AUTOLABELLING JAPANESE TOBI

Nick Campbell

ATR Interpreting Telecommunications Research Laboratories,
2-2 Hikaridai, Seika-cho, Kyoto 619-02, Japan. nick@itl.atr.co.jp

ABSTRACT

Focussing on the prosodic labelling of Japanese as an example, this paper describes the application of speech synthesis technology in a variety of speech processing tasks. It discusses first the use of synthesised utterances in the forced alignment and segmentation of a speech corpus, then the use of generated prosodic contours to determine the prosodic phrasing of an utterance, and finally the comparison with speech resynthesised using the prosodic transcription of the original utterance in order to check the transcription. It closes with an analysis of results from an auto-transcription of Japanese ToBI, and discusses some limitations of the proposed J-ToBI system.

1. INTRODUCTION

This paper describes a process for extracting prosodic information from a speech waveform. It has applications both in the labelling of prosody, and in the preparation of a corpus of speech for use in a high-definition concatenative speech synthesiser. The work is being tested on several languages, including English and Korean, but we limit our discussion here to the simpler case of Japanese. We discuss some details of the labelling of Japanese prosody and suggest some limitations of a recently proposed transcription system.

1.1. High Definition Speech Synthesis

For the synthesis of really natural-sounding speech, we need to prepare large databases of source units, preferably containing examples of every phone of the language, in each prosodic context, from continuous natural speech. By maximising the variety of the source database, we minimise the signal processing required at the waveform concatenation stage and ensure that all the fine detail of voice quality variation is preserved in the output synthesis.

The production of synthetic speech is thus reduced to i) the indexing of a corpus, and ii) the indentification within that corpus of a sequence of appropriate waveform segments that a) together form the desired utterance b) match the prosodic target specification, and c) join smoothly with each other. The indexing is performed as an off-line pre-process, allowing the retrieval and concatenation of segments to be done in less than real-time on a medium-power workstation.

Our present system [2, 3] [SOUND A897S01.WAV] uses z-score normalised values of pitch, power, and duration to store the prosodic variation of each phone in the database but, as more prosodically labelled corpora are becoming available, we are also testing the validity of using higher-level phonological labels to encode the prosodic variation directly for unit selection. Phonological encoding of prosody is also important for training the modules within the synthesiser that predict the target pitch, power, and duration for text-to-speech and concept-to-speech pre-processing.

1.2. Aligning the corpus

In order make use of the prosodic parameters extracted from a speech signal, it is necessary to associate them with linguistic features of the speech by aligning the waveform with phone labels describing the content of the utterance. If this phone sequence is not available it must be generated. The minimal requirements for processing a speech corpus are thus a) a representation of the speech waveform, and b) an orthographic representation of the words it contains. If the latter is not available, then it must be produced manually. Speech recognition technology can be of assistance at this stage, but is not yet robust enough to produce a reliable transcription unaided.

Given the words, we can then use the grapheme-to-phoneme component of the speech synthesiser to produce a sequence of phones. By mapping the synthesiser-generated phone sequence to the speech waveform, we can ensure that its labels will match the phone set of the synthesiser for the language, and can then use prosodic and contextual information to distinguish the allophonic variants of each phone type under a range of different utterance contexts.

The synthesiser is also used in the alignment of the phone sequence to the speech waveform. By choosing the closest speaker of the same language from the synthesis database, we generate an equivalent utterance having the same sequence of phones, and produce a waveform equivalent for the utterance being labelled. Because the waveform is generated by concatenation of raw waveform segments, typically each of the size of a single phone, it preserves the fine detail of acoustic characteristics at segmental boundaries and serves as a reference waveform from which to perform DTW matching. As each portion of the synthesised speech is DTW-aligned with its equivalent waveform segment in the original speech, the boundary information can be mapped from one to the other and the phone durations adjusted. When the optimal mapping has been determined between the two waveforms, the phone label information can be carried across from the synthesised version to identify each segment of the original.

Because the prosodic information is not of interest at this stage, it does not matter that the speech in the two waveforms may actually sound quite different; similarity of the acoustic sequence in the spectral domain is of primary concern at this stage. It is important, however, to ensure that any pause in the corpus speech is matched by a silence in the synthesised speech, and to pay particular attention to any portions of the speech that contain disfluencies, coughs, or non-speech noise.

