*	6.2	!H-	0.3
H-H%	6.1	%H	0.3
L*+H	5.3	X*?	0.1
L-H%	4.1	H+!H*	0.1
H-L%	3.7	!H-H%	0.1
H-	2.1		

whether we treat the break-index and tone tiers separately or jointly. As a result, we anticipate that such a statistical model should substantially aid automatic prosodic labeling by greatly reducing the branching factor during recognition.

3.3. Analysis by Sentence Type

We subdivided the corpus into five categories: yes-no questions, either-or questions, other questions (hereafter, “wh-questions”), exclamations, and other declarative sentences. Tables 4, 5, and 6 break down tonal element occurrence data according to these categories, omitting declarative because it is the overwhelming majority of the overall data listed above, and either-or because only five sentences were classified as such.

Table 5: *Percentages of word occurrences marked with each ToBI tonal element, for wh-questions, 74 sentences / 777 words.*

Tone	%	Tone	%
none	40.9	L+!H*	2.1
H*	14.0	H-	1.8
L+H*	12.2	L-	1.5
L-L%	11.5	*?	0.8
L*	6.6	%H	0.5
!H*	6.3	!H-L%	0.5
L*+!H	5.3	!H-	0.4
L*+H	4.8	H+!H*	0.3
L-H%	4.6	!H-H%	0.3
H-H%	4.5	%r	0.1
H-L%	4.1		

Table 6: *Percentages of word occurrences marked with each ToBI tonal element, for exclamations, 25 sentences / 515 words.*

Tone	%	Tone	%
none	48.0	L+!H*	1.6
H*	16.7	!H-L%	1.2
L-L%	13.4	L*+!H	1.0
L+H*	11.3	H-	0.8
!H*	9.3	!H-	0.8
L-H%	6.2	H-H%	0.6
L*+H	2.9	H+!H*	0.6
L*	2.9	*?	0.2
H-L%	2.5	%H	0.2
L-	2.3		

Compared to a largely declarative corpus, in yes-no questions H-H% becomes more common and L-L% less, as expected; however, here we find that H-H% still hasn’t caught up with L-L%. In fact, only 43% of yes-no boundary tones are H%, not much more than the 31% in the corpus overall; we attribute this to the length of the yes-no questions, averaging more than 18 words per sentence. With 24% of words receiving boundary tones, we find an average of 4.5 full intonational phrases per question, only one of which is necessarily a questioning phrase, explaining why H-H% does not account for a majority of phrase endings. Similarly, phrase accents shift toward H- but not as decisively as one might have originally expected; 48% of phrase accents are H- or !H-, compared to 16% for the corpus overall.

The prosodic patterns of wh-questions appear to be somewhat intermediate between those of yes-no questions and the corpus overall. 37% of boundary tones are H%, and 40% of phrase accents are H- or !H-. The pitch accent distribution of wh-questions is very similar to that of yes-no questions, with 35% of words marked with a version of H* and 17% L*.

Statistics for exclamations are remarkably similar to the overall corpus, with 72% of boundary tones being L%, 82% of phrase accents being L-, and 7% of words getting versions of L* and 39%, H*. We believe the primary difference between exclamations and other declarative sentences, is in the pitch range, as represented by HiF0, which is not analyzed in this study.

3.4. Analysis by Word Frequency

As mentioned above, when preparing a corpus for concatenative speech synthesis, the phonetic sequences are nearly known, while the prosodic labeling is a recognition task with unknown labels. However, it is reasonable to expect that knowledge of the words may help skew the likelihoods of the prosodic elements. For example, one would expect that, on average, rare words would be more likely to be accented than common words. Toward such future automatic recognition of prosodic labels, we analyze the present corpus in terms of the frequencies of occurrence of words receiving various tonal labels.

To analyze word frequency, we took occurrence statistics from 600 million words of text gathered from news, office correspondence, medical and legal documents, and other sources. The top 260,000 words were rank-ordered, and counts were converted into probabilities by dividing each by the sum of all words’ counts. Remaining words were deemed too rare to be meaningfully analyzed using this text, and so were assigned probability zero and rank 260,000. Each word in the speech corpus was then tagged with its occurrence rank and probability. Then, for each tonal element, average and standard deviation of probability and rank were computed for the word occurrences

Table 7: *Probability and rank-order statistics for the word occurrences labeled with each tonal element. Columns show tonal element ("tone"), number of distinct words marked with that tone ("DW"), total number of word occurrences marked with that tone ("Occ."), average and standard deviation of 10,000 × word probability for those word occurrences, and average and standard deviation of word-frequency rank divided by 100 for those word occurrences.*

