Synthesizing Intonation and Stress for English

Daniel M. Albro

December 14, 1995

1 Introduction

The goals of this project were threefold: to learn about the inner workings of reasonably current speech synthesizers, to study the English sytem of intonation, and to provide the public domain with a reasonably natural-sounding speech synthesizer. In order to achieve these goals, I have attempted to add a model of English intonation and stress patterns to a public-domain speech synthesizer. The end result was a text-to-speech synthesizer which models intonation conservatively based on punctuation, and which uses a stress model based on lexical lookup.

1.1 Choosing a Synthesizer to Modify

There are, as far as I know, no reasonably natural-sounding speech synthesizers available in the public domain. The two synthesizers I was able to find for Linux, the operating system I use, range from one that simply has a sound file for each phoneme to a fairly reasonable one based on Klatt's 1980 synthesizer. I modified the latter.

2 Model Used

I have modeled stress, and then intonation, in two steps. Thus I will go over the two models separately.

2.1 Stress

I have loosely based the model of stress effects on Fry's article "Duration and Intensity as Physical Correlates of Linguistic Stress" (Fry 1955), wherein the author tests the effect of duration and intensity on perception of stress. The important part of the article for this purpose was the appendix, which listed the durations and intensities of five different words ("object," "subject," "digest," "contract," "permit") when used as nouns and when used as verbs. This provided data about minimal pairs of stressed versus unstressed syllables. I analyzed the data with the intent of figuring out what modifications I could make to the basic duration and intensity parameters of vowels to make them sound stressed. The

results of this analysis are attached to the end of this paper. For each word, I compared the stress intensities and durations between the stressed and unstressed versions of each syllable. I took the ratio of the unstressed durations and intensities to the stressed, and also the differences. In order to decide whether I should multiply the unstressed durations and intensities by some ratio, add a constant, or simply set them to a constant value, I compared the standard deviations for the stressed syllable values themselves, for the ratios, and for the differences. Judging from the standard deviations in this data, at least, it seems as though intensity (in decibels) increases by a ratio of approximately 1.5, and duration increases by the addition of about five centiseconds. As it turns out, I did not actually use those exact numbers, but I did decide based on these results to vary intensity multiplicatively and duration additively.

In my implementation the actual numbers chosen were based upon testing the approximate values given above, and modifying them for naturalness. In the final model, I added three centiseconds to the duration vowels with primary stress and one centisecond to vowels with secondary stress. I reduced the duration of stressless vowels by three centiseconds when they follow syllables with primary stress, and by one centisecond otherwise. With respect to intensity, I discovered that the synthesizer can only handle a fairly narrow range of intensities, and given that Fry concludes that intensity is not particularly salient in the perception of stress, I increased the intensity of primary stress only 1.1 times, and that of secondary stress 1.05 times.

2.2 Intonation

The model of intonation used here follows fairly closely that described by Pierrehumbert in her article "Synthesizing Intonation" (Pierrehumbert 1981), with additional information from (Pierrehumbert 1980) and (Akers and Lennig 1985). The following constraints affected the model:

- The synthesizer is a text-to-speech system, so no information can be directly specified about which words are being especially emphasized.
- Discourse analysis and syntactic parsing were not available as inputs to the process.
- As a text-to-speech system, the synthesizer must sound fairly neutral and attempt to avoid any jarring or blatantly incongruous patterns of intonation.
- Categorical and stress information was available in the form of a dictionary file created by Roger Mitton at the University of London based on the *Oxford Advanced Learner's Dictionary of Current English*[1].

Based on these constraints, it was decided to use four basic intonation patterns: neutral declarative, exclamatory, neutral questioning, and continuative. In these intonations (at least as I modeled them), each intonational phrase (intonational phrase divisions will be

[1]The use of this dictionary file gives the synthesizer a marked English accent

discussed further below) begins either with a low boundary tone (L%) or spreads the final F_0 value of the previous phrase, if that phrase was a continuation. From there, all content words (nouns, verbs, and adjectives) with primary stress marked in the dictionary are assigned a pitch accent. Except for the nuclear stress, defined as the last pitch accent in the phrase, and the penultimate pitch-accented syllable in the case of exclamations, all pitch accents assigned were high tones (H*), as specified by Pierrehumbert (Pierrehumbert 1981) for neutral declarative intonation. The final pitch accents were assigned according to the following table:

Phrase Type	Tonal Pattern
Declarative	H* L- L%
Continuative	H* L- H%
Exclamatory	L* H* L- L%
Questioning	L* L- H%

Intonational phrases were assigned on the basis of punctuation. Any member of the following set of punctuation marks was considered to end the current intonational phrase and begin a new one: $\{'!','?','.','"',':','-',';',',','(',')'\}$. The particular character that ends each phrase determines the type of intonation assigned to the phrase. Currently, a period assigns declarative intonation, a question mark assigns questioning, an exclamation point assigns exclamatory, and anything else assigns continuative.

