Synthesis of Prosodic Styles

Chilin Shih, Greg Kochanski

Bell Laboratories, Lucent Technologies

{cls,gpk}@research.bell-labs.com

Abstract

A text-to-speech system can effectively imitate distinctive speaking styles when a few critical prosodic features are modeled and controlled. This paper demonstrates the methodology with a number of examples, including the ornamental notes and the amplitude profile that define the singing style of Dinah Shore, the phrase curve that sets off the dramatic speaking style of Martin Luther King Jr, and the variations of accent shapes between two American English speakers. The styles are described by Stem-ML tags (soft template mark-up language), which offers the flexibility needed to control accent shapes, phrasal pitch contours, and amplitude profiles, for speech as well as for singing.

1. Introduction

While the value of a style is subjective and involves personal, social and cultural preferences, the existance of style itself is objective and implies that there is a set of consistent features. These features, especially those of a distinctive, recognizable style, lend themselves to quantitative studies and modeling. A human impressionist can deliver a stunning performance by dramatizing the most salient feature of an intended style. Likewise, we show that a text-to-speech system can successfully convey the impression of a style when a few distinctive prosodic features are properly modeled. This can be effective even without matching the voice quality.

Much of the style of a speaker can be expressed in terms of features in f_0, amplitude, spectral tilt, and duration. [1, 2, 3, 4]. Personal style is conveyed by repeated patterns of these features occurring at characteristic locations. For example, a speaker may use the same feature patterns at the beginning or the end of each phrase, or at emphasized words. In this paper we focus on the modeling of f_0 and amplitude.

We chose three examples to illustrate the control of prosodic styles: the singing style of Dinah Shore, the speaking style of Martin Luther King Jr, and the accent shapes of two American English speakers. In the following sections, we first analyze the prosodic features of these styles, then we go into the technical details of describing these features in prosodic tags based on Stem-ML [5], which offers the flexibility needed to control accent shapes, phrasal pitch contours, and amplitude profiles. Similar features can be used to support other stylistic variations and emotional speech [6, 7, 8]. Our singing synthesis program also focus on style and performance rules rather than on voice quality [9, 10].

2. Features of Styles

Figure 1 shows the amplitude profiles of the first four syllables *Dai-sy Dai-sy* from the song *Bicycle built for two* by the singer Dinah Shore, who was described as a "rhythmical singer" [11]. A bow-tie-shaped amplitude profile expands over each of the four syllables, or notes. The second syllable, centered around 1.2 second, is the clearest example. The increasing amplitude toward the end of the note contrasts with most singers, whose amplitude tends to decline toward the end. This style of amplitude profile shows up very frequently in Shore's singing. The consistent delivery and and the clash with the listener's expectation mark a very distinct style.

Figure 2 shows the f_0 trace of phrases from the speech "I have a dream" delivered by Dr. Martin Luther King Jr. A dramatic pitch rise consistently marks the beginning of the phrase and an equally dramatic pitch fall marks the end. The middle section of the phrase is sustained on a high pitch level. The pitch profile shown in Figure 2 is found in most phrases in Martin Luther King's speech, even though the phrases differ in textual content, syntactic structure, and phrase length.

In the f_0 traces of typical English sentences from typical speakers, as in Figures 3 and 4 from two different persons, the dominant features reflect word accent and emphasis. The phrasal component, if any, is a smooth decline. In King's distinctive rhetorical style, word accent and emphasis modifications are present, but the magnitude of the change is relatively small compared to the f_0 change marking the phrase. The f_0 profile over the phrase is one of the most important marks of this style.

Speakers 1 and 2 in Figures 3 and 4 convey different speaking styles by using distinct rising contours. Speaker 1's rising accent, shown on the words *Tennessee* and *Maryland*, rises early and levels off in the final section of the word. Speaker 2's rising accent, shown on the words *fight*, *Seattle* and *Albuquerque*, has a fairly straight rising slope, which starts from the center of the word and peaks at the end of the word. We will return to these two examples in Section 6.