			Prob.×10,000		Rank / 100	
Tone	DW	Occ.	Avg.	S.D.	Avg.	S.D.
none	1104	15951	134	158	5	46
H*	3341	7245	6	20	76	246
L-L%	3596	7112	4	17	98	289
L+H*	2851	6447	8	21	84	274
!H*	2764	4883	4	16	81	265
L-H%	1977	3433	4	16	88	269
L*	1261	2035	6	19	54	198
L*+H	819	1267	7	23	85	266
L+!H*	836	1073	3	8	82	222
H-L%	643	848	6	19	77	238
L-	588	771	10	46	76	240
H-	527	771	5	11	59	179
L*+!H	400	530	3	10	79	225
*?	241	319	10	27	32	110
!H-L%	202	234	8	23	79	275
H-H%	137	177	8	26	170	519
!H-	164	177	3	6	54	106
H+!H*	152	164	3	8	120	366
%H	67	155	66	89	10	42
!H-H%	21	23	12	39	66	147
X*?	14	14	4	5	59	174
-X?	1	1	.02	0	418	0
%r	1	1	0.3	0	30	0

bearing that label, counting each distinct word each time it occurred with that label in the corpus. Results are in Table 7.

As can be seen, a word occurrence with no accent is, on average, a word accounting for 1.34% of the words in a large text; that is, a common word. As expected, this is far more common than words which have any of the frequently-occurring tones, as shown by the upper portion of the average probability column, although it appears a few common words get each type of accent, as evidenced by the high standard deviations of probability. This is reasonable, as common words on occasion will get, *e.g.*, contrastive emphasis. The average rank of unaccented words is 501 (shown in table as rank/100 = 5), indicating that more than just function words are spoken unaccented. However, average-rank statistics also show a large disparity between the unaccented category and all accent categories, though again we note the large standard deviations. Additional strong patterns are not evident in this data set, suggesting that more-detailed analysis is needed.

4. Conclusions

A professional speaker produces a rich variety of prosodic events, as categorized by a system like ToBI. But ToBI labels occur with very skewed frequencies, as expected, as a trigram "language model" of occurrence statistics has a perplexity of 3.23 on the 34 labels observed. This result suggests that such statistics will likely help in the difficult task of automatic recognition of ToBI labels. Significant patterns relate ToBI label occurrences to sentence type and word frequency. However, such associations must be modeled carefully, because of issues such as the majority of phrases in questions not necessarily being sentence-final or exhibiting a questioning pattern. Modeled appropriately, the observed associations imply that prosodic-label recognition will benefit from use of text information when it is available, for example, when labeling a corpus to prepare it for use in concatenative speech synthesis. Such future recognition capability should facilitate creation of large prosodically-annotated speech corpora, enabling high-quality concatenative speech synthesis with a rich variety of expressions.

5. Future Work

Performing a study like this using average speakers and additional professional speakers for comparison would clearly be desirable. Doing so would enable exploring whether the prosody of a professional is so much different than that of an average speaker that, even stopping short of speaker-dependent models, it would be worth developing separate models for professional and average speakers for automatic recognition of prosody. Unfortunately, the high cost of ToBI labeling makes such studies difficult to undertake. We also hope to expand the current analysis to relationships among occurrences of particular ToBI elements.

6. Acknowledgments

The author appreciates the assistance of Andy Aaron, who collected the corpus, and Ellen Eide, who administered the transcription and provided suggestions improving this work, and Michael Picheny for his suggestions.

7. References

[1] Conkie, A.; Riccardi, G.; Rose, R. C., 1999. Prosody Recognition from Speech Utterances using Acoustic and Linguistic Based Models of Prosodic Events. In *Proceedings of Eurospeech, Budapest, Hungary*.

[2] Pitrelli, J. F.; Beckman, M. E.; Hirschberg, J., 1994. Evaluation of Prosodic Transcription Labeling Reliability in the ToBI Framework. In *Proceedings of ICSLP, Yokohama, Japan*, 123-126.

[3] Pitrelli, J. F.; Eide, E. M., 2003. Expressive Speech Synthesis Using American English ToBI: Questions and Contrastive Emphasis. In *Proceedings of IEEE ASRU, St. Thomas, U.S. Virgin Islands*.

[4] Silverman, K.; Beckman, M.; Pitrelli, J.; Ostendorf, M.; Wightman, C.; Price, P.; Pierrehumbert, J.; Hirschberg, J., 1992. TOBI: A Standard for Labeling English Prosody. In *Proceedings of ICSLP, Banff, Alberta, Canada*, 867-870.

[5] Syrdal, A. K.; McGory, J., 2000. Inter-Transcriber Reliability of ToBI Prosodic Labeling. In *Proceedings of ICSLP, Beijing, P.R.C.*

[6] Syrdal, A. K.; Hirschberg, J.; McGory, J.; Beckman, M., 2001. Automatic ToBI Prediction and Alignment to Speed Manual Labeling of Prosody. *Speech Communication*, 33(1-2), 135-151.

[7] http://www.ling.ohio-state.edu/˜tobi

[8] Wightman, C. W., 2002. ToBI or Not ToBI. In *Proceedings of Speech Prosody, Aix-en-Provence, France*.