INTERNATIONAL JOURNAL OF SPEECH TECHNOLOGY 6, 393–408, 2003

Modeling of Vocal Styles Using Portable Features and Placement Rules

CHILIN SHIH AND GREG KOCHANSKI
Bell Laboratories, Lucent Technologies, Murray Hill, NJ, USA
cls@prosodies.org
gpk@alum.mit.edu

Abstract. This paper presents a mathematical description of style in speech and singing. These styles are represented as a set of portable prosodic features along with a set of rules to choose where the features are to be applied. Speakers and singers make creative choices to express their personal style, which may involve specific phrase curve, accent shape, or, similarly, musical embellishment. Therefore a quantitative model of style needs to support unconstrained accent and phrase curve description, and to solve potential conflicts that arise from this freedom. Our current implementation modifies two acoustic parameters: f_0 and amplitude. We use an articulator-based model, Stem-ML, to resolve conflicts between intended accents or embellishments and their environment. We present several examples to illustrate the modeling of accents and phrase curves, as well as the usefulness of style/content separation, and the similarity between speech and music.

Keywords: speech synthesis, Stem-ML, quantitative models, pitch, accent, music

1. Introduction

The sense of a style can be expressed in terms of recurrent, salient features. These salient features are often rare relative to a random sampling of speech or song, or are distributed in atypical patterns. The features are not normally placed arbitrarily, but instead have a close relationship to the underlying content and structure. This relationship can be expressed by saying there are a set of rules which specify where to place the salient features, given the underlying content and structure.

Matching a style does not require everything to be similar, only the salient features and their patterns of use. For instance, an impersonator or comedian 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.

The feature/location description of style we adopt here is similar to that presented in Bloch (1953), and is applicable to a broad range of speech, art and music. For example, story-telling styles can be described at a high level in terms of features and location rules:

> One stylistic device in this tale, employed as a connective between the episodes ... 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).

Prosodically, much of the style of a speaker can be expressed in terms of features in f_0, amplitude, spectral tilt, and duration (Murray and Arnott, 1993; Higuchi et al., 1997; Schroder, 2001). In this paper, we are concerned with low-level prosodic styles that can be implemented fairly directly in terms of acoustic parameters such as f_0 and amplitude. For example, for speech, this paper discusses the detailed rendition of the intonation of a phrase after the words have been chosen. For music, we model performance factors that are not part of the musical score.

We treat prosody and music together because it is desirable to have a unified model. The existence of intermediate vocal forms between speech and singing implies that speech, singing, and intermediate forms should all be treated by variants of the same model. Pragmatically, a unified model also allows us to model both speech and singing using the same algorithms and similar parameters. It also allows us to model mixtures of song and speech.

Our goal is to present some techniques that provide a mathematical description of style in speech and singing. The techniques allow us to separate recurrent, salient features that define a style from the textual content, and then later to place them where needed when speech or song is synthesized.

Expression of style is a creative process where speakers and singers introduce a wide range of individualized prosodic modifications. We find it useful to have modeling support for both local and global style features. We review and discuss the prosody representation issues in Section 2.

The prosodic model needs to have the flexibility to handle unconstrained style variations, and must resolve any potential conflicts. We discuss the mathematical basis of the model in Section 3, where we allow unconstrained representation of accent and phrase curve, and resolve potential conflicts with an articulatory-based model. Some of the salient features are local, which we will capture with the Stem-ML *stress* tag (Section 3.2). Some other features involve changes over a broad scope, such as an entire phrase of speech, which we capture with the Stem-ML *step_to* tag. We then follow with case studies presented in Section 4.

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

2. Strategies for Representing Intonation

In both singing and speech, there are strong arguments for representing pitch as a set of accents or embellishments that can be placed in arbitrary combinations on top of a background. In music, we treat the score as one component, and embellishments as another. Thus, a performance is treated as beginning with a mechanical, precise, naïve interpretation of the score, which is then transformed to a professional, artistic performance by adding embellishments and adjusting duration. In speech, we treat the phrase curve as one component, and accents as another. The phrase curve describes the larger structures, and depends primarily on how the text is broken up into phrases by pauses, while accents are associated with particular words. We draw an analogy between speech and music by modeling musical embellishments in the same way as accents in speech, and treating the phrase curve in the same way as the mechanical melody derived directly from a musical score. In both music and speech, we wish to be able to describe enough detail to convey performance styles.

Figure 1 shows the f_0 trace of phrases from the speech "I have a dream" delivered by 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. We suggest that the general shape of this phrase curve is the signature style of Martin Luther King. The pitch profile is found in many phrases in the speech, even though the phrases differ in textual content, syntactic structure, and length. On top of the phrase curve, one can identify the pitch movement due to accents associated with words. To capture the details it is desirable to recognize both the phrasal and accent components and model them accordingly.

Fujisaki's model (Fujisaki, 1983, 1988) treats surface sentence intonation as the combination of two components: phrase commands and accent commands. These commands are filtered by the muscle's time response (which is assumed to be a time-independent kernel), and added in the log scale to yield the surface pitch trajectory. His phrase command is modeled after the phenomenon of declination, which is most suitable for declarative sentences. To describe a phrase curve that deviates from the declination shape, one needs to use extra pulses that are not easily linked to linguistic attributes. Later models such as that of van Santen and Möbius (2000) also have a rigid view of the phrase curve, allowing few possibilities for the shape of phrase curve in the implementation. We see a need to develop this idea further, both to include the capability to implement unrestricted variations of the phrase curves in speech, and to describe music scores.

In contrast to the above models which build an f_0 curve by superimposing multiple components, there is also a tradition of single-component models for f_0 (Liberman and Pierrehumbert, 1984; Hirst et al., 2000; Taylor, 2000) that do not decompose it into phrase curves and accents. Under this view, all pitch movement in a sentence is accounted for by a linear string of

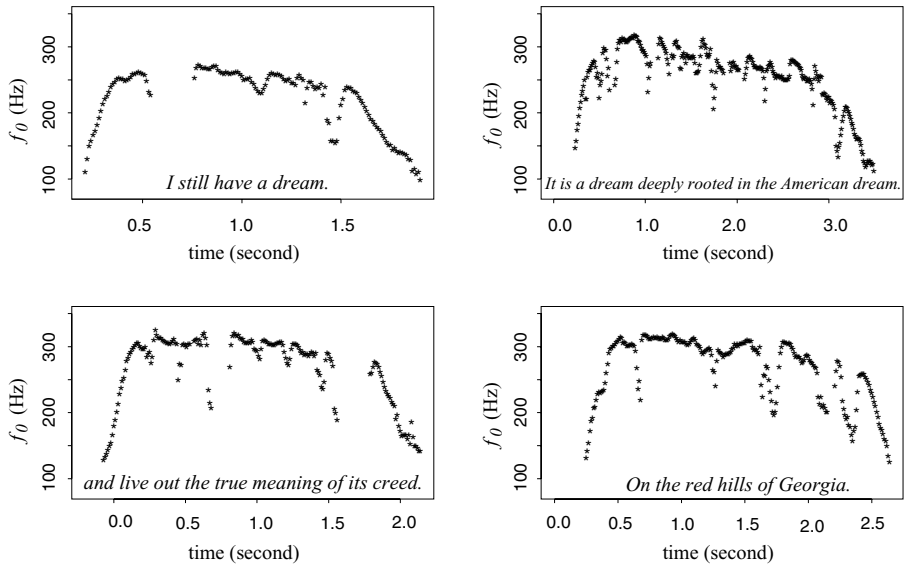

Figure 1. Phrasal f_0 profiles from the speech of Martin Luther King Jr.