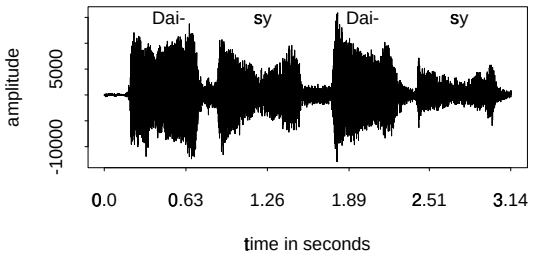

Figure 1: *Dinah Shore's signature amplitude profile*

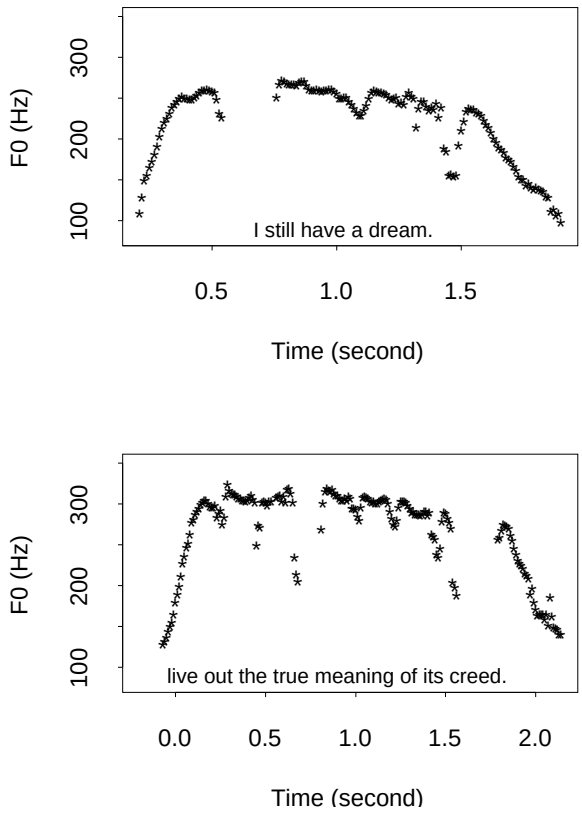

Figure 2: *Phrasal f_0 profiles from the speech of Martin Luther King Jr.*

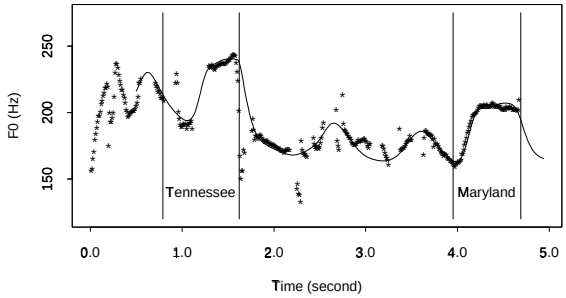

Figure 3: *A sentence from Speaker 1 with two rising accents. "I live in Nashville Tennessee and I'd like to go to Baltimore Maryland."*

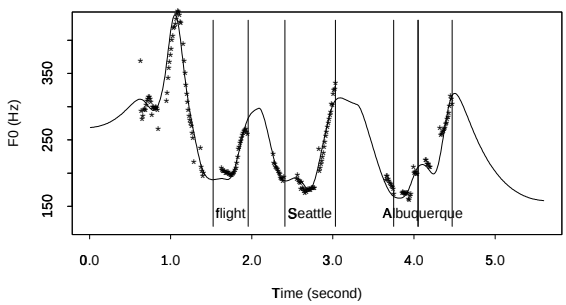

Figure 4: *A sentence from Speaker 2 with multiple rising accents. "Um I would like a flight to Seattle from Albuquerque."*

3. Prosody Tags

The prosody control of speech and song described in this paper is done in Stem-ML tags (Soft TEMplate Mark-up Language) [5][12]. Stem-ML is a tagging system with mathematically defined algorithm to translate tags into quantitative prosody. The system is designed to be language independent, and furthermore, it can be used effectively for both speech and music.

