Synthesis of Prosodic Styles

Chilin Shih, Greg Kochanski

Bell Laboratories, Lucent Technologies

Murray Hill, NJ

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

Abstract

A speech synthesizer can effectively imitate distinctive speaking styles when a few critical prosodic features are modeled and controlled. We represent a style of speech or singing as a set of localized prosodic features (ornaments or accents), along with a set of rules to choose where the features are applied. This paper shows a number of examples, including the ornamental notes and the characteristic amplitude profile that defines the singing style of Dinah Shore, and the rise-plateau-fall frequency contour that characterizes the public speech of Martin Luther King Jr. 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

The sense of a style is expressed in recurrent, dominant features. These features are not normally applied arbitrarily, but instead have a close relationship to the underlying content. Moreover, the salient features of a style are usually rare in the typical population, or are placed in an unusual pattern. Matching a style does not require everything to be similar, only the few salient features and their patterns of use. For instance, a human impersonator can deliver a stunning performance by dramatizing the most salient features of a politician's speaking style without actually duplicating the speech of the person he/she is impersonating. 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 feature/location description of style is essentially a restatement of (Bloch, 1953) and is applicable to a broad range of speech, art and music. It can be used to describe storytelling styles:

> One stylistic device in this tale, employed as a connective between the episodes, and commented on by Thompson [reference deleted] as a convention of Märchen, is the direct question addressed to the audience ... (Dorson, 1960)

Here we have a style defined by a feature ("direct question") and the location ("connective between the episodes"). Or, describing a style at a more detailed, phonetic level: "The humor of dialect is present throughout. Instances are the use of aspirated h's before consonants, ... " (Dorson, 1960). Similarly, one can trivially define a style of typography by specifying the forms of the letters, as there is an implicit set of rules that specify when one uses each letter.

We are concerned here with low-level prosodic styles, describing the detailed implementation of a phrase after the words have been chosen. Much of the style of a speaker can be expressed in terms of features in f_0, amplitude, spectral tilt, and duration (Kitahara and Y., 1989; Higuchi et al., 1997; Maekawa, 1999; Erickson et al., 2000). 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 on emphasized words. In this paper we focus on the modeling of f_0 and amplitude. Some of the salient features are local while others involve changes over a broad scope, such as an entire phrase of speech.

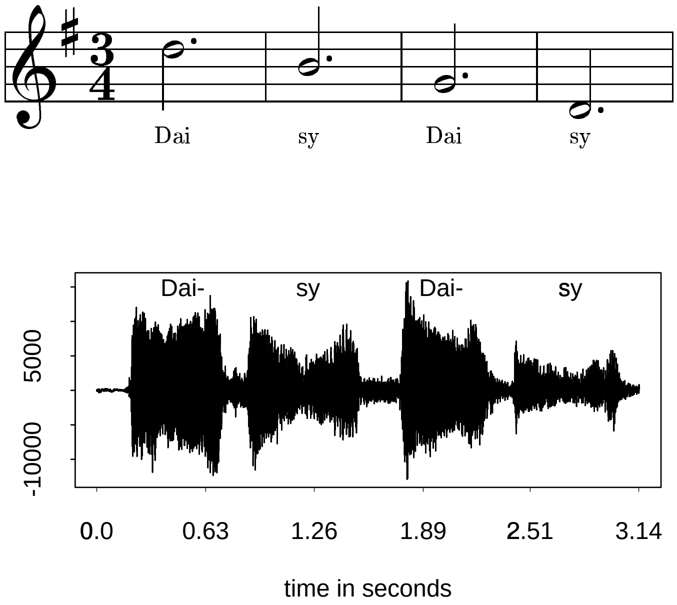


Figure 1: *Dinah Shore's signature amplitude profile*

To further support this definition of style, we point to two examples which we will discuss in detail later: musical ornamentation in the singing style of Dinah Shore, and characteristic phrase-scope f_0 contours in the public speeches of Martin Luther King Jr. Figure 1 shows the amplitude profiles of the first four syllables *Dai-sy Dai-sy* from the song *On a Bicycle Built for Two* by the singer Dinah Shore (Shore, 1999), who was described as a "rhythmical singer." (Jackson, 1999). She uses a bow-tie-shaped amplitude profile over each of the four syllables, or notes. (The second syllable, centered near 1.2 second, is the clearest example.) The increase in amplitude toward the end of the note contrasts with most singers, who have amplitude profiles that tend to decline across each note. This bow-tie amplitude profile shows up very frequently in Shore's singing. Her consistent use of this profile and the contrast with the norm mark the amplitude profile as an important component of her very distinct style.