1.3. Determining prosodic phrasing

Given a segmental alignment of the speech waveform, the next task is to determine its prosodic phrasing and to identify the accentuation patterns within each phrase. Since the phrasing of an utterance has a strong influence on its fundamental frequency contour, we can use an inverse mapping from the observed F_0 as a guide to the phrasing.

The synthesiser contains modules for the prediction of an F_0 contour by rule, from the text of an input sentence. The shape of a natural F_0 contour is dependent on a number of factors, including speaking style and the illocutionary force of the utterance, but is also correlated with the syntactic and semantic bracketing. This structural and more easily derived information is therefore used in the prediction of a synthetic F_0 contour for matching. By generating several contours, one for each possible semantic phrasing of the utterance, and then comparing each with the observed F_0 contour of the original, we can determine the closest and by implication the most likely phrasing of that utterance.

2. JAPANESE TOBI

Japanese ToBI (J-ToBI) [6] was proposed as an extension of the ToBI (Tones and Break Indices) prosodic labelling system originally designed for English [1]. It provides a systematic phonological transcription for recording the F_0 events and prosodic boundaries in Tokyo Japanese speech. Like its predecessor, it consists of four tiers:

Word Tier

The word tier contains the romanised transcription of the words of utterance. A minimal dictionary entry is used as the working definition of a "word", and as such we mark postpositions and particles as separate words. Accented words are marked with an apostrophe (') after the vowel of the relevant mora.

Tone Tier

The tone tier in J-ToBI marks the distinctive pitch events in the F_0 contour, and is consistent with the work on Japanese intonation by Beckman and Pierrehumbert [5]. The following is a list of the core labels in this tier.

- H*+L bitonal pitch accent marking the lexically specified accent of accented phrases.
- H- phrasal tone marking the high F_0 of unaccented phrases, also used in some accented phrases in which the phrasal H is higher than the accentual H. It is one of the two tones that delimit the accentual phrase (most commonly bunsetsu) in Japanese (see break index 2 below). In Tokyo Japanese, this tone usually occurs on the second mora of the phrase.
- L% Along with the phrasal H-, this final low boundary tone characterizes the accentual phrase in Japanese. Together, these two tones produce the familiar rise-fall pattern of the accentual phrase. There is also a "weak" variant of this tone (wL%) used in cases where the next phrase begins with a long syllable, or is initially accented.
- %L initial low boundary tone marked at the beginning of post-pausal phrases. It provides an anchor from which the F_0 rises at the beginning of utterances and after pauses. As with the final low boundary tone, this initial tone also has a "weak" variant (w%L), used in the same contexts.
- H% final high boundary tone marking the final high rise common in interrogative utterances, and also in some declaratives.

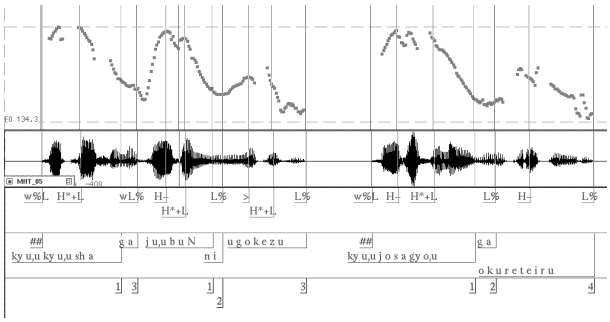

Figure 1: J-ToBI transcription of ATR-B-MHT-05

In addition to these core tones, the tone tier also includes labels for marking restarts after disfluent regions (%r) and uncertainty about the de-phrasing of accented words (*?).

Break Index Tier

Break indices are a measure of the degree of association between two sequential units. They indicate the prosodic grouping of words at various levels. These are measures of perceived juncture that have observable physical correlates, such as tonal markings and pre-boundary lengthening. J-ToBI distinguishes 5 degrees of disjuncture in the prosodic structure of Japanese.