accents. Pitch movement as shown in Fig. 1 will then be represented in the standard ToBI transcription convention (Silverman et al., 1992; Beckman and Ayers, 1997) as having a strong rising accent (L*+H) near the beginning, a strong falling accent (H*+L) near the end, followed by a phrase accent (L−) and a boundary tone (L%). This transcription assumes that all medial words are de-accented, and the high pitch plateau is a result of pitch interpolation. This is fine as a first approximation, but one can hear pitch movements in this section which are due to word accents, and these word accents cannot be captured in a ToBI transcription that shows the phrasal structure. We note that ToBI doesn't begin to provide a useful transcription of music.

Multi-component modeling of music is valuable because it allows us to capture the transient nature of embellishments, adding, deleting, and moving them independently of the musical scores. One may argue that embellishment can be expanded and written into music scores, therefore rendering the two components as one, but once a embellishment is turned into a sequence of notes, one loses the distinction between melody and embellishment. This restricts the ability to change performance styles.

The need for multi-component modeling for speech is accepted by some recent works that incorporate pitch range modeling into the ToBI framework (Jilka et al., 1999; Möhler and Mayer, 2001), by relaxing the strictly linear sequence view of intonation modeling. There are implementation constraints in these works, allowing only uniform compression or expansion of the pitch range across the whole intonation phrase. This kind of approach will not work cleanly for King's speech, as the wide swings of the rise and fall at the edges of a phrase are closely related to the compressed pitch range in the center of each phrase.

If we set our goal to be capturing personal styles by modeling music embellishment and speech accents, traditional techniques of accent modeling using a fixed inventory of pre-defined accents (Anderson et al., 1984; Jilka et al., 1999) will not be sufficient. So, we propose a model where there is no restriction on accent shape, and allow the user to define accent shapes as they see fit.

Unrestricted accent shapes, combined with the possibility that accents can be placed anywhere opens up the possibility of conflicting requests. What would happen if at a given time, one accent wants to be high and its neighbor wants to be low? What if accents overlap?

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 explain the mathematical basis of resolving accent conflicts. We then show examples of how to use Stem-ML on speech and music.

We start by describing a phrase from Dinah Shore's singing to illustrate the procedure of annotation,

automatic fitting and generation. We then discuss the modeling of amplitude profile, phrase curve, and accents. Similar features can be used to support other stylistic variations and emotional speech (Monaghan and Ladd, 1991; Abe, 1997; Cahn, 1998). Our singing synthesis focuses on style and performance rules rather than on voice quality (Bennett and Rodet, 1991; Cook, 1991; Macon et al., 1997). Note that this paper is an expanded version of Shih and Kochanski (2001).

3. Describing Prosody with Stem-ML

In this paper, the control of pitch and amplitude in speech and song is achieved by using Stem-ML tags (Soft TEMplate Mark-up Language) (Kochanski and Shih, 2000, 2003; Kochanski et al., 2003). Stem-ML provides prosody mark-up tags that can be correlated with linguistic features, and which have approximately local effects on acoustic parameters. The tags are mathematically defined, along 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.

3.1. Background

We rely heavily on two of the Stem-ML features to describe 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* (⇳), sets the pitch to a specified value (interpolating between *step_to* tags, as needed). When it is described by a sequence of *step_to* tags, the phrase curve is a piece-wise linear function. We use this method to describe both Martin Luther King's phrase curve, and notes in music.

Secondly, Stem-ML separates the placement of accents 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 between neighboring accents and between accents and the phrase curve. 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 embellishment templates for music. Additionally, it allows the heuristic rules for accent placement to be simple and clean, because the rules do not have to work around limitations concerning which accents can follow what. From a linguistic point of view, the flexibility of the phonetics allows for a simpler phonology.

Some combinations of accent and embellishment templates may conflict or be impossibly difficult to realize precisely; Stem-ML accepts conflicting specifications and returns a smooth surface realization that best satisfies all constraints.

The muscle motions that control prosody are smooth (i.e., they have finite first and second time derivatives) because muscles are physical objects and cannot accelerate instantaneously (cf. Huxley, 1957). We observe that when a section of speech material is unimportant, the speaker may not expend much effort to realize the targets (Lindblom, 1963; Shih and Kochanski, 2000). In general, the speaker is simultaneously trying to do several incompatible things: He wants to carefully produce the correct pitch contour so that the listener will understand. He is forced (by his own muscles) to generate a smooth pitch contour. Finally, he generally wants to execute the speech with minimum effort. Our model recognizes that sometimes one goal wins, sometimes the other, depending on the relative importance of the goals, but typically the result is a compromise. Loosely speaking, we assume that the speaker balances the effort required to speak against the possibility of being misunderstood.

3.2. Mathematical Definition

We start out by assuming that speakers and singers have in their mind a set of ideal targets which they are trying to communicate to the listeners. Some of these targets may correspond to local movements such as tones, accents, or music embellishments, and some to non-local movements such as phrase curves and musical scores.

Local movements typically have clear linguistics functions. We describe them with the Stem-ML *stress* tag. The *stress* tag specifies the accent component, and it normally corresponds to a syllable or a word; the phrase curve is specified by several *step_to* tags. The most important attribute of the *stress* tag is the *shape* template, which draws the ideal shape of an accent. The *stress* tag can define the pitch at one or more points, and so can be used to implement slopes, peaks, or valleys, in addition (or instead of) specifying a particular pitch at a particular time. The *stress* tag has other attributes to be explained momentarily, such as *strength*, *type* and *atype*, that control the way the specified *shape* is realized in different environments.

In this work, each pitch target, y_i, consist of an accent component (embellishment) added to the underlying phrase curve (musical score):

$$y_i = P + Y_i \cdot atype_i \cdot {s_i}^{|atype_i|} \quad (1)$$

where P is the phrase curve, Y_i is the shape of the ith accent, and $atype \cdot {s_i}^{|atype|}$ is a scale factor on the ith accent's pitch range, which normally expands the range of high strength accents. We assume that the *atype* parameter is shared amongst all instances of a particular accent (embellishment); it controls how the pitch range of the template scales with the tag's strength (Eq. (1)). Note that y_i and P and Y_i are all functions of time, and we have suppressed the t subscript, for clarity. P is defined across the entire phrase, but y_i and Y_i are defined only over the scope of the tag.

These targets are subject to performance constraints during production. We can represent the surface realization of prosody as the solution to an optimization problem, minimizing the sum of two functions: a physiological constraint G, which forces the pitch curve to be smooth, and a communication constraint R which is the sum of errors r_i between the pitch and the pitch target (y_i) for each tag:

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

The "effort" term, which, when the sum is minimized, forces the solution to be smooth and continuous.

$$R = \sum_{i \in \text{tags}} s_i^2 r_i \quad (3)$$

The "error" term. This is a weighted sum of template-by-template errors.

$$r_i = \sum_{t \in \text{tag}i} \alpha((p_t - \bar{p}_i) - (y_{i,t} - \bar{y}_i))^2 + \beta(\bar{p}_i - \bar{y}_i)^2 \quad (4)$$

The error in template i; this drives the pitch, p, to be close to the target, y.

