COMMON PATTERNS IN WORD LEVEL PROSODY

Frode Holm and Kazue Hata
Speech Technology Laboratory
Panasonic Technologies, Inc.
Santa Barbara, CA 93105

ABSTRACT

The task of generating natural human-sounding prosody for text-to-speech (TTS) has historically been one of the most challenging problems that researchers and developers have had to face. TTS systems have in general become infamous for their "robotic" intonations. This paper describes an approach to this problem which endeavors to capture as much detail as possible from speech data, but in a way that avoids the "black boxes" typical of neural networks and some vector clustering algorithms. Unlike these latter methods, our approach may give feedback as to exactly what the crucial parameters are that determine the successful choice of pattern. Focusing on the notion of prosody templates, we confirmed that a representative F0 and duration pattern can be extracted based on stress pattern for a target proper noun which occurs in sentence-initial position.

1. INTRODUCTION

In this study, we are attempting to extract from a body of recorded speech, commonly occurring word level prosodic patterns of both F0 and duration. These templates should be general enough to be used with an arbitrary word to be synthesized, yet have a quality of naturalness and detail that is very close, if not identical to that of a human speaker. These goals are somewhat conflicting and the difficulty lies in striking a balance between them.

To this end, we have taken as our fundamental assumption that the syllable forms the most basic unit of prosody in spoken English. Furthermore, we assume that the stress pattern determines the most perceptually important characteristics of both F0 and duration [1, 2]. At this level of granularity the template set will be small in size and easily implemented for a target/frame-sentence type system. If our assumption should prove too weak to capture all important variations, we can expand the set to allow for more feature determiners both at the syllable and the word level. For instance, we suspect that certain types of patterns could be determined by microscopic F0 perturbations caused by consonant type, voicing, intrinsic pitch of vowels and segmental structure in a syllable, as pointed out by van Santen and Hirschberg [3].

In this paper we will report on common patterns found using only the stress-pattern as our grouping function. Indeed, we have already found that our basic assumption is true in most, but not all, stress groups, and these results will be detailed below. Due to space limitations, we will only report on 1-3 syllable words, and only F0 contours will be detailed.

2. DATA COLLECTION

Our methodology has its basis in the notion of prosody templates; the idea that a general prosodic pattern can be extracted from and later applied to a given linguistic context. The current aim is to investigate word-level prosody for various sentence positions for proper nouns in announcement-type, emotionally-least-loaded statements. The output will be initially used in the frame sentence and target word synthesis scheme.

For this study, we have recorded about 3,000 sentences with proper nouns in sentence-initial position. This database was collected from a single female speaker of American English.

The preprocessing of data involved several steps. The segmentation of sentences was done by using an in-house HMM-based automatic labeling tool, followed by a manual correction for each target word string. Then, the stress pattern for the target words was assigned by ear using three levels of stress: primary ('1'), secondary ('2') and no stress ('0'). Also, syllabification was manually done using a set of rules we developed for this project. For F0 samples, we first used an automatic tracker followed by an in-house semi-automatic F0 editing program to increase the accuracy of the tracking.

This preprocessing was time-consuming, but in our approach, accuracy was necessary in order to reduce the noise level in our statistical measures.

Finally, all the processed data was incorporated into a relational database, which was designed to allow us to sort, search and combine this information according to any feature and criterion we might choose. We have only barely begun to tap the power of this immensely useful tool.

3. NORMALIZATION AND STATISTICAL ANALYSIS

In order to be able to compare the F0 curves, we first performed a comprehensive normalization procedure designed to eliminate from the raw data as much irrelevant information as possible.

First, the curves were normalized with respect to time. Each segment of the curve belonging to a syllable, as determined by the aforementioned labeling step, was resampled to a fixed number of F0 points (30 per syllable). Thus, each syllable was given the same length on an abstract unit-less time line.

Next, we tried to eliminate as many of the arbitrary constant offsets in baseline pitch as possible. Quite often we observed that our speaker produced identical contours, but at different elevations. Our method consisted of first transforming the F0 points for the entire sentence to the log domain and then computing the linear regression line for that sentence [5, 6, 7]. The point at which the regression line intersects the word end-boundary was then used as the elevation point for that target word. It was subsequently shifted down (or up) to a common reference point at 100Hz.