Following the outline in Figure 5, text or music scores are passed to the tag generation component, which uses heuristic rules to select and to position prosodic tags. Style-specific information is read in to facilitate the generation of tags. Style-specific attributes may include parameters controlling breathing, vibrato, and note duration for songs, in addition to Stem-ML templates to modify f_0 and amplitude. The tags are then sent to the prosody evaluation component, Stem-ML, which produces time series of f_0 or amplitude values.

We rely heavily on two of the Stem-ML features to describe speaker styles in this paper. First, Stem-ML allows the separation of local (accent templates) and non-local (phrasal) components of intonation. One of the phrase level tags *step_to* moves f_0 to a specified value which remains effective until the next *step_to* tag. When it is described by a sequence of *step_to* tags, the phrase curve is being treated as a piece-wise differentiable function. We use this method to describe Martin Luther King's phrase curve and music notes. Secondly, Stem-ML accepts user-defined accent templates with no shape and scope restrictions. This feature gives users the freedom to write templates to describe accent shapes of different languages as well as variations within the same language. We write speaker-specific accent templates for speech, and ornament templates for music.

The specified accent and ornament templates may result in physiologically implausible combination of targets. Stem-ML accepts conflicting specifications and returns smooth surface realizations that best satisfy all constraints.

We observe that the muscle motions that control prosody are smooth because it takes time to make the transition from one intended accent target to the next. We also observe that when a section of speech material is unimportant, the speaker may not expend much effort to realize the targets [13]. We then represent the surface realization of prosody as an optimization problem, minimizing the sum of two functions: a physiological constraint G, which imposes a smoothness constraint by minimizing the first and second derivatives of the specified pitch p, and a communication constraint R, which minimizes the sum of errors r between the realized pitch p and the targets y. In other words, one should speak precisely if one really wants to be understood.

The errors are weighted by the strength S_i of the tag which indicates how important it is to satisfy the specifications of the tag. If the strength of a tag is weak, the physiological constraint takes over and in those cases, smoothness becomes more important than accuracy. S_i controls the interaction of accent tags with their neighbors by way of the smoothness requirement, G. Stronger tags exert more influence on their neighbors. Tags also have α and β, which control whether errors in the shape or average value of p_t is most important, these are derived from the Stem-ML *type* parameter. In this work, the targets, y, consist of an accent component riding on top of a phrase curve.

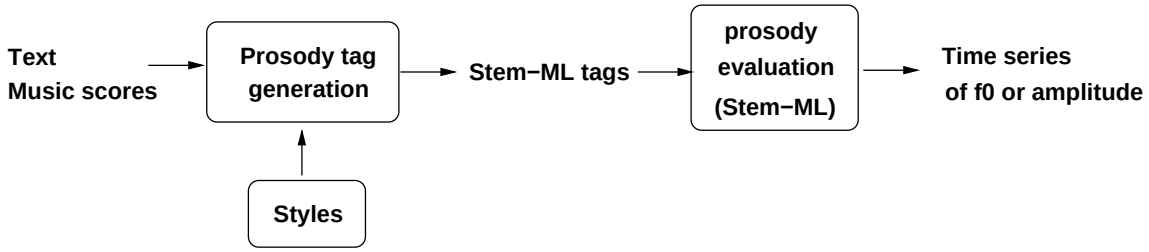

Figure 5: *System diagram of style-dependent prosody tag generation and evaluation.*

$$G = \sum_t \dot{p}_t^2 + (\pi\tau/2)^2 \ddot{p}_t^2$$

$$R = \sum_{i \in \text{tags}} S_i^2 r_i$$

$$r_i = \sum_{t \in \text{ tagi}} \alpha(p_t - y_t)^2 + \beta(\bar{p} - \bar{y})^2$$

The generated f_0 and amplitude contours is used in a text-to-speech system to generate speech and songs. In the current implementation, amplitude modulation is applied to the output of the TTS system.

4. Phrase Curve

Martin Luther King's speech has a strong phrasal component with an outline defined by an initial rise, optional stepping up to climax, and a final fall. This outline is described with Stem-ML *step_to* tags. The argument to, introduced by *to=* at the end of each line below, specify the intended f_0 as $base + to \times range$, where $base$ is the baseline and $range$ is the speaker's pitch range.