The errors are weighted by the *strength*, s_i, of each *stress* tag, which indicates how important it is to satisfy the specifications of the tag. We do not claim that G provides a detailed representation of muscle behavior, but it captures the damped mass-and-spring dynamics of real muscles, and provides results similar to the classic Hill (1938) model of muscle behavior.

If the strength of a tag is low, the physiological constraint dominates, and smoothness is more important than accuracy. Each tag's strength 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.

Stress tags also have a parameter which controls whether errors in the shape or average value of the pitch relative to the target is more important. (This is the Stem-ML *type* parameter.) We write this parameter as $\alpha = \cos(type \cdot \pi/2)$ and $\beta = \sin(type \cdot \pi/2)$, so $\alpha^2 + \beta^2 = 1$.

In Eqs. (2) to (4) above, p_t is the normalized pitch at time t, that is, the pitch relative to the speaker's normal range. Also, $\bar{p}_i$ is the average of p over the scope of tag i, and $\bar{y}_i$ is the average of y_i over its scope.

For the speech modeling, we simply scale p to get f_0: $f_0 = p \cdot range + base$, where *range* and *base* are speaker-dependent constants that give the normal range of f_0 variation and the speaker's typical f_0. For the singing examples, we use an exponential scaling to make defining the phrase curve (i.e., the notes) more convenient: $f_0 = 2^{(p/12)} \cdot base$. The range of f_0 in the examples presented here is small enough so that the two representations are not too different.

3.3. Notation

Local movements such as accents, tones, and musical embellishments are described by Stem-ML *shape* templates in the *stress* tags. In this paper, we define and use bow-tie ($\bowtie$), wiggle ($\approx$), rise ($\triangle$), fall ($\triangledown$) and droop ($\triangleright$) shapes. Each shape is specified as line segments connecting a set of points $[(x_1, y_1), (x_2, y_2), \ldots]$, and α (see Eq. (4)). The subscript in $shape_{strength}$ specifies the strength of the tag, which is the s_i in Eqs. (1) and (3). These can be used to describe word accents in speech and embellishment in singing. Each tag has a scope (over time), 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 4.1, 4.2, and 4.4, 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 ($\updownarrow$), such that $\updownarrow_{value}$ pins the phrase curve to *value* at the time of the tag, and the pitch will follow. Section 4.3 shows an example describing larger scope features such as a phrase curve with Stem-ML *step_to* tags.

4. Examples

In this section, we give several examples illustrating the use of Stem-ML.

4.1. Musical Embellishments—Changing Pitch

We use Stem-ML in two directions, both to evaluate prosody from tags and, in reverse, to deduce the values of numerical parameters of tags from the data. The Stem-ML evaluation component takes tag and attribute values as input and generates time series data such as f_0 or an amplitude curve. The Stem-ML optimizer takes data and partial tag annotation as input, and it finds the best description of the data in terms of the tags' parameters. One feeds it Stem-ML tags with free parameters (e.g., a tag with an undetermined strength attribute), and it finds the values of the parameters that lead to the best fit to the data. We show here how this works with a single phrase from Dinah Shore's rendition of *A Bicycle Built for Two*, originally written by Dacre (1892).

This song is historically important in the text-to-speech synthesis tradition (Olive, 1998). John Kelly and colleagues at Bell Labs synthesized *A Bicycle Built for Two* in the early 1960's (Mathews, 1963). It was the first computer synthesized song. The work was the inspiration behind the movie *2001: A Space Odyssey* (Kubrick, 1968), where the rebellious computer HAL was singing this song as he was being disconnected, claiming (historically correctly) that this is the first song his master taught him. We chose Dinah Shore's recording because she gave several variations of the same song, with light accompaniment, so that f_0 and amplitude could be reliably extracted.

Musical scores do not completely specify the sound, 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 embellishments (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 naïve, 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 Fig. 2, 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 the performance is pleasant and sounds in tune.

Given a song and the corresponding musical score, we manually annotate notes and their locations as shown in Fig. 2. 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 embellishment fitting and allows us to align the glide-up embellishment (△) with the beginning of the note.

In Stem-ML models, musical notes are treated analogously to the phrase curve in speech: both are built with *step_to* tags. For music, the Stem-ML *pitch range* is set

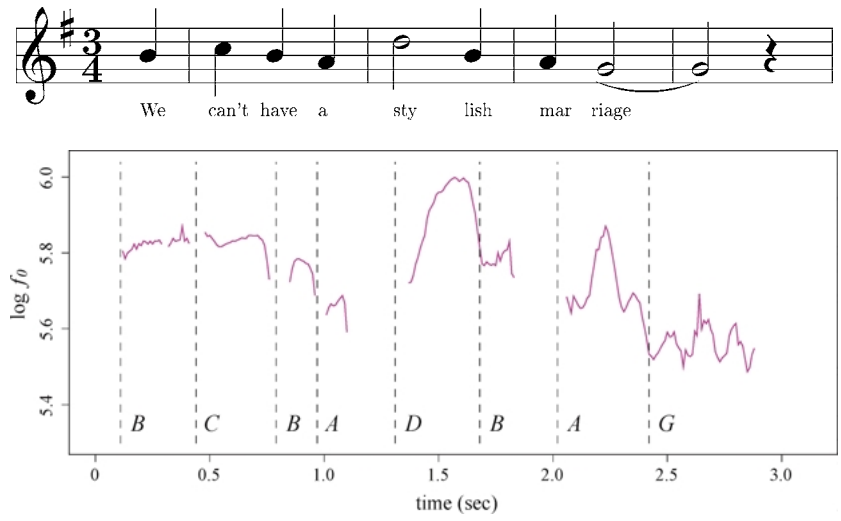

Figure 2. A musical phrase and its score.

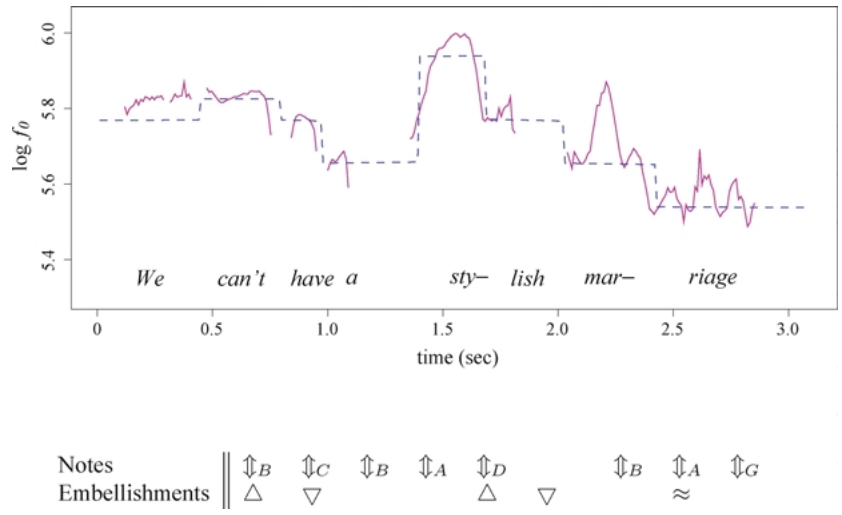

Figure 3. The difference between singing performance and the musical score.

to be an octave and we use an exponential mapping between the Stem-ML *strength* values and f_0. Note that the *pitch range* doesn't limit the pitch: It merely sets the scaling.