Figure 2 shows the f_0 trace of phrases from the speech "I have a dream"

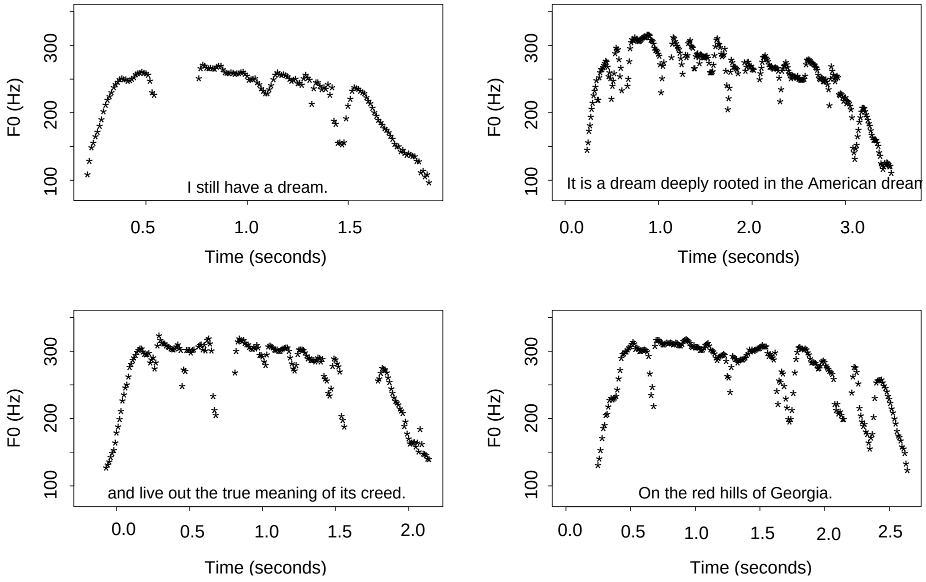


Figure 2: *Phrasal f_0 profiles from the speech of Martin Luther King Jr.*

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 length.

In the following sections, we first explain how to describe prosodic features with Stem-ML, a prosody description language that offers the flexibility needed to control accent shapes, phrasal pitch contours and amplitude profiles. We start by describing a phrase from Dinah Shore's singing to illustrate the procedure of annotation, automatic fitting and generation. Similar features can be used to support other stylistic variations and emotional speech (Cahn, 1998; Abe, 1997; Monaghan and Ladd, 1991). Our singing synthesis focusses on style and performance rules rather than on voice quality (Macon et al., 1997; Bennett and Rodet, 1991).

2 Describing Prosody with Stem-ML

In this paper, the control of pitch and amplitude in speech and song is done via Stem-ML tags (Soft TEMplate Mark-up Language) (Kochanski and Shih, 2000; Kochanski and Shih, 2002). Stem-ML provides prosody markup tags in the form of TAG_{value}. All tags are mathematically defined with an algorithm for translating tags into quantitative prosody. The system is designed to be language independent, and furthermore, it can be used effectively for both speech and music.

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* ($\Updownarrow$) 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 separates the placement of accents (ornaments) from their detailed shape. Any accent template can be inserted at any point, without much consideration of the environment, because Stem-ML calculates coarticulation effects for the user. 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.

Some combinations of accent and ornament templates may conflict or be impossibly difficult to realize precisely; 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 (Shih and

Kochanski, 2000). 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. Loosely speaking, we assume that the speaker balances the effort required to speak against the possibility of being misunderstood.

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 low, 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 are realized more accurately and also 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.

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

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

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

In summary,

- Local movements such as accents, tones, and musical ornaments are described by Stem-ML *shape* tags. In this paper, we define and use bowtie

($\bowtie$), wiggle ($\approx$), rise ($\triangle$), fall ($\bigtriangledown$) and droop ($\triangleright$) shapes.

Each shape is specified as a set of points $[(x_1, y_1), (x_2, y_2), \ldots]$, and α.

The subscript in $shape_{strength}$ specifies the strength of the tag.

These can be used to describe word accents in speech and ornament notes in singing. 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 sections 3.1, 3.2, and 3.4, we explore several examples where local f_0 or amplitude modification is controlled by Stem-ML *shape* templates.