Numerical values are assigned to the tones as a percentage of the difference from the declination top line to the baseline. That is, two boundary lines are assigned for each intonational phrase. The top line begins at the speaker's basic fundamental frequency plus seventy Hertz and declines to the basic frequency plus ten Hertz. The baseline begins at the speaker's basic fundamental frequency plus twenty Hertz and ends at the basic frequency minus twenty Hertz. Low tones are given a frequency of ten percent of the difference between the baseline and top line, and right-edge border tones descend ten percent below the baseline. High tones alternate between one-third of the distance between the lines and two-thirds, and the nuclear pitch is always placed at the top line. The alternation is assigned such that the pitch before the nuclear pitch will be one-third the distance between the baseline and the top line. The alternating values are then adjusted to add ten percent to the figure for nouns.

Once frequency values have been assigned to the relevant syllables, the actual fundamental frequency contour can be specified. High tone peaks are placed midway into the stressed vowel with which they are associated, except for the nuclear pitch, for which the tone is placed at twenty percent of the vowel duration. Following Pierrehumbert, high tones are left level for six centiseconds. To link the tones, a parabolic path is followed between low tones and other tones. Between high tones the contour dips, so a dip point is calculated and a simple path is calculated to go through that dip point.

2.3 Minor Additions

In addition to modeling stress and intonation, the program was modified also to extend the duration of phrase-final syllables and to pause between sentences and phrases (100 centiseconds for sentences, and 70 centiseconds between other types of phrase boundaries).

3 Implementation

The speech synthesizer works in the following manner. The user specifies some text to be read. The program takes that text and converts it into phones by looking up words in a dictionary file and by rule for strings that are not found in the dictionary. The phones are then converted to structures ("elements") containing synthesis parameters (formant frequencies, spectral tilt, etc.). The "holmes" procedure then takes the parameters for each phone and interpolates between them and sends one centisecond's worth of data to a different procedure to be synthesized. This procedure adds a glottal pulse and creates basic sound sample data, which is sent to the audio device to be played. As originally implemented, this entire process was indefinitely repeated over sentence-sized chunks of text.

The first modification made was to keep track of the syntactic category of each word as it is looked up in the dictionary. This information had been available, but had been left unused. Each word was assigned a character indicating its category: 'N' for nouns, 'V' for verbs, 'A' for adjectives, 'F' for function words, and 'U' for unknown. These were accumulated in an array. The second modification made was to split the text up into "intonational phrases" rather than sentences. In the original code, the phones were accumulated in a buffer as they were filled out from the text. This buffer would be sent on to be synthesized and then flushed when the punctuation symbols '!', '.', or '?' were encountered before a space or at the end of the file. This process was modified to do all of this after any of the characters listed above as well. The character that ends each phrase was added to the array of category labels.

3.1 Stress Assignment

Originally, the array of phones also contained stress indicators, but they were simply skipped over, and no way was provided for transferring the stress information to the other procedures. In order to preserve the stress information, an array of stress values was created, with the same same number of entries as the array used to hold elements. This array was then filled out while assigning elements to phonemes. Non-vowel elements were assigned a stress code of 4, stressless vowels were assigned 3, secondarily stressed vowels were assigned 2, and vowels with primary stress were assigned 21 for nouns, 11 for verbs or adjectives, or 1 for functional words and words of unknown category. Another modification was made after the element assignment loop to make sure that the last content word of the phrase has a primary stress. Nuclear stress was then assigned. This was done by finding the last stress with a value over 10 (*i.e.*, pitch accents) and changing it to 31. If no pitch accents have

been assigned, nuclear stress is instead assigned to the final syllable with primary stress. If no syllables have primary stress, then nuclear stress is assigned to the final syllable of the phrase. Finally, 5 is added to the stress value for the last vowel in the phrase, to indicate that the vowel should be lengthened.

3.2 Stress Effects

A structure was added to the program to keep information about segment intensity and duration changes, as well as to keep a pitch contour segment for each segment. The structures were filled out according to the following algorithm:

```
ASSIGN-STRESS()
 1  for i ← 1 to elements
 2      do if element represents a word break
 3            then previous_stress ← false
 4                 stress_level ← stress[i] mod 10
 5         Segment[i].intensity ← 1.0
 6         Segment[i].duration ← element[i].duration
 7         if element = vowel
 8           then if stress_level > 5
 9                   then stress_level ← stress_level − 5
10                        Segment[i].duration ← Segment[i].duration + 3
11                        switch
12                   case stress_level = 1 : Segment[i].intensity ← 1.1
13                        Segment[i].duration ← Segment[i].duration + 3
14                        previous_stress ← true
15
16                   case stress_level = 2 : Segment[i].intensity ← 1.05
17                        Segment[i].duration ← Segment[i].duration + 1
18                        previous_stress ← false
19
20                   case default  : if previous_stress
21                                      then Segment[i].duration ← Segment[i].duration − 3
22
23                                      else  Segment[i].duration ← Segment[i].duration − 1
24                                            previous_stress ← false
```