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-embellished notes (derived directly from the musical score), study the differences between the performance and the scores, and classify the differences into embellishments. Figure 3 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 $\triangledown$. The *wiggle* shape, perhaps the perceptually most obvious feature, is marked with $\approx$, and occurs near 2.2 seconds in Figs. 2 and 3. We avoid conventional terms for these embellishments, because we wish to avoid the rigid definitions of musical ornaments. For instance, the *wiggle* shape is similar to a classical inverted mordent, but without any particular intervals, and it can have freer movement. 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

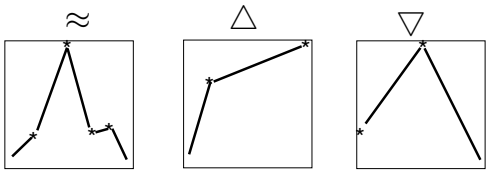

Figure 4. Best-fit shapes of the musical embellishments.

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 embellishment categories $\approx$, $\triangle$, $\triangledown$ (Fig. 4), along with the strength values of each instance (Fig. 5). From these annotations, including musical notes, embellishment types and fitted strength values, Stem-ML generated the f_0 curve shown in Fig. 5.

The training cleanly separates the melodic component of the song from the embellishments, resulting in tags that describe portable embellishments that can be moved around and used as building blocks of new renditions. For instance, by moving $\approx$ from marriage to *can't*, we generate a different rendition of the same musical phrase as shown in Fig. 6.

We can follow this method to build a library of musical embellishments. With such a library, we can change embellishments, shift the embellishments to different locations, or change their strengths to write the song in a different style. Currently, embellishment placement is handled by heuristic rules. For example, $\approx$ is used by Dinah Shore on an accented syllable with a strong beat, in a sequence of phrase final descending notes.

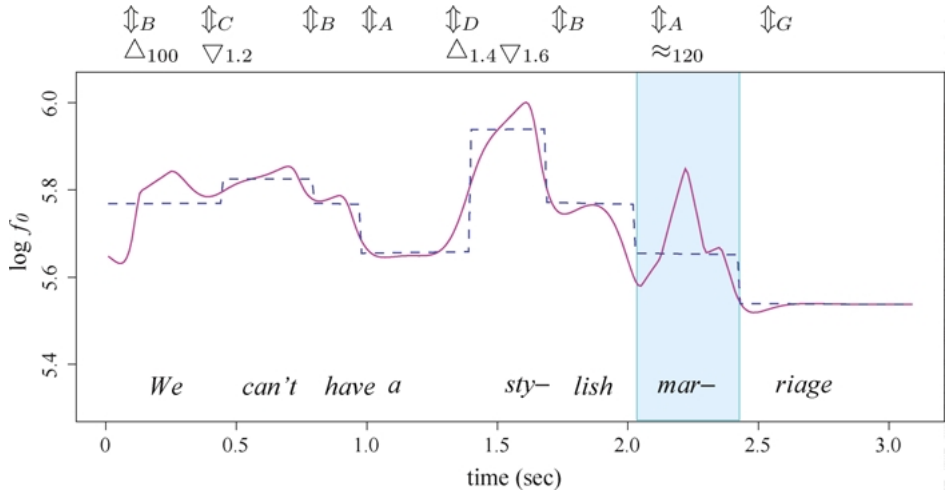

Figure 5. Embellishments with fitted strength values, and the resulting generated f_0 curve.

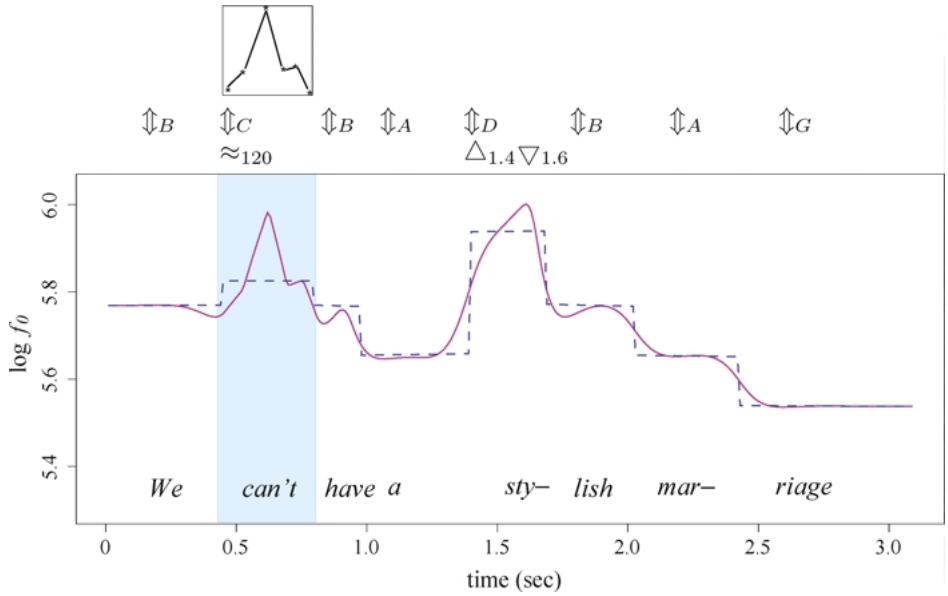

Figure 6. Moving embellishment around to generate a different performance.

Changing these rules is part of changing the musical style.

When deciding where to place an embellishment, one should follow musical conventions as well as reflecting a personal style. For instance, placing an embellishment 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, because Stem-ML models the low level physiological interactions between tags, but makes no attempt to model aesthetic judgments.

Shore's wiggle (≈) also has characteristic amplitude profile. This embellishment 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. Figure 7 shows these two templates on the same time axis.

Shore sang nine wiggles in the three variations of the song *On a Bicycle Built for Two*. These renditions had different tempos, keys, and improvisation, thus providing an interesting range of contexts for this

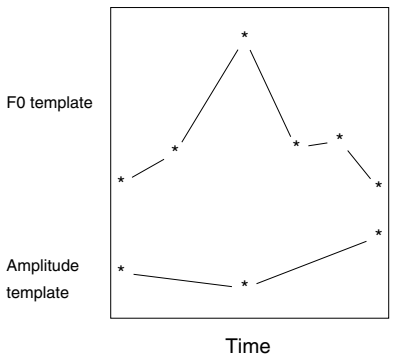

Figure 7. Templates for the embellishment wiggle (≈). The figure shows the f_0 (top) and amplitude (bottom) templates for this embellishment.

embellishment. Table 1 lists the duration of these wiggles in seconds, the length of the measure, the estimated length of the note, and the textual context.