- Non-local movements, including musical notes and phrase curve, are controlled by Stem-ML *step-to* tags ($\Updownarrow$). $\Updownarrow_{value}$ moves the phrase curve to $value \times pitchrange$, and the pitch will follow.

  Section 3.3 shows an example of describing larger scope features with Stem-ML phrase tags.

The generated f_0 and amplitude contours are 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.

3 Examples

3.1 Musical Ornaments - Changing Pitch

We use Stem-ML in both directions, both to evaluate prosody from tags and, in reverse, to deduce tags from the data. The Stem-ML evalation component takes tag and attribute values as input and generates time series data such as f_0 or amplitude curve. The Stem-ML optimizer takes data and partial tag annotation as input, and finds the best description of the data in terms of the tags. One feeds it Stem-ML tags with free parameters (*e.g.,* an undetermined strength

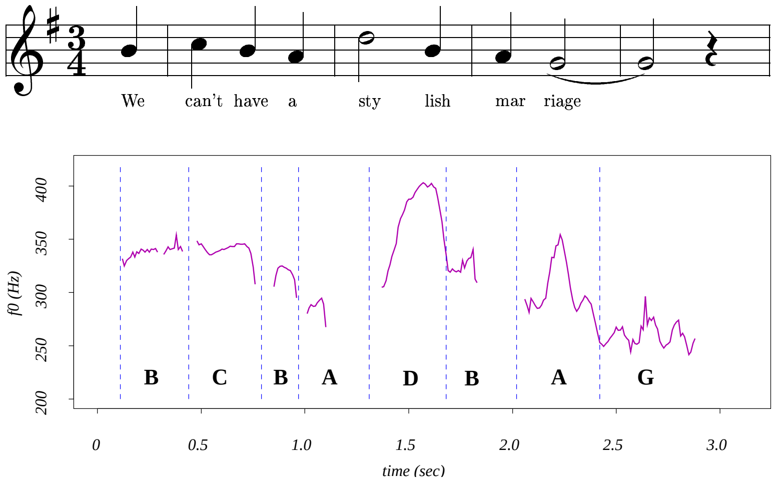


Figure 3: *A musical phrase and it's score.*

attribute), and it fills in the values that lead to the best fit to the data. In this section, we show how this works with a single phrase from Dinah Shore's rendition of *On a Bicycle Built for Two*, originally written by (Dacre, 1892).

Musical scores are under-specified in the sense that performers may have very different renditions based on the same scores. We make use of the musical structures and phrasing notation to insert ornaments (Garretson, 1993) and to implement performance rules, which include the default rhythmic pattern, retard, and duration adjustment (Sundberg et al., 1983; Friberg, 1995)

Indeed, real performances may differ enough from a naive, mechanical interpretation of the score so that even the identification of a note with a particular time interval may be ambiguous or difficult. For example, in Figure 3, none of the musical notes fall on expected frequencies, neither do they show step-like frequency jumps as implied by the musical score, despite the fact that it sounds "in tune."

Given a song and the corresponding musical score, we manually annotate notes and their locations as shown in Figure 3. We place the note boundaries close to the beginning of voicing onset, therefore the half note D is annotated as being shorter than one would expect from the music, because it begins with a voiceless consonant cluster *st.* This definition of note boundary works better with ornament fitting and allows us to align the glide-up ornament ($\triangle$) with the beginning of the voicing onset.

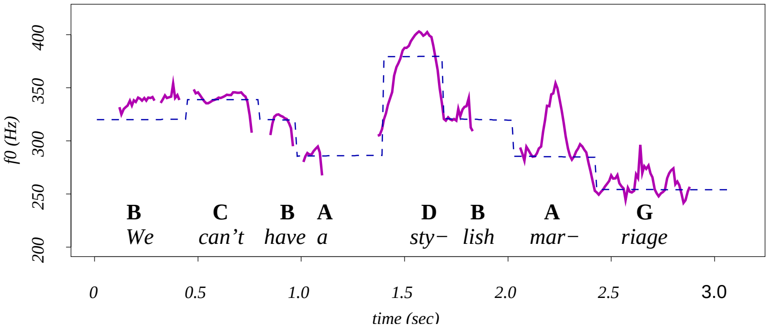
In our Stem-ML models, musical notes are treated analogously to the phrase curve in speech: both are built with *step_to* tags. For music, the *pitch range* is defined as an octave, and each semitone is 1/12 of an octave on a logarithmic scale.