We use heuristic grammar rules to place the tags. Each phrase starts from the $base$ value (*to=0*), stepping up on the first stressed word, remaining high till the end for continuation phrases, and stepping down on the last word of the final phrase. At every pause, return to 20% of the pitch range above $base$ (*to=0.2*), and stepping up again on the first stressed word of the new phrase. The amount of *step_to* correlates with sentence length. Additional stepping up is used on annotated, strongly emphasized words.

```
Cname=step-to;pos=0.21;strength=5; to=0;
# Step up on the first stressed word ‘‘nation’’
Cname=step-to; pos=0.42; strength=5; to=1.7;
Cname=step-to; pos=1.60; strength=5; to=1.7;
# Further step up on rise
Cname=step-to; pos=1.62; strength=5; to=1.85;
Cname=step-to; pos=2.46; strength=5; to=1.85;
# Beginning of the second phrase
Cname=step-to; pos=3.8; strength=5; to=0.2;
# Step up on the first stress word live
Cname=step-to; pos=4.4; strength=5; to=2.0;
Cname=step-to; pos=5.67; strength=5; to=2.0;
# Step down at the end of the phrase
Cname=step-to; pos=6.28; strength=5; to=0.4;
```

The *step_to* tags above produce the phrase curve shown in dotted lines in Figure 6 for the sentence *This nation will rise up, and live out the true meaning of its creed.* The solid line shows the generated f_0 curve, which is the combination of the phrase curve and the accent templates, to be discussed momentarily in Section 6.

5. Musical Notes

Musical scores are under-specified. Performers may have very different renditions based on the same scores. We make use of the musical structures and phrasing notation to insert ornaments [14] and to implement performance rules, which include the default rhythmic pattern, retard, and duration adjustment [15, 16]

The musical input format is given below, showing the first phrase of *Bicycle Built for Two* [17]. This information specifies notes and octave (columns 1), nominal duration (column 2), and text (column 3, expressed phonetically). Column 3 also contains accent information from the lexicon (strong accents are marked with double quotes, weak accents by periods). The letter [t] in the note column indicates tied notes, and a dash links syllables within a word. Percent signs mark phrase boundaries. Blank lines mark measure boundaries, and therefore carry information on the metrical pattern of the song.

```
3/4 b=260
%
g2 3 “dA-

e2 3.0 zE

%
c2 3 “dA-

g1 3.0 zE

%

a1 1.00 “giv
b1 1.00 mE
c2 1.00 yUr

a1 2.00 “an-
c2 1.00 sR

g1t 3.0 “dU-

g1 2.0
g1 1.0 *
%
```

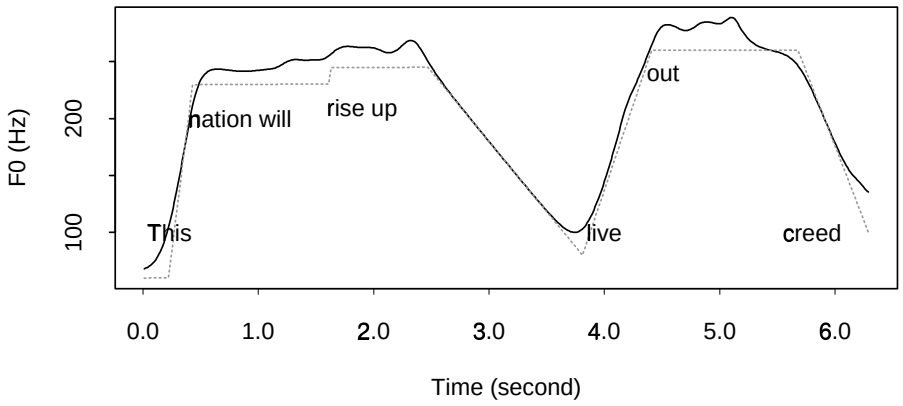

Figure 6: *Generated phrase curve with accents in the styles of Martin Luther King.*