After these transformations, we were now in a position to do some statistical measures on the samples. The essential question we wanted to ask was: how similar or different are they from each other? This is a deep issue that could warrant a research project in itself. However, we have initially taken a simple-minded approach that tabulates how far from the arithmetic mean each sample is. We have used a measure computing the area difference between two vectors (see equation below). Although this measure is by no means guaranteed to always yield perceptually valid results, we have found that it is usually quite good at producing the information we seek.

$$d(Y_i) = c\sqrt{\sum_{k=1}^{N} (y_{ik} - \overline{Y}_k)^2 v_{ik}}$$

d = measure of the difference between two vectors

i = index of vector being compared

Y_i = F0 contour vector

$\overline{Y}$ = arithmetic mean vector for group

N = samples in a vector

y = sample value

v_i = voicing function. 1 if voicing on, 0 otherwise.

c = scaling factor (optional)

For each pattern, this distance measure is tabulated and the resulting histogram plot will then reveal how close to each other the F0 contours are. In other words, this will tell us whether our grouping function (stress pattern) adequately accounts for the observed shapes. A wide spread shows us that it does not, while a large concentration near the average indicates that we have found a pattern determined by stress alone. In the following section we will detail some of these results.

4. RESULTS OF F0 STUDY

As pointed out by Gimson [4], there are common word accentual/rhythmic patterns in English. For 1-to 3-syllable words, the patterns we found in our database were as follows:

1-syllable words: 1
2-syllable words: 10, 01, 12, 21
3-syllable words: 001, 010, 012, 100, 102, 120, 201, 210, 221

Certain patterns occur more frequently than others. For instance, we found that among two syllable words, the 10 pattern is the most frequent. The occurrence of the 001 pattern is small, and occurs usually with loanwords, for instance from French and Spanish.

Please note that in the following figures, all syllable lengths are normalized to be the same length (see above).

4.1. One-syllable words

Fig. 1 shows distance measures of F0 contours from the average contour, which resides at 0. We predict that if there is a representative average contour for each pattern, the distribution would be a half of a bell curve. The population distribution for the 1 pattern is close to this shape, but with some anomalies.

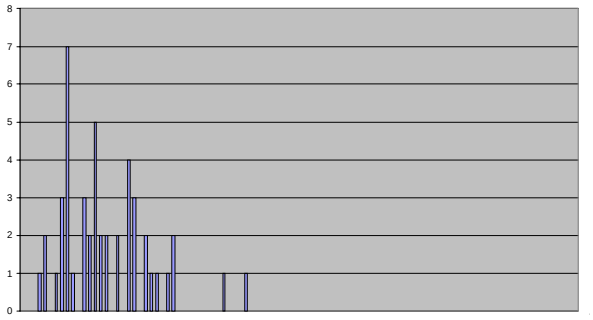

Figure 1: Distribution plot for stress-pattern '1'. The X-axis is on an arbitrary scale. The Y-axis is the count frequency for a given distance. Dissimilarities become significant around 1/3 on the X-axis.

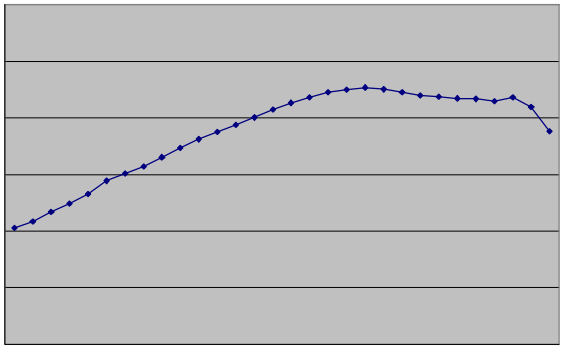

Figure 2: Average F0 contour for pattern '1'. Plot is in normalized log coordinates. The bottom, middle and top correspond to 50Hz, 100Hz and 200Hz respectively.

Fig. 2 shows the average F0 contour for this pattern, which is a slow rising contour. The two anomaly cases in Fig. 1 were 'Earls' and 'Hayes'. Both of these exhibit a fall at the end of the syllable as can be seen from the F0 contour of 'Earls' (Fig. 3).