We use the Stem-ML optimizer to find the base frequency, so that we can identify the key and the tuning. With base frequency known, we can then draw in the un-ornamented notes (derived directly from the musical score), study the difference between the performance and the scores and classify the differences into ornaments. Figure 4 plots the f_0 curve of the singing performance in solid lines, and the notes in the score as dashed lines.

We marked locations where a note glides up with $\triangle$ and when a note glides down, we marked $\bigtriangledown$. The *wiggle* shape, perhaps the perceptually most obvious feature, is marked with $\approx$, and occurs near 2.2 seconds in Figures 3 and 4. The pitch undulation on the last note (G) is a vibrato. We handled vibrato separately in our song program, because the neural and physiological mechanisms may be different, so we did not annotate it for the fitting.

Given f_0 and annotations expressed in Stem-ML tags, we again use the optimizer to fit parameter values of shapes and strengths that best describe the observed f_0. We fixed the strength value of the musical *step_to* notes to 8. This large value helps to maintain the specified frequency as the tags pass through the prosody evaluation component. We obtained from the fitting process the best shape for each of the abstract ornament categories $\approx$, $\triangle$, $\bigtriangledown$ (Figure 5), along with the strength values of each instance (Figure 6).

From these annotations, including musical notes, ornament types and fitted strength values, Stem-ML generated the f_0 curve shown in Figure 6. The F_0



Notes	$\Updownarrow_B$	$\Updownarrow_C$	$\Updownarrow_B$	$\Updownarrow_A$	$\Updownarrow_D$		$\Updownarrow_B$	$\Updownarrow_A$	$\Updownarrow_G$
Ornaments	$\bigtriangleup$	$\bigtriangledown$			$\bigtriangleup$	$\bigtriangledown$		$\approx$	

Figure 4: *The difference between singing performance and the musical score.*

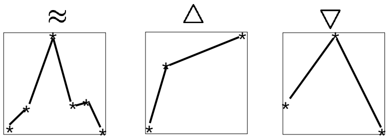


Figure 5: *Best-fit shapes of the musical ornaments.*

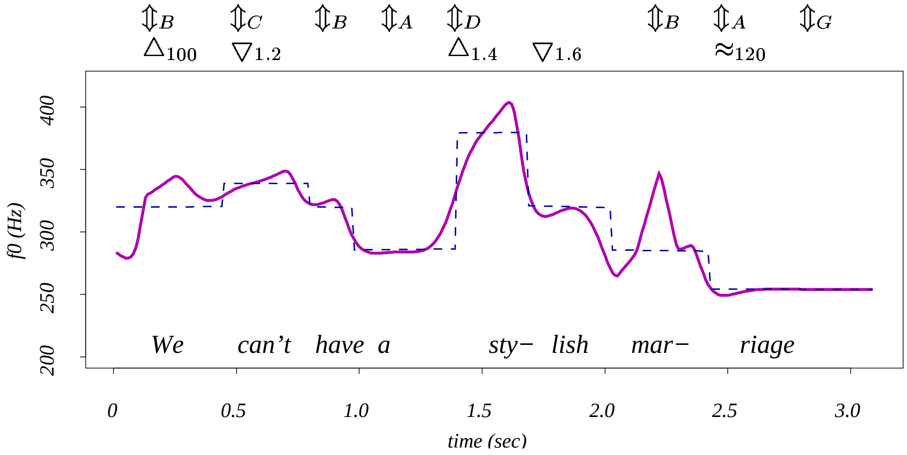


Figure 6: *Ornaments with fitted strength values, and the resulting generated f_0 curve.*