Musical notes are treated analogously to the phrase curve in speech. Both are built with Stem-ML *step_to* tags. In music, the *pitch range* is defined as an octave, and each *step* is 1/12 of an octave in the logarithmic scale. Each musical note is controlled by a pair of *step_to* tags. The first four notes of *Bicycle Built for Two* is shown below:

```
# Dai- (Note G)
Cname=step-to; pos=0.16;strength=8; to=1.9966;
Cname=step-to; pos=0.83;strength=8; to=1.9966;
# sy (Note E)
Cname=step-to; pos=0.85;strength=8; to=1.5198;
Cname=step-to; pos=1.67;strength=8; to=1.5198;
# Dai- (Note C)
Cname=step-to; pos=1.69;strength=8; to=1.0000;
Cname=step-to; pos=2.36;strength=8; to=1.0000;
# sy (Note G, one octave lower)
Cname=step-to; pos=2.38;strength=8; to=0.4983;
Cname=step-to; pos=3.20;strength=8; to=0.4983;
```

The strength specification of the musical *step_to* is very strong (*strength=8*). This help to maintain the specified frequency as the tags pass through the prosody evaluation component.

6. Tag templates

Word accents in speech and ornament notes in singing are described in style-specific tag templates. Each tag has a scope, and while it can strongly affect the prosodic features inside its scope, it has a decreasing effect as one goes farther outside its scope. In other words, the effects of the tags are more or less local. These templates are intended to be independent of speaking rate and pitch. They can be scaled in amplitude, or stretched along the time axis to match a particular scope.

Distinctive speaking styles may be conveyed by idiosyncratic shapes for a given accent type. We examine the ARPA Communicator travel reservation database, where subjects interact with a dialogue system trying to make flight reservations, and find many examples of speaker-specific accent shapes. One

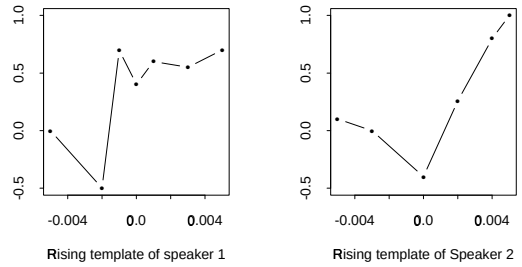

Figure 7: *Rising accent templates from two different speakers*.

of the most common intonation patterns associated with a request of flight origin and destination is the rising intonation (L* H- H%). Different instances of the rising shapes by the same speaker are fairly consistent, but there are substantial differences between speakers. Figure 7 shows two different rising templates from two speakers.

We used these two templates to generate the rising contours in Figures 3 and 4, respectively. Figure 3 shows the sentence *... I live in Nashville Tennessee and I'd like to go to Baltimore Maryland.* The rising intonation in question shows up on the words *Tennessee* and *Maryland*, where the pitch rises early and peaks before the end of the word. The final section of these two words has relatively flat f_0. These attributes are reflected in the first template of Figure 7. The template is aligned around the center of the word and stretched to match the duration of the word. Figure 4 shows the sentence *Um I would like a flight to Seattle from Albuquerque.* The speaker used the rising accent on *flight*, *Seattle*, and twice on *Albuquerque*, where both *Al-* and *-quer-* are accented. In contrast to the first speaker, the rising slopes from the second speaker are fairly straight, rising from the valley near the center of the word and peak at the end of the word. The four rising contours in Figure 4 are all generated from the second rising template in Figure 7. In both figures, the natural f_0 tracks are plotted in stars and the generated f_0 tracks in solid lines.