- 0 break index marking junctures common in fast speech phenomenon, *e.g.,* /kore+wa/ → [korya].
- 1 marks the juncture between sequential words, and as such is the most common break index.
- 2 marks the juncture between prosodic units corresponding to the accentual phrase [5]. This unit is delimited by the rise-fall of the phrasal H- and L% boundary tones. It often consists of a noun plus following postposition (bunsetsu). However, it is also common to find two or more content words grouped together into a single accentual phrase, delimited by these two tones.
- 3 marks the boundary between successive intermediate (major) phrases. This is the domain in which the high-tone line is specified, and therefore at a break index 3 boundary, a new pitch range is chosen for the following phrase. The prosodic juncture marked by a 3 is stronger than that marked by a 2 (accentual phrase), but lacks the percept of "finality" which accompanies the stronger break index 4.
- 4 marks the boundary of an intonation phrase. It is a strong juncture marked by a sense of "finality". This may be cued by a variety of factors, including lowering, lengthening, or "final" contours. In read speech, break index 4 is reserved for the ends of utterances.

In addition to the 5 labels described above, the J-ToBI break index tier also contains labels for marking labeller uncertainty (e.g. 1-, 2-, 3-, 4-) about the strength of the boundary, and also labels for marking hesitations or other disfluencies (e.g. 1p, 2p, 3p) which often occur in spontaneous speech.

Miscellaneous Tier

This tier is used for other phenomena present in the speech signal which cannot be properly described by the phonological events marked in the tone and break index tiers. Such phenomena include repairs, disfluencies, laughing, etc.

Figure 1 shows a J-ToBI transcribed utterance of read speech [4] [SOUND A897S02.WAV]. This utterance is one

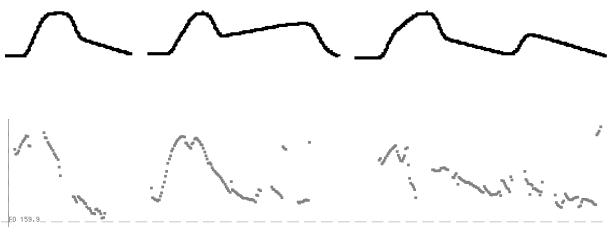

Figure 2: F_0 generated from chatr. The underlying contour is as specified by the labelling (top), but because chatr concatenates tokens of real speech, the micro-segmental perturbations are inherited from the selected tokens (bottom)

intonation phrase (Break Index 4), composed of three intermediate phrases (BI 3). The first intermediate phrase has only one accentual phrase (BI 2), while the second two have 2 accentual phrases each, with downstep applying. All but the last accentual phrase contain an accent, marked by the H*+L. The peak F_0 of the phrase [ugokezu] is not realized within the accented mora, and so the actual peak is marked using ">", an early F_0 event. The H- phrasal tone on the last unaccented phrase is also marked at the F_0 peak, disregarding segmental perturbations. The "weak" variant of the low boundary tone (w%L) is marked in cases where the following mora is a long syllable.

3. AUTO-TRANSCRIBING A CORPUS

A program was written to perform automatic J-ToBI labelling, given a speech waveform and a representation of the orthography of each utterance as input. The program uses the text-to-speech system components described above to predict a phone sequence for each utterance and 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. A direct implementation of the Algorithm for Phonetic Realisation ([5] Ch.7) is used to generate intonation contours in the synthesis. This module produces a representation of the underlying contour (Fig. 2, top) when given a sequence of ToBI labels as input. The F_0 contours are predicted iteratively according to the likely label sequences for the input sentence, and the one that is closest to the observed contour determines the optimal labelling to be assigned to the utterance. The initial (default) accentuation is predicted from the lexicon by the synthesiser. In read Japanese, BI 0 rarely occurs, BI 1 is redundant, and BI 4 is required at end-of-sentence; only BI 2 & BI 3 need be predicted/tested at sentence-internal accentual phrase boundaries. The rest of the (tonal) transcription can then be produced by rule. Thus the number of contours to be generated is small enough to make this iterative analysis-by-synthesis practical.

Table 1: Number of labels per class

a: 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%
b: 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%

key: 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 1 or 2) required by the grammar. d) inserted label (by labeller 1 or 2) not required. e) same label sequence but with very different timing. f) exact time alignment but different label.

Table 2: Agreement for hand labels (n=50)

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

tones	%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

Table 3: human-machine labelling agreement

breaks	.	2	3	4
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

tones	.	%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

(n=503, column: machine, row: human)

To test the system, the 503 sentences read by a professional announcer [4] were hand-labelled in accordance with the J-ToBI prosodic labelling conventions by two labellers. A subset of 50 of these utterances were jointly labelled as a check on transcription consistency. The utterances were then auto-aligned as described above.

4. RESULTS

Results are presented here that first compare the hand-labelled transcription consistency and then show the degree to which the automatic labelling agrees.