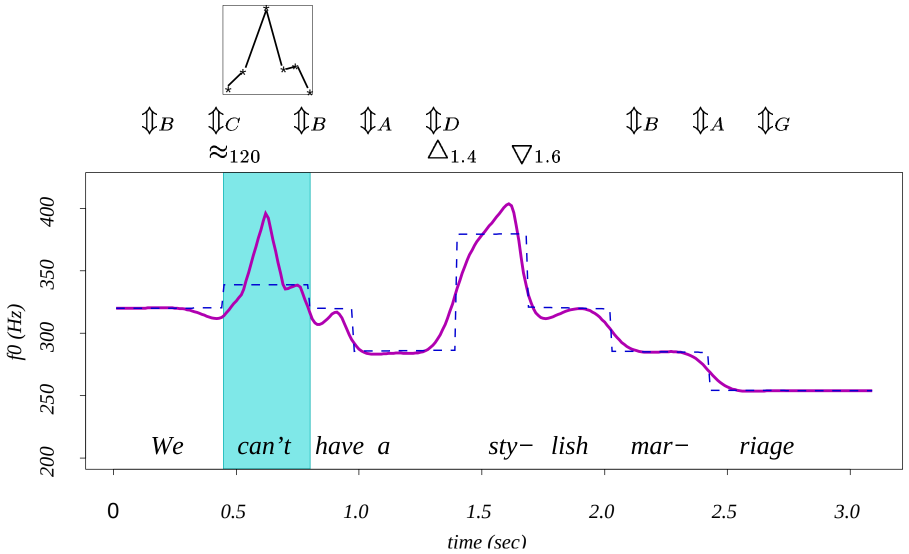


Figure 7: *Moving ornaments around to generate a different performance.*

track re-generated from the codes closely matches the original.

The codes cleanly separate the melody component of the song from the ornaments, resulting in pieces that can be moved around and used as building blocks of new songs. For instance, by moving ≈ from *mar*riage to *can't*, we generated a different rendition of the same musical phrase as shown in Figure 7.

We can follow this method to build a library of musical ornaments. We can change ornaments, shift the ornaments to different locations, or change their strengths to write the song in a different style. Currently, ornament placement is handled by heuristic rules. For example, ≈ is used by Dinah Shore on an accented syllable with a strong beat in a sequence of phrase final descending notes. Changing these rules is part of changing the musical style.

Ornament placement follows musical conventions as well as reflecting a personal style. For instance, placing an ornament on any note gives a melody that can be sung and sounds "natural", but many choices do not make good music. This is not unexpected, as Stem-ML models the low level physiological interactions between tags, but makes no attempt to model aesthetic judgements.

3.2 Musical Ornaments - Changing Amplitude

To model the local amplitude changes seen in Figure 1, we describe the shapes of the amplitude profile with templates the same way as we describe the shapes of the pitch ornaments.

The amplitude control for the first five measures of Shore's *On a Bicycle Built for Two* is shown in Figure 8. A bowtie shaped template ($\bowtie$) is applied to long, non-final notes as on each syllable of the word *Daisy.* A droop template ($\triangleright$) is applied to short notes and to the last note of the phrase to yield a declining amplitude profile.

Figure 9 shows another example of amplitude control in the singing style of Dinah Shore. Shore's ornamental wiggle ($\approx$) has two humps in the f_0 trajectory, where the first f_0 peak coincides with the amplitude valley. We use an amplitude template in tandem with the f_0 template to coordinate these two channels.

The length of the $\approx$ 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.

3.3 Speaking Styles - Phrasal Scope.

In the f_0 traces of typical English sentences from typical speakers, the dominant features reflect word accent and emphasis. The phrasal component, if any, is a smooth decline. Two examples of "normal" speech can be seen (and will be discussed later) in Figures 12 and 13. They are very different from Martin Luther King's distinctive rhetorical style (Figure 2), where word accent and emphasis modifications are also 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 salient features of this style.