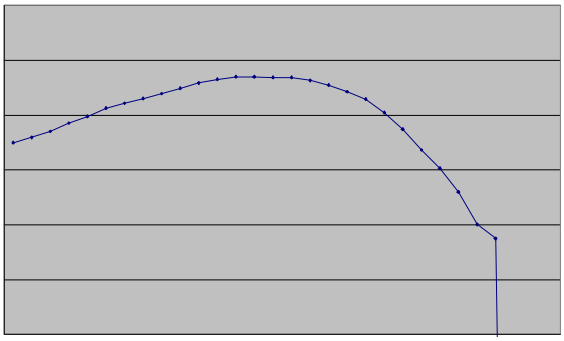

Figure 3: F0 contour of the word 'Earls'.

It's an open question whether this is an accidental choice of the speaker's or examples of a more consistent pattern, perhaps determined by segmental factors.

4.2. Two-syllable words

If we compare the 10 pattern in Fig. 4 and the 01 pattern in Fig. 5, the location of the peak is obviously different, as is the overall F0 contour shape. The 10 pattern shows a rise-fall pattern with a peak at about 80% into the first syllable, whereas the 01 pattern is a flat-rise-fall pattern with a peak at about 60% into the second syllable.

The 12 pattern in Fig. 6 is very similar to the 10 pattern, but once F0 reaches the target point of the rise, the 12 pattern has a longer stretch in this higher F0 region, implying perhaps the existence of the secondary stress. This assumption has to be tested by running a perceptual evaluation using synthesized words with these patterns.

The pattern involving the secondary stress on the first syllable is less straightforward. The population distribution of the 21 pattern (and also the 201 and the 210 pattern in three syllable words) failed to cluster. This implies that maybe the speaker used other acoustic cues to realize this pattern.

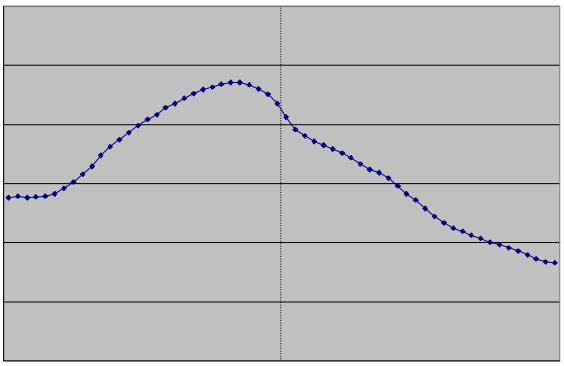

Figure 4: Average contour for pattern '10'.

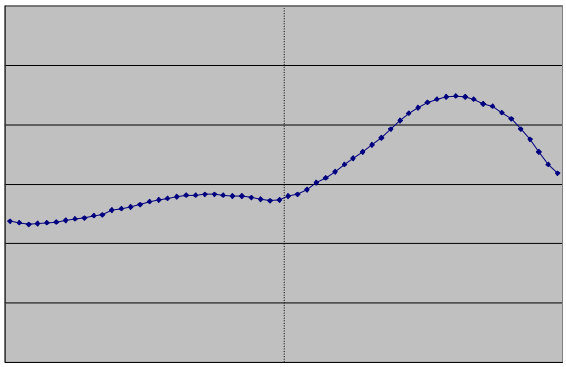

Figure 5: Average contour for pattern '01'.

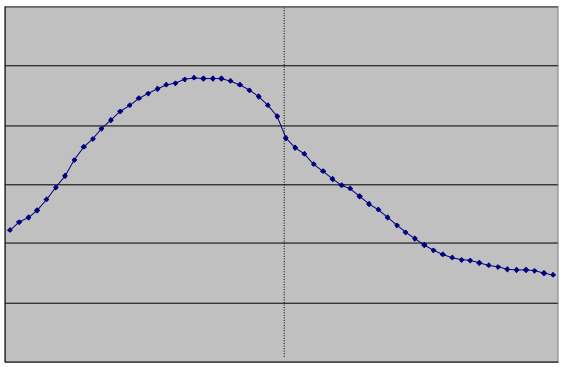

Figure 6: Average contour for pattern '12'.

In the sample furthest from the average, 'Quail Ridge', there is a clear peak in the first syllable and the fall is in the second syllable. The F0 contours of other words in the 21 pattern show that the level of the first syllable does not have to be higher or lower than that of the second syllable. We suspect that maybe durational cues are important in this case. Again, this has to be confirmed with a perceptual evaluation with synthetic words.