Since the note boundary in the signal is ambiguous and difficult to label with standard tools such as Xwaves (ESPS/Waves, 2002), we asked three subjects to tap the beat (one tap per measure) while listening to the music. Subjects can perform this task easily, and it gives a clear mark of the boundary of each measure. Subjects also agree on their transcriptions of note-to-measure ratio (e.g., 2-to-1 or 3-to-1). We then estimate the note length from the measure length and the transcription of the note-to-measure ratio.

We fit a regression model predicting the length of wiggle from the consonant voicing status and from the length of the measure and the length of note respectively:

$$wiggle = -0.05\text{s} + 0.42 \cdot measure + 0.07 \cdot voice \tag{5}$$

The prediction based on measure length and voicing is better (Pearson's $r = 0.63$) than the prediction from note length and voicing (Pearson's $r = 0.35$).

The fit implies that the length of a wiggle is longer at slower tempos. Voicing also has an effect; a voiceless onset to a note shortened the wiggle's length. It is interesting that one can predict the length of wiggle better from the tempo than from the note to which it is applied. It appears that there is a minimum length requirement of this embellishment. If the note length is too short it lengthens to accommodate the embellishment.

4.2. *Musical Embellishments—Changing Amplitude*

Figure 8 shows the amplitude profiles of the first four syllables *Dai-sy Dai-sy* in our example by Dinah Shore. She merged a de-crescendo and crescendo in the same note, creating a bow-tie-shaped amplitude profile (The second syllable, centered near 1.2 seconds, is the clearest example.). The decrease of amplitude in the middle of a note contrasts with notes from most singers. For instance, Fig. 9 shows the more even, slowly changing amplitude profile of another singer. The 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 distinct style.

To model the local amplitude changes seen in Fig. 8, we describe the shapes of the amplitude profile with templates the same way as we describe the shapes of the pitch embellishments. The same modeling techniques are applicable, because (at least during vowels, and if one normalizes for the vowel), the amplitude is primarily controlled by the sub-glottal pressure (Strik and Boves, 1992), and that pressure is controlled in turn by the dynamics of the chest, diaphragm, and abdominal muscles.

A note should have at least two beats to allow sufficient time to realize this pattern (minimally one beat for de-crescendo and one beat for crescendo). In all of observed cases the note starts as the first beat of the measure. In addition, Shore didn't use bow-tie on notes

Table 1. Lengths of wiggle embellishments in *Daisy* as they relate to the duration of the notes and measures that contain them.

	$\approx_1$	$\approx_2$	$\approx_3$	$\approx_4$	$\approx_5$	$\approx_6$	$\approx_7$	$\approx_8$	$\approx_9$
Wiggle	0.33 s	0.30 s	0.34 s	0.37 s	0.28 s	0.32 s	0.44 s	0.45 s	0.37 s
Measure	0.82 s	0.80 s	0.79 s	0.80 s	0.80 s	0.99 s	0.98 s	0.91 s	0.91 s
Note	0.27 s	0.26 s	0.40 s	0.26 s	0.26 s	0.50 s	1.47 s	0.30 s	0.30 s
Text	marr(iage)	carr(iage)	dai(sy)	do	carr(iage)	for	love	marr(iage)	carr(iage)

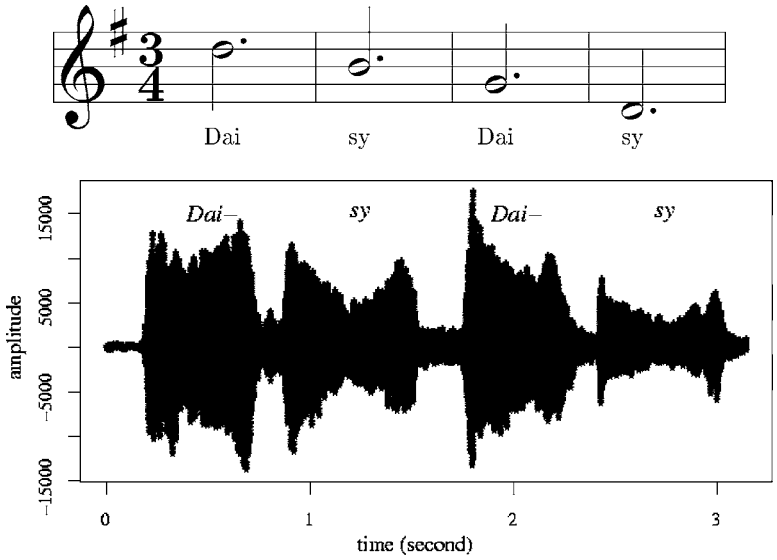

Figure 8. Dinah Shore's signature amplitude profile.

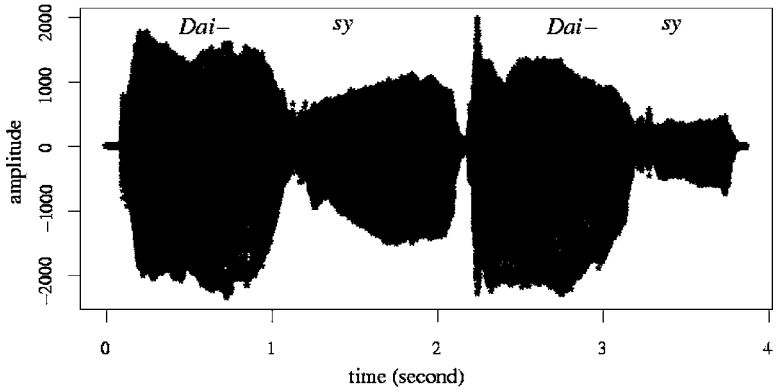

Figure 9. Amplitude profile of another singer.

with heavy pitch embellishment, vibrato, or on the very last note.

The bow-tie amplitude is used frequently. Out of the 50 notes in the first repetition of the song, there are 22 notes that are 2 beats or longer. Among them, 14 have the bow-tie amplitude profile. In the second repetition, Shore inserted words which shorten some notes. Consequently, the number of long notes is reduced to 16, out of which 10 have the bow-tie amplitude pattern. The third repetition is in slow tempo where Shore opted for crescendo and vibrato instead of bow-tie on long notes. Out of 24 candidates, 7 have the bow-tie pattern. In contrast, we didn't find any bow-tie amplitude profile in the recording of the same song by two other singers, one amateur and one professional.

The amplitude control for the first phrase of Shore's *On a Bicycle Built for Two* is shown in Fig. 10. A bow-tie shaped template ($\triangleright\!\!\triangleleft$) is applied to long notes as on each syllable of the word *Daisy*, the stressed syllable of *answer*, and the final note *do*. A droop template ($\triangleright$) is applied to short notes.