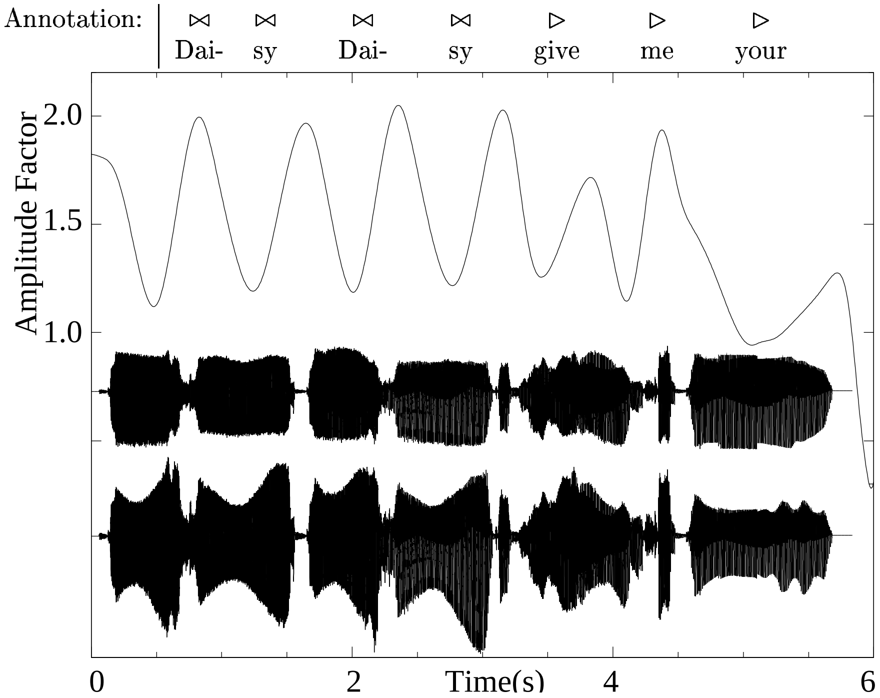


Figure 8: *Amplitude control in synthesized song. Stem-ML is used to produce a time series of amplitude vs. time, which is used to multiply the amplitude profile of TTS-generated sound to implement the style. This figure displays (from top to bottom), the amplitude control time series, speech produced by the synthesizer without amplitude control, and speech produced by the synthesizer with amplitude control.*

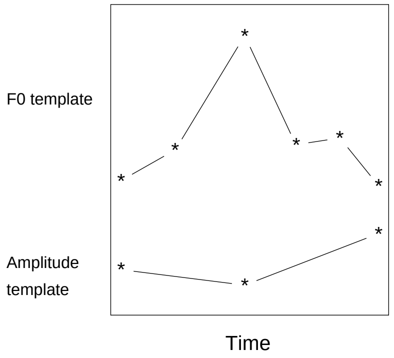


Figure 9: *Templates for the wiggle (≈) ornament. The figure shows the* f_0 *(top) and amplitude (bottom) templates for this ornament.*

King's speech has a strong phrasal component with an outline defined by an initial rise, optional stepping up to climax, then a final fall. To model this style, we use *step-to* tags ($\Updownarrow$) to control the rise and fall in the phrase curve. The argument value of the tag specifies 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 ($\Updownarrow_0$), step 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*, and step up again on the first stressed word of the new phrase. The amount of *step-to* ($\Updownarrow$) correlates with sentence length. Additional stepping up is used on annotated, strongly emphasized words.

The *step-to* tags above produce the phrase curve shown in dotted lines in Figure 10 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.

Figure 11 displays the accent templates used to generate Figure 10. 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.

The generated f_0 contour in Figure 10 is produced by heuristic rules, with the resulting contour noticeably different from King's original in Figure 2, however people still recognize the style despite the differences. This is an example where dominant features of a style can be used successfully in style imitation. The features and rules are portable due to their simplicity. The rules refer to the edges of a sentence or phrase with minor adjustment to sentence length, without resorting to complex information such as sentence structure and the part of speech of words.

$\Updownarrow_{0}$ $\Updownarrow_{1.7}$ $\Updownarrow_{1.85}$ $\Updownarrow_{1.85}$ $\Updownarrow_{0.2}$ $\Updownarrow_{2.0}$ $\Updownarrow_{2.0}$ $\Updownarrow_{0.4}$

This nation ... rise ..., live out ... creed.

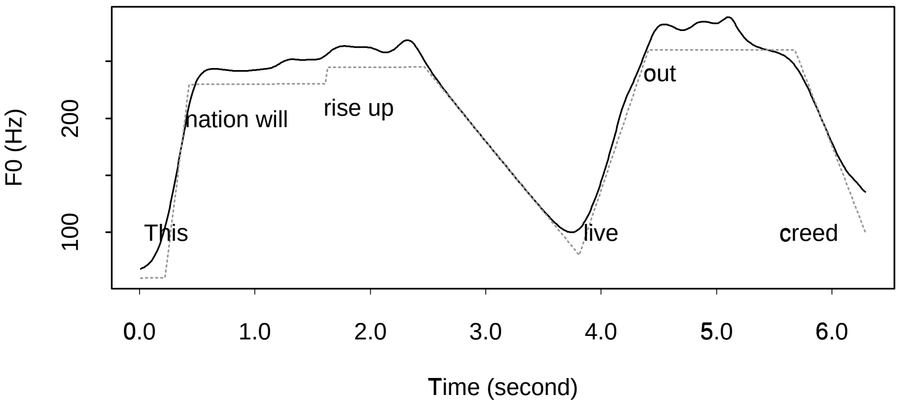


Figure 10: *Generated phrase curve and pitch countour in the style of Martin Luther King.*

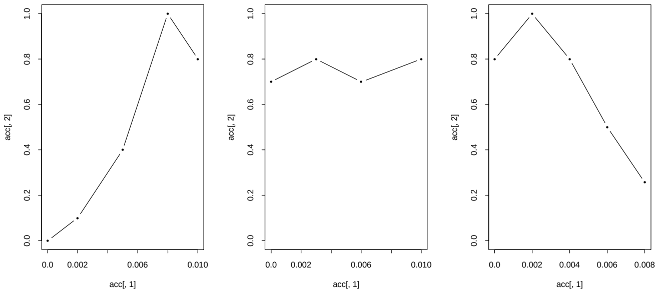


Figure 11: *Accent templates for King's prosody.*

3.4 Speaking Styles - Local Scope

Distinctive speaking styles or moods may be conveyed by idiosyncratic shapes for a given accent type. We examine the DARPA 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 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.