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 (see Table 1). 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 fundamental frequency around the point of interest.

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.

5. DISCUSSION

We see a very high degree of uniformity in the hand-labelled transcriptions (Table 2), which indicates either that there is little room under the present system for individual interpretation of fine variation, 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.

Because of the constrained nature of Japanese intonation, once the phrasing has been decided, and the accentuation known (derived by rules from the lexicon), the rest of the transcription can be almost done by rule; with really only two important 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 the confusion matrix in Table 3 that the automatic transcription seems to be quite effective at distinguishing break index 2 from break index 3, which is perhaps the most important labelling decision. However, two differences are immediately obvious from Table 3: 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 ($\prec$). 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 (see 'e' in Table 1b)). The reason for this is not yet obvious, but may be a consequence of phrase-final devoicing, or resulting from poor F_0 tracking in these regions.

6. LISTENING TO THE TRANSCRIPTION

Finally, perhaps the greatest advangtage of having a synthesiser included as part of the labelling process is that it allows the human labeller to listen to the results of the labelling and to work interactively, checking the transcription by deciding not just whether the labels are appropriate according to a prescribed syntax, but by listening to see if they produce a functionally equivalent rendering of the utterance in the synthesised speech. [SOUND A897S03.WAV]

Because the synthesiser uses a fundamental-frequency prediction module that takes the same transcription labels as input, the labelling conventions themselves can be tested at the same time. If by varying the transcription, the labeller is not able to produce an equivalent utterance interactively, then either the labels are inadequate, or the synthesiser prediction module itself needs improvement. In informal tests labelling Osaka Japanese, we have found that several changes to the proposed J-ToBI standards will be required.

7. CONCLUSION

We have presented a system for the automatic labelling of prosodic features in read speech, taking as input a recorded waveform and an orthographical representation of its contents. The system has been tested with clearly-spoken Tokyo-style Japanese using the proposed J-ToBI connventions. While there is good agreement with hand-labelled trasncriptions of the same corpus, it is not as high as that between two human transcribers, and it is clear that a post-processing stage will still be required. Extensive use is made of speech synthesis technology in this labelling, both in the prediction and alignment of a phone sequence, and in the generation of fundamental frequency contours for matching with the original. The synthesiser is also useful in audio-checking of transcription results, facilitating looping between an initial auto-transcription and audio-assisted 'polishing', allowing the labeller to not only see the speech and visually compare fundamental frequency waveforms, but also to listen to the results of his/her labelling for a functional comparison of the transcription.

It would seem that the main (and perhaps only) decision to be made when labelling an utterance of read speech in Tokyo Japanese is whether each sentence-internal phrase boundary is Break-index 2 or 3. That decided, then everything else follows according to the transcription syntax or can be predicted from the lexicon. If so much is predictable by rule, then it is surely against the basic ToBI principles to require it to be labelled explicitly. Alternatively, we should ask what other information is present in the speech that is NOT being labelled, and that could not be predicted by rule. J-ToBI seems inadequate for the labelling of prominence, focus, and speaking style, and it is these that convey the interpretation of an utterance that is carried by its prosody. The functional aspects of prosody can not be predicted by rule from the text, but make use of the freedom for individual expression. Future work will be required to see whether we can differenciate what can be predicted from what has been observed in order to interpret the more subtle signals in the spoken utterance.

REFERENCES

[1] Beckman, M. E. & Ayers, G. M., "The ToBI Handbook.", Technical Report, Ohio-State University, U.S.A. 1993.

[2] W. N. Campbell. Synthesis units for natural English speech. Technical Report SP 91-129, IEICE, 1992.

[3] W. N. Campbell and A. W. Black. Prosody and the selection of source units for concatenative synthesis. In J. van Santen, R. Sproat, J. Olive, and J. Hirshberg, editors, *Progress in Speech Synthesis*. Springer Verlag, 1996.

[4] Ohta, Y., & Campbell, W. N., "Labeling of Prosodic Structure in Japanese,", ATR Technical Report TR-IT-0062, 1995.

[5] Pierrehumbert, J., & Beckman, M. **Japanese Tone Structure.** Cambridge, MA: MIT Press. 1988.

[6] Venditti, J. J., "Japanese ToBI Labelling Guidelines", Technical Report, Ohio-State University, U.S.A. 1995.