4.3. Speaking Styles—Phrasal Scope

In this section, we switch to speech, exploring a way to model Martin Luther King's distinctive style. Technically, much of the style is carried by the phrase curve, which we control in the same way as we control music scores. The combination of accent and phrase curve

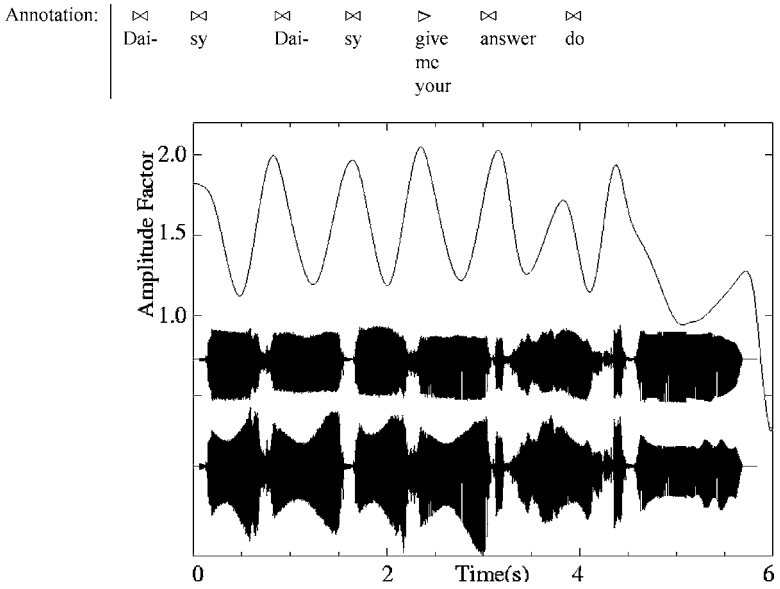

Figure 10. Amplitude control in synthesized song. Stem-ML is used to produce a time series of amplitude vs. time, which is used to multiply the amplitude profile of TTS-generated sound to implement the style. This figure displays (from top to bottom), the amplitude control time series, acoustic waveform produced by the synthesizer without amplitude control, and acoustic waveform produced by the synthesizer with amplitude control.

is the same as the combination of music scores and embellishments.

In the f_0 traces of typical English sentences from typical speakers, most of the f_0 movements reflect word accent and emphasis. The phrasal component, if any, is a smooth decline. They are different from Martin Luther King's rhetorical style (Fig. 1), where 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 salient features of King's style.

Figure 11 shows a set of histograms comparing snapshots of the phrase curves of Martin Luther King Jr. and another professional speaker (J). Speaker J's data are presented in the left column and King's data in the right column.

Each histogram shows the distribution of 10 voiced f_0 samples collected from different regions of phrases, where phrase is defined as speech signal followed by at least 250 ms of silence. Samples of f_0 were taken every 10 ms, and we excluded voiceless regions, so each region is at least 100 ms long. The rows, from top to bottom, show the changes of f_0 patterns as time progresses. The picture shows two distinct patterns of f_0 usage and their sensitivity to phrasal positions.

The plots show several regions of interest: the first 10 voiced samples of a phrase, from the 30th to the 40th samples, 10 samples from the mid point of the phrase, and the final 10 samples of the phrase. All sentences are long—therefore, the mid point always comes after the 30th sample.

The speech materials from both speakers are continuous. King's speech includes 12 minutes of *The American Dream* (King, 2000). Speaker J's speeches were movie critiques and commentaries. There are around 35,000 f_0 samples in each database. The two speakers have similar pitch range spanning from 50 to 300 Hz, but with very different patterns of f_0 usage.

Speaker J's pattern, shown on the left, exhibits a broad distribution in f_0 ranges in all but the final positions. The middle region has lower range than the earlier regions, which is consistent with declination effect and downstep effect (Fujisaki, 1983; Pierrehumbert, 1979). The final region is markedly lower than previous regions, where most of the f_0 samples are below 100 Hz. This pattern is consistent with the final lowering effect and the sentence final low boundary tone (Liberman and Pierrehumbert, 1984).

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. His initial and final f_0 patterns are similar, both dominated by f_0 values around

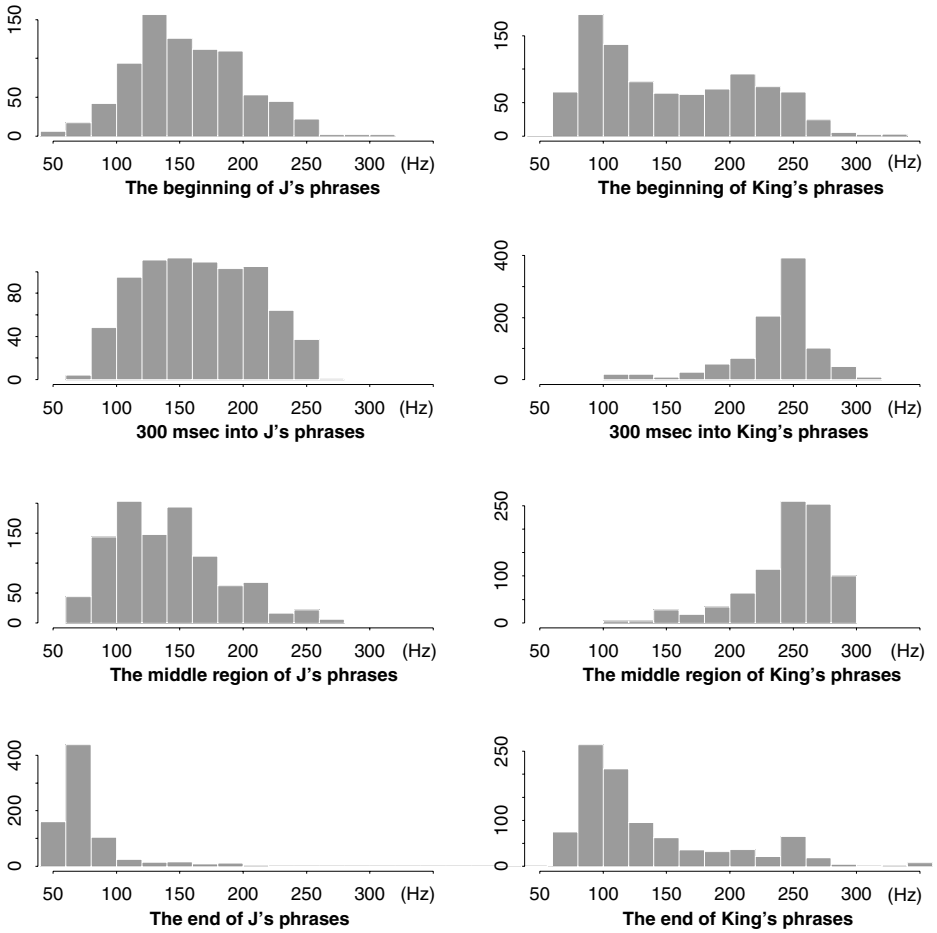

Figure 11. Histograms showing the contrast between Martin Luther King Jr. (right) and a professional speaker J (left). Each column of plots forms a time-series of how the pitch behaves through a phrase, where time increases downwards. Successive rows are the first 100 ms of the phrase, the region from 300 ms to 400 ms, the 100 ms following the midpoint, and the last 100 ms of the phrase.

100 Hz. As early as 300 msec into the phrase, and persistent throughout the phrase, the f_0 range is characterized by a narrower band around 250 Hz. King may use pitch step-up to emphasize words, causing pitch to rise rather than to decline. This may account for the higher pitch range around the mid point region, compared to the earlier 300 msec region.

