PERCEPTUALLY BASED AUTOMATIC PROSODY LABELING AND PROSODICALLY ENRICHED UNIT SELECTION IMPROVE CONCATENATIVE TEXT-TO-SPEECH SYNTHESIS

Colin W. Wightman[1]   *Ann K. Syrdal*   *Georg Stemmer*[2]   *Alistair Conkie*

Mark Beutnagel

AT&T Labs - Research, Florham Park, NJ, U.S.A.

[1]also Dept. of Computer and Information Sciences, Minnesota State University, Mankato, MN, U.S.A.

[2]also Informatik Dept., University of Erlangen, Erlangen, Germany

ABSTRACT

Prosody is an important factor in the quality of text-to-speech (TTS) synthesis. Typically, acoustic parameters such as f_0 and duration are the only variables related to prosody that are used to determine unit selection. Our study explored adding the explicit use of linguistically and perceptually motivated prosodic categories in unit selection-based TTS. One of our goals was to automate the process of prosodically labeling our TTS inventory. However, reliability among labelers for some ToBI[6] (Tones and Break Indices) categories was too low[9] for successful training of an automatic prosody recognizer. We developed a prosody labeling system simpler and more robust than standard EToBI (English ToBI). This "ToBI Lite" system was used successfully for automatic labeling of the acoustic inventory and in prosodically enriched unit selection. A formal listening test was conducted to compare subjective quality ratings for several variations of the AT&T unit selection concatenative TTS system that differed only in their method of prosodic labeling of the inventory or their use of prosody for unit selection. The use of simple prosodic categories in unit selection significantly improved ratings, and automatic prosodic labeling resulted in higher ratings than manual labeling.

1. INTRODUCTION

The relationship between acoustic properties and perceived prosody is not yet well understood, but it is nevertheless critical for natural sounding synthetic speech. Our study explored the use of perceptually motivated prosodic categories in unit selection-based TTS.

2. ROBUST ToBI LITE

One of our goals was to automate the process of prosodic labeling the TTS inventory. However, the reliability among labelers for some EToBI categories[9] was too low for successful training of an automatic prosody recognizer using the full EToBI system. Another problem was a sparse data set for some rarely occurring prosodic events.

Transcriber agreement[9] was high (> 50%) for only two to four of eight pitch accent types (which mark syllable prominence) and for three of nine edge tone types (which mark prosodic phrase boundaries). Perceptual judgments of syllable prominence and phrase boundaries, together with inter-transcriber reliability results[9], provided guidelines for developing a prosody labeling system that is simpler, more robust, and easier to use[11] than standard EToBI.

2.1. Perceptual Prominence and EToBI

Simple scalar ratings by groups of non-expert listeners have been used by several speech researchers to estimate for various languages the relative auditory prominence in continuous speech of syllables[3][5] and of phrase boundaries[2]. Ratings were found to be highly consistent among listeners, and correlations with various acoustic parameters related to prosody were significant.

A small scale perceptual experiment was run in order to compare listeners' ratings of the perceptual prominence of syllables and phrase boundaries for the same recorded speech material that was independently labeled by four ToBI labelers[9]. The 644-word speech database was recorded from one female professional speaker reading text representing different prosodic styles. Three adult native speakers of American English served as listeners and independently rated syllable prominence (using a 3-point scale) and phrase boundaries (using a binary judgment – presence versus absence of phrase boundary). Listeners made judgments on the basis of auditory cues only. They navigated through the speech file by means of an interactive waveform display with attached time-aligned label files containing words and syllable boundaries, with a blank tier for the listeners to enter their ratings.

Figure 1 combines results from the current perceptual prominence study with results from the separate ToBI reliability study[9] that used four different individuals to independently ToBI label the same speech data. It displays the perceptual ratings given to syllables that previously had been assigned ToBI tonal categories. Ratings for each of the three listeners are plotted individually, along with the group mean.
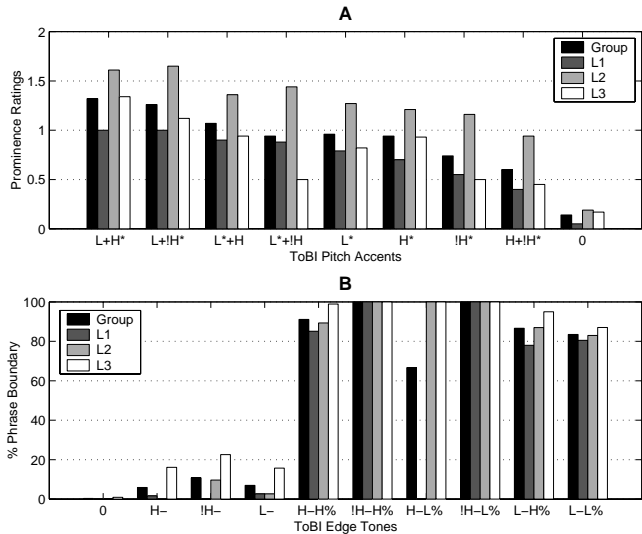
Figure 1: Perceptual Prominence Ratings for Pitch Accents (A) and Percent of Phrase Boundaries Perceived for Edge Tones (B). The group mean ratings (black bar) and ratings for individual labelers (L1, L2, L3) are shown.