Speakers 1 and 2 in Figures 12 and 13 convey different speaking styles[1] by using distinct rising contours. In both figures, the natural f_0 tracks are plotted in stars and the generated f_0 tracks as solid lines. The distinct accent shapes are captured in the accent templates, which are shown above the figures. We set the scope of the template to be equal to the scope of the word.

Figure 12 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.

Figure 13 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 second speaker's rising slopes are fairly straight, rising from the valley near the center of the word to a peak near the end of the word. The four rising contours in Figure 13 are all generated from the rising template above the figure.

[1] We interpret these differences as stylistic, rather than as different meanings because the speakers are, broadly speaking, making the same request to the system, they both know it is a machine that cannot understand any linguistic subtleties, and because no clear difference in intent could be heard in the recordings.

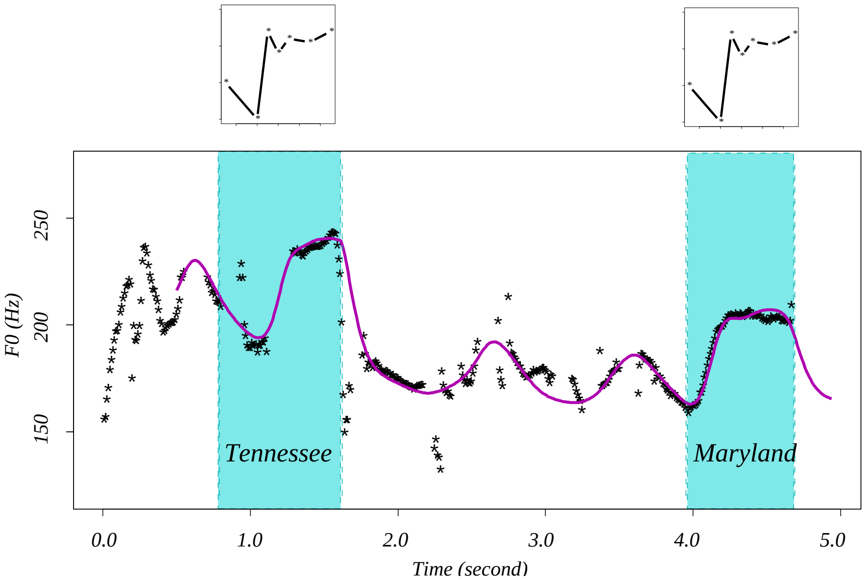


Figure 12: *A sentence from Speaker 1 with two rising accents. "I live in Nashville Tennessee and I'd like to go to Baltimore Maryland."*

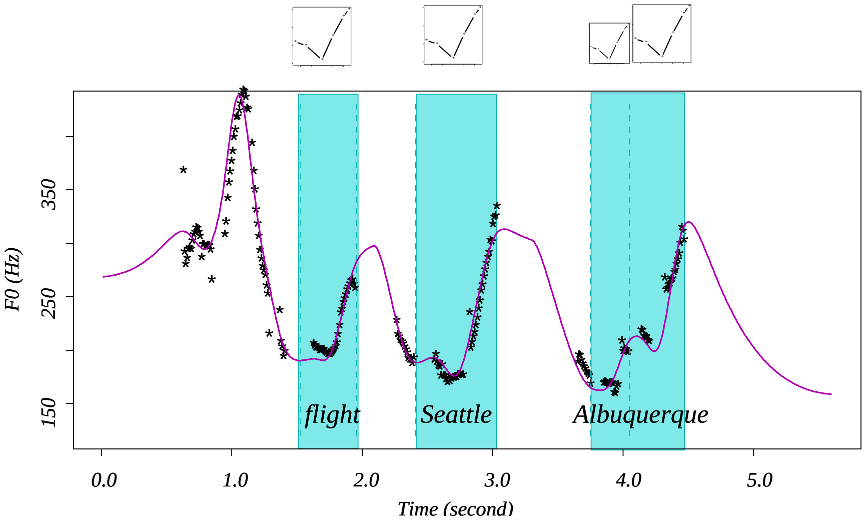


Figure 13: *A sentence from Speaker 2 with multiple rising accents. "Um I would like a flight to Seattle from Albuquerque."*