In the song program there are templates of ornament notes

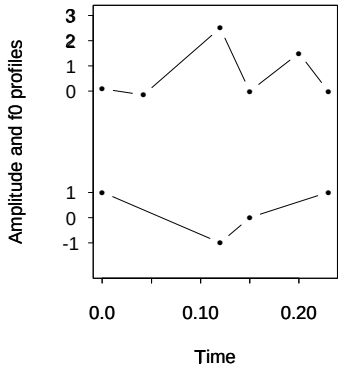

Figure 8: *Ornament templates.*

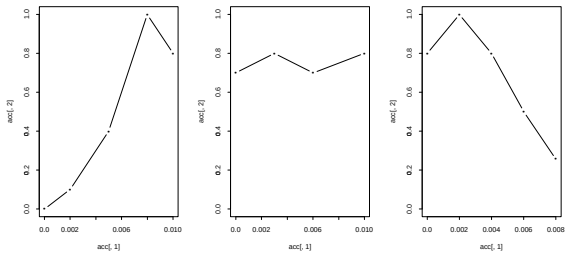

Figure 9: *Accent templates for King's prosody.*

which can be placed in specified locations, superimposed on the musical note. Figure 8 shows the f_0 (top) and amplitude (bottom) templates of an ornament in the singing style of Dinah Shore. Her ornament has two humps in the f_0 trajectory, where the first f_0 peak coincides with the amplitude valley. The length of the ornament stretches elastically with the length of the musical note within a certain limit. On short notes (around 350 msec) the ornament stretches to cover the length of the note. On longer notes the ornament only affects the beginning.

Shore often used this ornament in a phrase final descending note sequence, especially when the penultimate note is one note above the final note. She also used this ornament to emphasize rhyme words.

In King's speech, there are also reproducible, speaker-specific accent templates. Figure 9 displays accent templates used to generate Figure6. King's choice of accents is largely predictable from the phrasal position: a rising accent in the beginning of a phrase, a falling accent on emphasized words and in the end of the phrase, and a flat accent elsewhere.

Once tags are generated, they are fed into the prosody evaluation unit, which interprets Stem-ML tags into the time series of f_0 or amplitude.

7. Implementation Examples

The output of the tag generation component is a set of tag templates. We show a truncated but operational example displaying the tags that control the amplitude. Other prosodic parameters are similar, but not shown to save space.

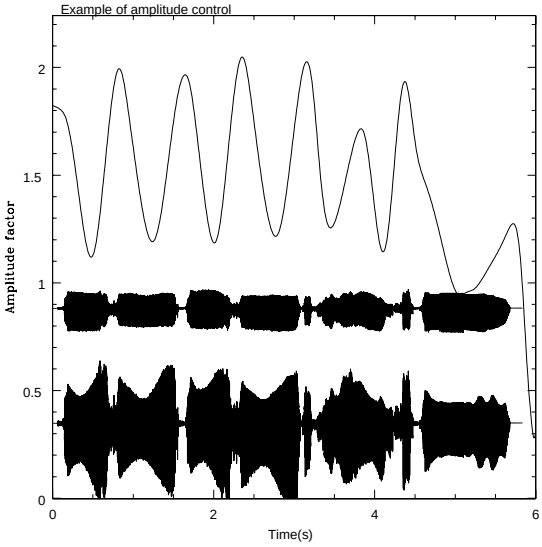

Figure 10: *Amplitude control in synthesized song.*

The first two lines consist of global settings that partially define the style we are simulating. The next section ("user-defined tags") is the database of tag templates for this particular style. After the initialization section, each line corresponds to a tag template.

```
# Global settings
add=1;base=1;range=1;smooth=0.06;pdroop=0.2;adroop=1

# User-defined tags
name=SCOOP; shape=-0.1s0.7,0s1,0.5s0,1s1.4,1.1s0.8
name=DROOP; shape=0s1,0.5s0.2,1s0;
name=ORNAMENT; shape=0.0s1,0.12s-1,0.15s0,0.23s1

# Amplitude accents over music notes
# Dai-
ACname=SCOOP; pos=0.15; strength=1.43; wscale=0.69
# sy
ACname=SCOOP; pos=0.84; strength=1.08; wscale=0.84
# Dai-
ACname=SCOOP; pos=1.68; strength=1.43; wscale=0.69
# sy
ACname=SCOOP; pos=2.37; strength=1.08; wscale=0.84
# give
ACname=DROOP; pos=3.21; strength=1.08; wscale=0.22
# me
ACname=DROOP; pos=3.43; strength=0.00; wscale=0.21
# your
ACname=DROOP; pos=3.64; strength=0.00; wscale=0.21
```

Finally, the prosody evaluation module produces a time series of amplitude vs. time. Figure 10 displays (from top to bottom), the amplitude control time series, speech produced by the synthesizer without amplitude control, and speech produced by the synthesizer with amplitude control.