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 controls where the phrase curve should be relative to the speaker's pitch range. The intended f_0 value of the phrase curve at the time of the tag is calculated as $base + step\_to\ value \times range$, where *base* is the baseline and *range* represents the speaker's pitch range.

We use heuristic grammar rules to place the tags. Each utterance starts from the *base* value ($\Updownarrow_0$), steps up on the first stressed word, remains high till the end for continuation phrases, and steps down on the last word of the final phrase. At every pause, it returns to 20% of the pitch range above *base*, and steps 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 Fig. 12 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 13 displays the accent templates used to generate Fig. 12. King's choice of accents is largely predictable from the phrasal position: a rising accent in the

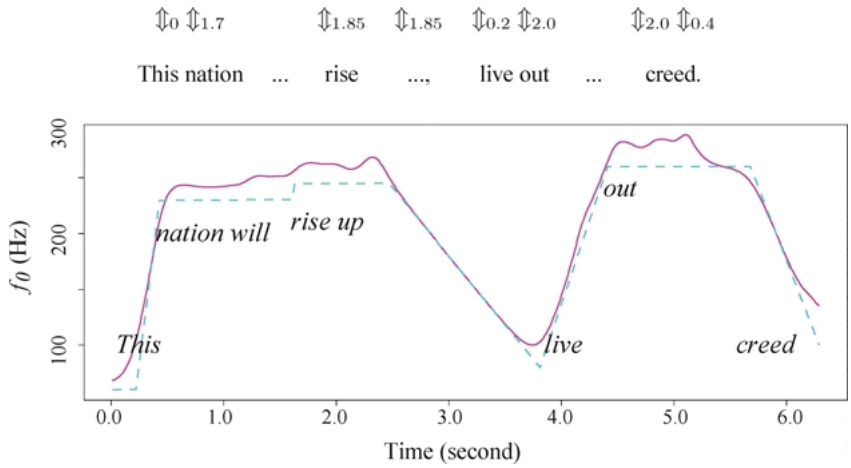

Figure 12. Generated phrase curve and pitch contour in the style of Martin Luther King.

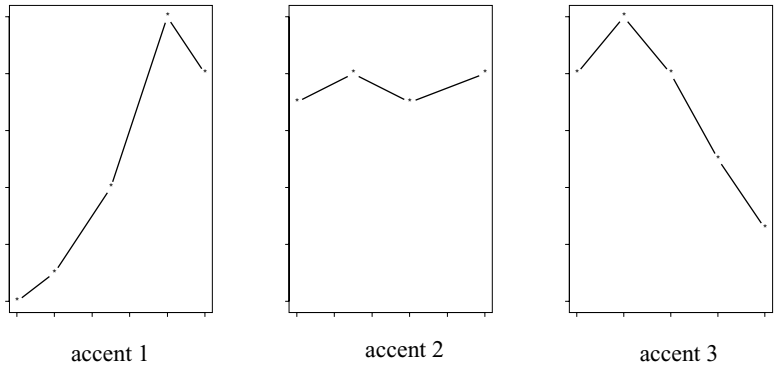

Figure 13. Accent templates for King's prosody.

beginning of a phrase, a falling accent on emphasized words and in the end of the phrase, and a third accent elsewhere.

Our emphasis is to explore portable prosodic features rather than copy synthesis, so in addition to reproducing the input f_0 curve, we also require that the features behave similarly as they are moved, that they produce physically possible f_0 curves, and more or less sound the same, no matter where they are placed.

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 adjustments for sentence length, without resorting to complex information such as sentence structure and the part of speech of words.

4.4. Speaking Styles—Local Scope

Speaker-dependent speaking styles may also be conveyed by idiosyncratic shapes for a given accent type. We examined the DARPA Communicator (NIST, 2000) travel reservation database, where subjects interact with a dialogue system trying to make flight reservations, and found 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 (Shih et al., 2001), which in ToBI notation would be annotated as having the tone sequence L*H-H%, a low accent followed by a high phrase accent and a high boundary tone. Different instances of the rising shapes by the same speaker are fairly consistent, but there are substantial differences between speakers.

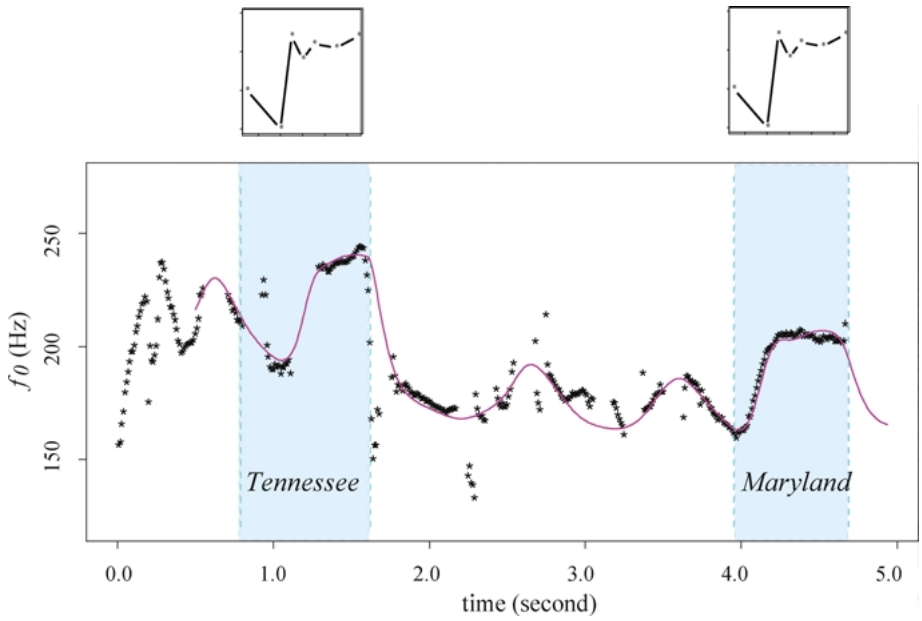

Figure 14. A sentence from Speaker 1 with two rising accents. "I live in Nashville, Tennessee and I'd like to go to Baltimore, Maryland."

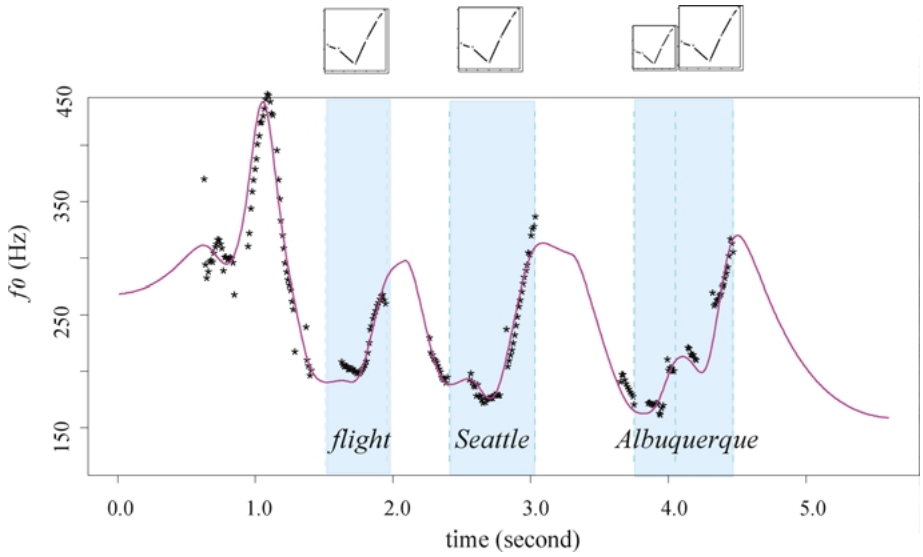

Figure 15. A sentence from Speaker 2 with multiple rising accents. "Um I would like a flight to Seattle from Albuquerque."