3.3 Intonation

In order to model intonation, an array is created large enough to hold one F_0 value for each centisecond of the current intonational phrase. The F_0 values at the stress points themselves were determined as described above, based on the stress values assigned in the previous section. The main implementation issue to be discussed here is the assignment of

contours between the assigned F_0 values. Between low tones and other tones, whether low or high, a "monotonic" path is taken, which is to say a more or less direct path. The exact curve is determined by a spline-fit algorithm, given an array of the positions in time of the two assigned points, plus an array of the frequencies at those points. First, a fake point is determined at the same frequency as the higher of the points and in mirror image to it, rotated around the lower of the points. Then, the three points are sent to the following algorithm, which generates the spline parameters:

```
SPLINE-FIT(x, y)
 1  d ← 2 · (x_3 − x_1)
 2  u_1 ← x_2 − x_1
 3  u_2 ← x_3 − x_2
 4  w ← 6.0 · ((y_3 − y_2)/u_2 − (y_2 − y_1)/u_1)
 5  p_1 ← 0.0
 6  p_3 ← 0.0
 7  p_2 ← w/d
 8  return u, p
```

The actual points are then determined as follows:

```
F(x)
 1  return x^3 − x
```

```
COMPUTE-F0-VALUE-SPLINE(x, y, u, p, point)
 1  while point > x_{i+1}
 2      do i ← i + 1
 3         t ← (point − x_i)/u_i
 4  return (t · y_{i+1} + (1 − t) · y_i + u_i^2 · (f(t) · p_{i+1} + f(1 − t) · p_i)/6.0)
```

For the contour between two high tones, we must first compute the location of the bottom of the dip between them (B is the baseline value at the lower of the two tones):

```
DIP-POINT(F0_1, F0_2, B, t_1, t_2)
 1  t ← t_2 − t_1
 2  if t < 20
 3     then F ← 1 − (0.005t)
 4
 5     else  F ← 0.9 − 0.015(t − 20)
 6           if |F0_2 − F0_1| < .1
 7              then return F
 8
 9              else  if F0_1 < F0_2
```

```
10                    then F_a ← F
11                         F_b ← F_a · (F0_1 − B)/(F0_2 − B)
12                         F ← (F_a + F_b)/2
13                         if F > F0_1
14                           then F ← F0_1
15
16                    else F_a ← F
17                         F_b ← F_a · (F0_2 − B)/(F0_1 − B)
18                         F ← (F_a + F_b)/2
19                         if F > F0_2
20                           then F ← F0_2
21
22  return F
```

Given this dip point, a standard polynomial fit is used, following Lagrange's interpolation formula:

$$\sum_{1 \le j \le N} y_j \prod_{1 \le i \le N, i \ne j} \frac{x - x_i}{x_j - x_i},$$

where the three points are the two assigned tones and the dip point.

Finally, once the entire contour has been generated, a variable is set to recall for the next time whether the border tone should be spread, and the generated intensity, duration, and contour values are fed into the normal synthesizer (the intensity ratio is multiplied against all amplitude parameters, such as the amplitude of the vowel formants and the nasal poles, etc.) The algorithms used above for curve fitting come from (Sedgewick 1990).

4 Results

With the intonation model described above, the speech synthesizer sounds much more human than it did before. The intonation is not perfect, so it tends to sound like "Arnold Schwarzenegger" English, but overall much more recognizable than before. As an evaluation metric, I will include a tape of the synthesizer reading a letter to Miss Manners[2], before and after the modification.

Bibliography

Akers, G., and M. Lennig. 1985. Intonation in text-to-speech synthesis: Evaluation of algorithms. *Journal of the Acoustical Society of America* 77(6):2157–2165.

Fry, D. B. 1955. Duration and intensity as physical correlates of linguistic stress. *Journal of the Acoustical Society of America* 27(4):765–768.

[2]It was a convenient text to use...

Pierrehumbert, J. 1980. *The Phonology and Phonetics of English Intonation*. PhD thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts.

Pierrehumbert, J. 1981. Synthesizing intonation. *Journal of the Acoustical Society of America* 70(4):985–995.

Sedgewick, R. 1990. *Algorithms in C*. Reading, MA: Addison-Wesley.

A Duration and Intensity Calculations

The duration and intensity calculations follow.

B Code

The following represents only the files that were modified for this project. The new code has been set off (more or less) by comments saying "INTONATION MOD".