Panel A shows syllable prominence ratings (on a scale of 0 to 2) received by syllables previously classified according to the eight different ToBI pitch accents plus the unaccented case ("0"). The modal perceptual ratings for the four bi-tonal pitch accents (L+H*, L+!H*, L*+H, and L*+!H) were higher than the ratings assigned to the other pitch accents. The prominence ratings for unaccented syllables were much lower than the ratings for the remaining pitch accent types. These results correspond well with previous psycholinguistic experiments on nuclear accent types and prominence[1]. For purposes of the present TTS experiment, ToBI pitch accents were clustered into two categories according to their associated perceptual prominence: ** for bi-tonals, * for other accents; "0" was assigned to unaccented syllables.

Panel B displays the percentage of perceived phrase boundaries associated with each of the nine previously assigned ToBI edge tones plus the unaccented case ("0"). It is evident that only major intonational phrases (marked by both a phrase accent (H-, !H-, or L-) and a boundary tone (H% or L%) were perceived reliably by listeners as phrase boundaries when making a binary choice. Consequently, for the present TTS experiment, only syllables marked by ToBI boundary tones were classified as phrase boundaries (%).

3. ToBI LITE RECOGNITION

This ToBI Lite system was used successfully for automatic labeling of the acoustic inventory and in prosodically enriched unit selection.

The automatic labeling algorithm[10] used to label the acoustic inventory utilizes a decision-tree based VQ that is de-

signed jointly with a HMM in which a single model state corresponds to each possible label. In this case, there were six possible labels (corresponding to the three levels of prominence on phrasal boundaries and on non-boundaries) and the underlying HMM was thus a fully-connected, 6-state HMM.

3.1. Training

Expert ToBI labelers prosodically transcribed several hours of a large TTS acoustic inventory spoken by a female professional speaker[8]. The training database of 860 utterances included read text representing different prosodic styles: business news, interactive prompts, and short laboratory sentences. The collapsed ToBI labels were used for training. That is, bi-tonal pitch accents were mapped to ** and other pitch accents were mapped to *, and only edge tones marking major phrases were mapped to %.

The input to the recognizer included automatic segmentation results and acoustic parameterization. Twenty-four linguistically motivated acoustic features were derived from the waveform and segmentation, and extracted at the syllable level. Some features were binary (e.g. stress, word-final, word-initial, schwa) and others were continuous (e.g. normalized duration, maximum/average pitch ratio). The desired output was perceptual labels for prominence and phrasing for each syllable, and thus a feature vector was generated for each syllable.

Maximum Mutual Information training of the VQ decision tree was done jointly with training of the HMM using the iterative method described by Wightman and Ostendorf[10], resulting a Maximum Likelihood design for the overall labeling model.

3.2. Testing

We first briefly describe performance results obtained for a test corpus. To objectively measure accuracy, the automatic ToBI Lite labels were compared to collapsed manual EToBI labels and also to manual perceptual (ToBI Lite) labels. We then focus on a listening test that compared subjective ratings of speech quality earned by experimental TTS systems that used differently labeled prosodic information explicitly for unit selection.

A test set of the 42 utterances used for both the ToBI reliability study and for the perceptual prominence study was held out from the training set. Syllable-level results on the test set were compared to collapsed manual EToBI labels and to perceptual ToBI Lite labels.

Compared to collapsed EToBI labels, accuracy for the extremes of the prominence/pitch accent scale was quite high: 83.5% of non-accented syllables and 84.9% of ** syllables were correctly recognized. Recognition accuracy for prominent (either ** or *) versus unaccented syllables was 69.3%, with a false alarm rate of 16.5%. Phrase boundaries

(%) were correctly recognized 93.4% of the time, with a false alarm rate of 2.0%.

