Auto-ToBI: a prosody-labelling workbench for Japanese Read Speech *

Nick Campbell
ATR Interpreting Telecommunications Research Laboratories

1 Introduction

This paper describes a prosody-labelling workbench that provides a first-pass automatic ToBI labelling for Japanese utterances, leaving the manual labeller the simpler task of checking and making fine adjustments. The implementation uses ATR's CHATR synthesiser to estimate a contour for several possible renditions of the utterance and selects the optimal one by comparison with the observed f0. The resulting prosodic labelling and f0 contour are then displayed using Entropic's Xwaves software. By thus allowing comparison between the predicted and observed contours, and between synthesised and original utterances, it assists the work of prosodic labelling considerably and allows us to create very large source databases for the synthesis of natural-sounding speech.

2 Methods and Data

A program was written that performs J-ToBI labelling, given a speech waveform and a representation of the orthography of each utterance as input. The program uses text-to-speech technology to predict a phone sequence for each utterance, and aligns it using speech recognition technology to determine an optimal alignment of the phone sequence to the speech waveform. It then extracts the fundamental frequency contour for each utterance and, using the text and segmental durations derived from the alignment, in conjunction with the intonation module of the synthesiser, predicts a series of candidate intonation contours from which the closest match is determined by comparison with the original. The contours are predicted iteratively according to the most likely label sequences and the one that is closest to the observed contour determines the optimal labelling to be assigned to the utterance. Because in the case of Japanese, the break indices need only be predicted at accentual phrase ($\approx$ bunsetu) boundaries, and the initial (default) accentuation can be predicted from the lexicon by the synthesiser, the number of contours to be generated is small enough to make this iterative analysis-by-synthesis possible.

To test the system, the ATR B-Set 503 sentences were hand-labelled in accordance with the J-ToBI prosodic labelling conventions [2] by two labellers. A subset of 50 of these utterances was jointly labelled as a check on consistency. The B-Set utterances were then auto-aligned as described above.

*自動 ToBI：日本語朗読発声用の韻律ラベリング・ワークベンチ
ニック キャンベル, ATR 音声翻訳通信研究所

Table 1 Number of labels per class

human-human							
a	b	c1	c2	d1	d2	e	f
637	404	2	3	20	10	28	7
57%	36%	0%	0%	2%	1%	3%	1%
human-machine							
a	b	c1	c2	d1	d2	e	f
816	6727	435	9	435	33	3346	44
6%	57%	4%	0%	4%	1%	28%	1%

Table 2 human-human break agreement

	.	2	2-	3	3-	4
.	1	5	4	.	.	2
2	.	119	.	.	1	1
2-	.	5	1	1	.	.
3	.	.	.	90	.	.
3-	.	.	.	3	2	.
4	.	.	.	1	.	47

(a dot indicates a missing label)

3 Results

Results are presented here that first compare the human-human transcription consistency and then show the degree to which the automatic labelling compares with these results. The reader is referred to [1] for a brief introduction to the J-ToBI transcription system, and to [2] for a more detailed explanation. No further details of J-ToBI labelling will be presented here.

In comparing two prosodic transcriptions, we have not only to account for insertions, deletions, and matches, but also to include a measure of the accuracy of the matches by showing the time difference between similar labels assigned to a given event. Since the purpose of this labelling is to enable extraction of information about the intonational characteristics of each utterance, a significant difference in the timings assigned to the labels can result in a different value for the retrieved fundamental frequency around the point of interest.

We distinguished the following criteria for measurement (see Table 1 for counts): a) exact match (same label and exact timing at the centisecond level. b) approximate match (same label and timing within 10 csec). c) missed label (by labeller A or B). d) inserted label (by labeller A or B). e) same label sequence but with very different timing. f) exact time alignment but different label.

Because of the time it takes a human labeller to do a ToBI transcription, we limited the human-human comparison to a subset of the first 50 utterances, labelled in common, after which the labellers shared the remaining 453 between them. The machine-human comparison is performed between all 503 utterances.

Table 3 human-human tonal agreement

	%L	%wL	≺	H*L	H-	L%	wL%
%L	54	.	.	.	.	.	.
%wL	1	45	.	.	.	.	.
*?	.	.	.	1	1	.	.
≺	.	.	11	4	.	.	.
H*L	.	.	2	197	4	.	.
H-	.	.	.	6	156	.	.
L%	.	.	.	.	.	245	.
wL%	.	1	.	.	.	3	73
.	.	.	2	7	3	5	6

4 Discussion

We see a very high degree of conformity in the human transcriptions (Tables 2 & 3), which indicates either that there is little room under the present transcription system for individual interpretation, or that the reading style of these texts is so uniform that there is little prosodic variation of interest to mark. Perhaps both are true.

Where there is a difference in interpretation (showing freedom in the transcription system) is primarily in the assignment of break indices, reflecting the decision whether an accentual phrase is 'complete in itself' or 'part of a greater unit'. This decided, the rest of the transcription is almost by rule; with really only two decisions left for the labeller to make:

a) whether or not a prescribed accent is fully realised in the utterance (as shown by *'*?'* marking uncertainty),

and b) whether or not the 'elbow' in the intonation contour is located clearly on the prescribed mora (indicated by '≺' on the actual point of descent).

We can see from Table 4 that the automatic transcription seems to be quite effective at distinguishing

Table 4 human-machine break agreement

	.	2	3	4
.	.	.	1	.
2	69	1189	208	26
2-	1	10	15	.
3	106	274	713	9
3-	.	22	4	1
3m	.	.	3	.
4	71	69	189	416

(column=machine, row=human)

Table 5 human-machine tonal agreement

	.	%L	H*L	H-	L%	wL%
%L	103	.	.	.	344	59
%wL	101	.	.	.	329	79
*?	5	.	2	2	.	.
≺	60	.	62	35	.	.
H*L	264	.	1794	171	.	.
H-	167	.	141	1548	.	.
L%	435	3	.	.	1964	269
wL%	126	1	.	.	446	243

(column=machine, row=human)

break index 2 from break index 3, which is perhaps the most important labelling decision.

With respect to the automated transcription, because much of the subsequent labelling is done by rule, there is less freedom; however, two differences are immediately obvious from Table 5: a) there is no marking of the phrase-initial tone (%L). This reveals a fault in the program, which is not sensitive to pauses in the utterance and therefore defaults to the end of the previous accentual phrase as the beginning of the next (and can be easily remedied), and b) that there is no sensitivity to differences in accent alignment (≺). This is a more serious problem that perhaps requires human intervention in a post-processing stage.

In general, alignment agreement was good, with median differences at 2 csec (25th and 75th percentiles: -7 csec, and 6 csec respectively), however, for the case of phrase-final tone markers (L%) it was noted that they are consistently being placed too early (on average by 11 csec). The reason for this is not yet obvious, but may be a consequence of phrase-final devoicing.

5 Conclusion

We have presented a system for the initial labelling of prosodic labels for clearly-spoken Tokyo-style Japanese. While there is good agreement with human-labelled speech, it is clear that a post-processing stage will still be required. Because the automatic system makes use of a speech synthesiser for predicting the optimal contours, further use of the synthesiser can also be made in audio-checking of transcription results, and so looping between an initial auto-transcription and audio-assisted 'polishing', allowing the labeller to not only see the speech and fundamental frequency waveforms, but also to listen to the results of his/her labelling.

Bibliography

[1] "J-ToBI: intonation labelling system for Japanese" W. N. Campbell & J. J. Venditti, ASJ Fall '95.

[2] "The J-ToBI Labelling Guidelines", J. J. Venditti, Ohio-State University Tech Rept, 1995.