4 Conclusion

We can represent styles of speech or performance styles in music by a set of distinctive features along with rules to show where the features are placed. With this approach, we can convey the impression of a particular speaker/singer by capturing the most salient prosodic features. We show how to do this with several examples. Speech and song demos are available from http://prosody.multimedia.bell-labs.com .

Practical applications of this technique might include implementation of quotes in news articles, multiple characters in games or dialog systems, or reading email with the prosodic characteristics of the sender.

References

Abe, M. (1997). Speaking styles: statistical analysis and synthesis by a text-to-speech system. In van Santen, J. e., editor, *Progress in Speech Synthesis*, pages 495–510. Springer-Verlag.

Bennett, G. and Rodet, X. (1991). Synthesis of the singing voice. In Mathews, M. V. and Pierce, J. R., editors, *Current Directions in Computer Music Research*, pages 19–44. The MIT Press, Cambridge, Massachusetts.

Bloch, B. (1953). Linguistic structure and linguistic analysis. In Hill, A. A., editor, *Report of the fourth annual round table meeting on linguistics and language teaching*, pages 40–44. Georgetown University Press, Washington. page 42.

Cahn, J. E. (1998). Generating pitch accent distributions that show individual and stylistic differences. In *The ESCA Workshop on Speech Synthesis.*

Dacre, H. (1892). *Daisy Belle, or A Bicycle Made for Two.* Francis, Day and Hunter. music composed by the author.

Dorson, R. M. (1960). Oral styles of american folk narrators. In Sebeok, T. A., editor, *Style in Language*, pages 39, 41. The M.I.T. Press, Cambridge, Massachusetts. ISBN 0 262 69010 1.

Erickson, D., Abramson, A., Maekawa, K., and Kaburagi, T. (2000). Articulatory characteristics of emotional utterances in spoken English. In *ICSLP*, Beijing, China.

Friberg, A. (1995). *A Quantitative Rule System for Musical Performance.* PhD thesis, Royal Institute of Technology (KTH), Sweden.

Garretson, R. (1993). *Choral Music: History, Style, and Performance Practice.* Prentice Hall.

Higuchi, N., Hirai, T., and Sagisaka, Y. (1997). Effect of speaking style on parameters of fundamental frequency contour. In van Santen, J. e., editor, *Progress in Speech Synthesis*, pages 417–428. Springer-Verlag.

Jackson, A. (1999). In *The Dinah Shore Collection, Columbia and RCA recordings, 1942-1948.* Vocalion, Watford, Hertfordshire, England. Liner notes.

Kitahara, Y. and Y., T. (1989). Prosodic components of speech in the expression of emotion. *JASA*, 84.

Kochanski, G. P. and Shih, C. (2000). Stem-ML: Language independent prosody description. In *Proceedings of the 6th International Conference on Spoken Language Processing*, Beijing, China.

Kochanski, G. P. and Shih, C. (2002). Soft templates for prosody mark-up. *Accepted by Speech Communications.*

Macon, M. W., Jensen-Link, L., Oliverio, J., Clements, M., and George, E. B. (1997). A system for singing voice synthesis based on sinusoidal modeling. In *Proc. of International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 435–438.

Maekawa, K. (1999). Phonetic and phonological characteristics of paralinguistic information in spoken Japanese. In *International Conf. Sp. Lg. Proc.* no. 0997.

Monaghan, A. I. C. and Ladd, D. R. (1991). Manipulating synthetic intonation for speaker characterization. In *ICASSP*, pages 453–456.

Shih, C. and Kochanski, G. P. (2000). Chinese tone modeling with Stem-ML. In *Proceedings of the sixth International Conference on Speech and Language Processing*, Beijing, China.

Shore, D. (1999). Bicycle built for two. In *The Dinah Shore Collection, Columbia and RCA recordings, 1942-1948.* Vocalion, Watford, Hertfordshire, England. Compact disc 2DCUS3003, 65837 30032.

Sundberg, J., Askenfelt, A., , and Frydén, L. (1983). Musical performance: A synthesis-by-rule approach. *Computer Music Journal*, 7:37–43.