Compared to perceptual ToBI Lite labels, 80.9% of non-prominent syllables and 84.8% of ** syllables were correctly recognized. Recognition accuracy for prominent (either ** or *) versus unaccented syllables was 76.2% correct, with a false alarm rate of 19.1%. Phrase boundaries were correctly recognized 93.0% of the time, with 1.6% false alarms.

4. ToBI LITE and UNIT SELECTION

Unit selection for synthesis is determined by a Viterbi search for the lowest cost path through a network of possible acoustic inventory units. The cost function is defined as the sum of target costs and concatenation costs. Concatenation costs estimate how smoothly the units in a sequence are perceived to join together. Target cost estimates the perceptual distance of a specific inventory unit from the desired target. Units are described by a feature vector, with features such as duration and f_0. For each target unit in the utterance to be synthesized, appropriate feature values are predicted. The target cost is calculated as the sum of weighted feature vector differences between the inventory unit and the target unit. Feature weights are trained during the creation of a TTS voice to optimize the mapping from feature vector differences between units to cepstral distances between units in the inventory. In the baseline TTS system used for this experiment, the only features related to prosody were duration, f_0, and syllable stress. These features are prosodically ambiguous, because there is a one-to-many mapping from them to prosodic structure.

5. PERCEPTUAL TEST

A formal listening test was conducted to determine whether or not the expansion of target cost feature vectors to include features explicitly related to prosodic categories improves perceived TTS quality. The test compared subjective quality ratings for several variations of an experimental AT&T unit selection TTS system that differed only in their method of prosodic labeling of the inventory or their use of prosody for unit selection.

5.1. Methods

Eight experimental TTS conditions were evaluated in the listening test:

BL (Baseline) TTS used the standard prosody module to predict prosody from standard punctuated text input, and did not use prosodic categories in unit selection. Except for the last TTS system listed below (BATS), this and the other experimental conditions used the same 80 minute acoustic inventory for synthesis.

BL+TA (Baseline with text annotated) TTS was identical to BL except that, instead of automatically predicted prosody, it used input text that was prosodically annotated to correspond with the ToBI labels assigned by transcribers to the naturally spoken test utterances. This was done to ensure more variety in the target prosody of the test utterances than would be predicted by TTS from text input alone. If the target prosody were too limited, there would be less opportunity to observe differences among the experimental TTS systems. All the remaining experimental TTS systems also used the same annotated text as input.

ASU (Accented/stressed/unstressed) TTS used three general prosodic categories of syllable prominence explicitly in unit selection: accented, stressed but unaccented, and unstressed. The presence of syllable accent was determined from manual ToBI labels.

HTL (Hand ToBI Lite) TTS used manually assigned ToBI labels collapsed and mapped into the simpler ToBI Lite format. Three levels of prominence and two categories for phrase boundary resulted in six possible combinations. In this and the other TTS conditions listed below, prosodic category and syllable stress were independent features used for determining target costs.

HTL+L (ToBI Lite plus L*) TTS used an expanded ToBI Lite model that included the L* pitch accent as a separate prominence category along with ** and *. It was separated because the falling pitch and lower f_0 range of L* distinguish it acoustically from the other accents with * perceptual prominence. HTL+L categories were mapped from manually assigned ToBI labels.

ATL (Auto ToBI-L) TTS used automatically labeled ToBI Lite categories in unit selection.

ATS (Auto ToBi-Lite and Segmentation) TTS used automatically labeled ToBI Lite categories in unit selection and an automatically phonetically segmented and aligned inventory.

BATS (Big ATS) TTS was like ATS but with an acoustic inventory five times larger (400 minutes). This condition took advantage of the speed of automatic labeling to label a much larger database than had been manually labeled.

Twelve test utterances were synthesized by each combination of TTS system and each of two methods of speech representation, HNM[7] and PSOLA[4]. Two different speech representations were used for greater generalizability. Half of the test utterances were interactive telephone service prompts, and half were sentences from business news articles. All were less than 10 seconds in duration, and all were present in the acoustic inventory. Recordings of the test utterances by the same female speaker used for the TTS inventory were included as a reference condition in the listening test. Stimuli were normalized for level, 40-6500 Hz bandpass filtered, and presented over calibrated headphones. There was a total of 204 test utterances in the listening test.

Listeners and Test Procedures: Forty-three listeners participated in the one-hour listening test. They were adult

American English speakers with no known speech or hearing deficits. They were tested in four groups of 10 or 11 listeners; each group was presented a different random order of test utterances. There were 8,772 observations in the experiment.