4.3. Three-syllable words

The 010 pattern shows a clear bell curve in the distribution and some anomalies. The average contour (Fig.7) is a low flat followed by a rise-fall contour with the F0 peak about 85% into the second syllable. It is worth noting that the some of the anomalies in this distribution include 'mispronounced' words in the sense of badly controlled F0.

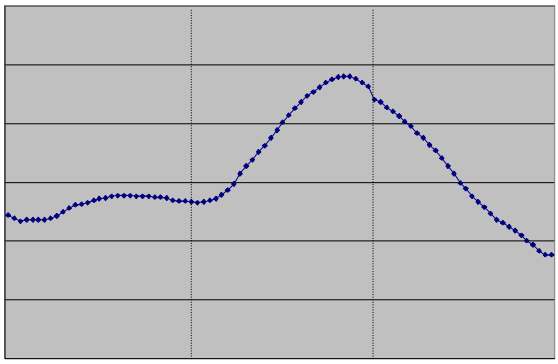

Figure 7: Average contour for '010'.

The 100 pattern and the 102 pattern show very similar contours. The secondary stress cue in the 102 pattern might be realized by a durational cue, just as in two-syllable words. It is interesting to note that the anomalies in these two patterns involve a nasal in the second or third syllable. It has been documented that when a nasal follows a vowel which is expected to have a F0 peak due to stress, the peak is often delayed into the nasal segment [8].

5. CONCLUSION

The continuing goal of our study is to produce word-level F0 contours and duration patterns for target words that are close or identical to those of a human speaker. We applied standard statistical procedures to uniform vector formats to which values for both F0 and duration were converted. Using only the stress pattern as the distinguishing feature, we have found that nearly all plots of F0-curve similarity distribution exhibit a distinct bell-curve shape. This confirms our working assumption, that the most important shapes could be found using this criterion alone.

However, the plots also show that there are significant deviations from this norm, exhibited by long 'skirts' and the placement of the bell-curve centroid some distance away from the average. The challenge for the future lies in correctly identifying additional grouping features that will adequately account for the patterns that fall outside the expected range. This will be an iterative process that should gradually provide us with a richer and more accurate template set, but will also be progressively more difficult as there will be fewer and fewer samples to generalize from.

Although we did not discuss duration in this paper, we would like to mention that our approach in this area follows a similar strategy as that for F0. This will be detailed in a future paper.

Also on our future agenda is the investigation of other positions such as sentence-medial within the announcement-type utterance. We suspect that the patterns in positions other than sentence-initial will be more complex. Moreover, since amplitude is also an acoustic cue to stress in English, we would like to apply the same strategy to this factor in the near future.

We are also hoping to incorporate our current methodology into phrase-level prosody, but do fully expect that the template concept will need to be enlarged upon, the higher up in the semantic hierarchy we go. However, we are confident that the current discovery is a good first step toward analyzing and generating prosody in English in general.

6. REFERENCES

1. Fry, D. B. *"Duration and Intensity as Physical Correlates of Linguistic Stress,"* JASA 27: 765-768, 1955.

2. Fry, D. B. *"Experiments in the Perception of Stress."* Language and Speech, 1: 126-152, 1958.

3. van Santen, J.P.H. and Hirschberg J. *"Segmental Effects on Timing and Height of Pitch Contours,"* Proceedings of ICSLP 94, Yokohama, Japan, pp. 719-722, Sept. 18-22, 1994.

4. Gimson, A.C. *An Introduction to the Pronunciation of English (2nd ed.)*, Edward Arnold Publishers, London, 1970.

5. Swerts, M., Strangert, E. and Heldner, M. *"F0 Declination in Read-aloud and Spontaneous Speech,"* Proceedings of ICSLP 96, Philadelphia, PA, USA, pp. 1501-1504, Oct. 3-6, 1996.

6. Lieberman, P., Katz, W., Jongman, A., Zimmerman, R. and Miller, M. *"Measures of the Sentence Intonation of Read and Spontaneous Speech in American English,"* JASA 77: 649-657, 1985.

7. Huber, D. *"A Statistical Approach to the Segmentation and Broad Classification of Continuous Speech into Phrase-sized Information Units,"* Proceedings of ICASSP 89, Glassgow, Scotland, pp. 600-603, May 23-26, 1989.

8. Ladd, D.R., and Silverman, K.E.A. *"Vowel Intrinsic Pitch in Connected Speech,"* Phonetica 41: 31-40, 1984.