8. Conclusion

We show that it is possible to convey the impression of a distinct speaker by capturing the most salient prosodic attributes. Speech and song demos are available from http://www.bell-labs.com/project/tts/stem.html.

The style-specific attributes are described in prosody tags written in the language Stem-ML, and are used to drive a TTS system. This language is designed as a multi-lingual intonation system. Stem-ML is flexible enough to handle both speech and songs, and provides enough control to differentiate speaker styles.

In practice, different prosodic styles could be used to mark different sections of a web page, to act as different persons in a dialogue system, and to read email with the prosodic characteristics of the sender.

9. References

[1] Y. Kitahara and Tohkura Y., "Prosodic components of speech in the expression of emotion," *JASA*, vol. 84, 1989.

[2] N. Higuchi, T. Hirai, and Y. Sagisaka, "Effect of speaking style on parameters of fundamental frequency contour," in *Progress in Speech Synthesis*, J. et.al. van Santen, Ed., pp. 417–428. Springer-Verlag, 1997.

[3] K. Maekawa, "Phonetic and phonological characteristics of paralinguistic information in spoken Japanese," in *International Conf. Sp. Lg. Proc.*, 1999, no. 0997.

[4] D. Erickson, A. Abramson, K. Maekawa, and T. Kaburagi, "Articulatory characteristics of emotional utterances in spoken English," in *ICSLP*, Beijing, China, 2000.

[5] Greg P. Kochanski and Chilin Shih, "Stem-ML: Language independent prosody description," in *ICSLP*, Beijing, China, 2000.

[6] Janet E. Cahn, "Generating pitch accent distributions that show individual and stylistic differences," in *The ESCA Workshop on Speech Synthesis*, 1998.

[7] M. Abe, "Speaking styles: statistical analysis and synthesis by a text-to-speech system," in *Progress in Speech Synthesis*, J. et.al. van Santen, Ed., pp. 495–510. Springer-Verlag, 1997.

[8] A. I. C. Monaghan and D. R. Ladd, "Manipulating synthetic intonation for speaker characterization," in *ICASSP*, 1991, pp. 453–456.

[9] M. W. Macon, L. Jensen-Link, J. Oliverio, M. Clements, and E. B. George, "A system for singing voice synthesis based on sinusoidal modeling," in *Proc. of International Conference on Acoustics, Speech, and Signal Processing*, 1997, vol. 1, pp. 435–438.

[10] Gerald Bennett and Xavier Rodet, "Synthesis of the singing voice," in *Current Directions in Computer Music Research*, Max V. Mathews and John R. Pierce, Eds., pp. 19–44. The MIT Press, Cambridge, Massachusetts, 1991.

[11] Dinah Shore, "Bicycle built for two," in *The Dinah Shore Collection, Columbia and RCA recordings, 1942-1948*.

[12] Greg Kochanski and Chilin Shih, "Soft templates for prosody mark-up," Tech. Rep., Bell Laboratories, Lucent Technologies, http://www.bell-labs.com/project/tts/stem-MLdefine.pdf, 2001.

[13] Chilin Shih and Greg P. Kochanski, "Chinese tone modeling with Stem-ML," in *ICSLP*, Beijing, China, 2000.

[14] Robert Garretson, *Choral Music: History, Style, and Performance Practice*, Prentice Hall, 1993.

[15] J. Sundberg, A. Askenfelt, , and L. Frydén, "Musical performance: A synthesis-by-rule approach," *Computer Music Journal*, vol. 7, pp. 37–43, 1983.

[16] A. Friberg, *A Quantitative Rule System for Musical Performance*, Ph.D. thesis, Royal Institute of Technology (KTH), Sweden, 1995.

[17] Harry Dacre, *Daisy Belle, or A Bicycle Made for Two*, Francis, Day and Hunter, 1892, music composed by the author.