5.2. Results and Discussion

As shown in Figure 2, two TTS conditions, ATL and ASU, were rated significantly higher than the others. The fact that ATL was significantly better than HTL indicates that automatic ToBI Lite labeling was superior to ToBI Lite mappings from manual ToBI labeling. ASU's superiority over HTL and BL+TA indicates that the inclusion within a single feature set of even simple binary information about accent in addition to lexical stress led to a significant improvement. There was no significant increase in ratings when prosodically annotated text was used (BL+TA) rather than automatic TTS assignment (BL) to determine tone placement and type. The expansion of ToBI Lite to include L* accents as a separate category (HTL+L) did not improve perceived TTS quality. Although automatic prosodic labeling improved TTS quality over manual labeling, the automatic segmentation of ATS and BATS conditions clearly resulted in significantly poorer quality. This result was due to very low ratings for one or two test utterances, rather than to consistently lower scores. The listening test allowed the problems to be identified and subsequently remedied.

A repeated measures ANOVA was used to analyze the listening test data. There was a significant main effect of TTS condition ($F(7,294) = 12.147$, $p < 0.0001$). Newman Keuls tests with $\alpha = 0.05$ were performed to compare the TTS conditions.

6. CONCLUSIONS

The explicit inclusion of very simple but robust and perceptually salient prosodic classifications in features used for unit selection significantly improved perceived TTS quality. Automatic prosodic labeling resulted in significantly higher opinion scores than manual labeling. This somewhat surprising result probably in part reflects the greater consistency of automatic recognition techniques. Automation of prosodic labeling also provides a tremendous practical advantage through reducing by several orders of magnitude the time needed to develop a new synthetic voice.

7. REFERENCES

1. G. Ayers. *Nuclear Accent Types and Prominence: Some Psycholinguistic Experiments*. PhD thesis, Ohio State University, 1996.

2. J. dePijper and A. Sanderman. On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues. *J. Acoust. Soc. Amer.*, 96:2037–2047, 1994.
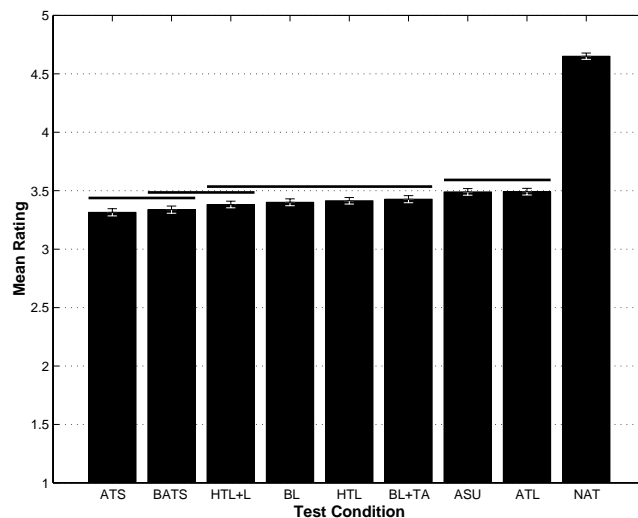
Figure 2: Mean Opinion Scores with standard error bars for the eight TTS conditions and the natural speech reference. Conditions whose ratings were not significantly different from each other are indicated by the same horizontal line above the bars.

3. G. Fant and A. Kruckenberg. Preliminaries to the study of Swedish prose reading and reading style. *STL-QPRS*, 2/89:1–83, 1989.

4. E. Moulines and F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9:453–467, 1990.

5. T. Portele and B. Heuft. Towards a prominence-based synthesis system. *Speech Communication*, 21:61–72, 1997.

6. K. Silverman, M. Beckman, J. Pierrehumbert, M. Ostendorf, C. Wightman, P. Price, and J. Hirschberg. ToBI: A standard scheme for labeling prosody. In *Proc. ICSLP*, pages 867–879, Banff, October 1992.

7. Y. Stylianou, T. Dutoit, and J. Schroeter. Diphones concatenation using a harmonic plus noise model of spe ech. *Proc. EUROSPEECH*, Sept. 1997.

8. A. K. Syrdal, J. Hirschberg, J. McGory, and M. Beckman. Automatic ToBI prediction and alignment to speed manual labeling of prosody. *Speech Communication*, (in press).

9. A. K. Syrdal and J. McGory. Inter-transcriber reliability of ToBI prosodic labeling. In *Proc. ICSLP*, Beijing, 2000.

10. C. W. Wightman and M. Ostendorf. Automatic labeling of prosodic patterns. *IEEE Trans. Speech and Audio Processing*, pages 469–481, October 1994.

11. C. W. Wightman and R. C. Rose. Evaluation of an efficient prosody labeling system for spontaneous speech utterances. In *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, Keystone, Colorado, December 1999.