Speakers 1 and 2 in Figs. 14 and 15 convey different personal styles by using distinct rising contours. 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. 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 14 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 15 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 Fig. 15 are all generated from the same rising template shown above the figure.

This treatment opens up the possibility that the same annotation in intonational phonology, such as L*H-H%, may map to substantially different pitch contours, because different speakers have different habitual execution of the same linguistic functions. A related example is intonation patterns that are part of a foreign accent. A non-native speaker may have the same linguistic intent as the native speaker, but may simply implement an accent differently under the influence of their native language.

This is similar to the style/content distinction shown earlier in the paper, where the phonology plays a role in the content, and the style is the individual's implementation of the accent. Technically, this treatment is no different than the modeling of musical embellishment and the modeling of accent types that have different phonological status. We simply allow the user to define unrestricted shapes for accents in the modeling process.

5. Conclusion

In this paper, we have described prosodic features that are related to personal styles. We have shown examples of modeling embellishments and amplitude in music, as well as phrase curves and accent shapes in speech.

We can represent styles of speech or performance styles in music by a set of prosodic 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.

These examples suggest a common theme in terms of prosodic modeling: There are local effects such as accent shape and musical embellishment, and longer term effects such as phrase curve and musical notes. The accents and embellishments should be portable, so that they can be placed arbitrarily, but still produce a physically possible f_0 curve, and have similar perceptual results. This portability of the accents allows the heuristic rules that place them to be simple and more intuitive, because they then do not have to work around illegal combinations of accents.

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

All the examples discussed in this paper are available on the web at: http://prosodies.org/papers/2003/IJST/styles.wav or http://kochanski.org/gpk/papers/2003/IJST/styles.wav.

Acknowledgments

We thank Bob Damper and three anonymous reviewers for extensive comments, Danielle Fosler-Lussier for music terminology clarification, and Patrick Regan for listening to numerous versions of *Daisy*.

References

Abe, M. (1997). Speaking styles: Statistical analysis and synthesis by a text-to-speech system. In J.P.H. van Santen, R. Sproat, J. Olive, and J. Hirschberg (Eds.), *Progress in Speech Synthesis*, Springer-Verlag, pp. 495–510.

Anderson, M., Pierrehumbert, J., and Liberman, M. (1984). Synthesis by rule of English intonation patterns. *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, San Diego, CA, Vol. 1, pp. 2.8.1–2.8.4.

Beckman, M.E. and Ayers, G. (1997, March). Guidelines for ToBI labeling (version 3). http://www.ling.ohio-state.edu/phonetics/ToBI/ToBI.0.html.

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

Bloch, B. (1953). Linguistic structure and linguistic analysis. In A.A. Hill (Ed.), *Report of the Fourth Annual Round Table Meeting on Linguistics and Language Teaching*. Washington, DC: Georgetown University Press, pp. 40–44.

Cahn, J.E. (1998). *A Computational Memory and Processing Model for Prosody*. PhD Thesis, MIT, Cambridge, MA.

Cook, P. (1991). Identification of Control Parameters in an Articulatory Vocal Tract Model, with Applications to the Synthesis of Singing. PhD Thesis, Stanford University.

Dacre, H. (1892). *Daisy Belle, or A Bicycle Made for Two*. London: Francis, Day and Hunter.

Dorson, R.M. (1960). Oral styles of American folk narrators. In T.A. Sebeok (Ed.), *Style in Language*. Cambridge, MA: MIT Press, pp. 27–51.

ESPS/Waves. (2002). http://www.speech.kth.se/esps/esps.zip.

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

Fujisaki, H. (1983). Dynamic characteristics of voice fundamental frequency in speech and singing. In P.F. MacNeilage (Ed.), *The Production of Speech*. Springer-Verlag, pp. 39–55.

Fujisaki, H. (1988). A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour. In O. Fujimura (Ed.), *Vocal Fold Physiology: Voice Production, Mechanisms and Functions*. New York: Raven, pp. 347–355.

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 J. van Santen, R. Sproat, J. Olive, and J. Hirschberg (Eds.), *Progress in Speech Synthesis*. Springer-Verlag, pp. 417–428.

Hill, A.V. (1938). The heat of shortening and the dynamic constraints of muscle. *Proceedings of the Royal Society B 126*, 136–195.

Hirst, D.J., Di Cristo, A., and Espesser, R. (2000). Levels of representation and levels of analysis for the description of intonation systems. In M. Horne (Ed.), *Prosody: Theory and Experiment. Studies Presented to Gösta Bruce*. Dordrecht, The Netherlands: Kluwer Academic Publishers, pp. 51–87.

Huxley, A.F. (1957). Muscle structure and theories of contraction. *Progress in Biophysics and Biophysical Chemistry*, *7*:257–318.

Jilka, M., Möhler, G., and Dogil, G. (1999). Rules for the generation of ToBI-based American English intonation. *Speech Communications*, *28*:83–108.

King, M.L. (2000). *Martin Luther King, Jr.: We Shall Overcome*. Rolling Bay, Washinton: SpeechWorks, SoundWorks Entertainment, Inc., JRCD 7036.

Kochanski, G. and Shih, C. (2003). Prosody modeling with soft templates. *Speech Communication*, *39*(3/4):311–352.

Kochanski, G., Shih, C., and Jing, H. (2003). Hierarchical structure and word strength prediction of Mandarin prosody. *International Journal of Speech Technology*, *6*:33–43.

Kochanski, G.P. and Shih, C. (2000). Stem-ML: Language independent prosody description. *Proceedings of the International Conference on Spoken Language Processing*, Beijing, China, vol. 3, pp. 239–242.

Kubrick, S. (1968). *2001: A Space Odyssey*. Turner Entertainment Company. Based on the book of the same title by A.C. Clarke.

Liberman, M.Y. and Pierrehumbert, J.B. (1984). Intonational invariance under changes in pitch range and length. In M. Aronoff and R. Oehrle (Eds.), *Language Sound Structure*, Cambridge, MA: MIT Press, pp. 157–233.

Lindblom, B. (1963). Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America*, *35*:1773–1781.

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. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Munich, Germany, vol. 1, pp. 435–438.

Mathews, M.V. (1963). Bicycle built for two. *Music from Mathematics*. Decca Records. DL 9103.

Möhler, G. and Mayer, J. (2001). A discourse model for pitch-range control. *4th ISCA Workshop on Speech Synthesis*, Pitlochry, Scotland, pp. 11–15.

Monaghan, A.I.C. and Ladd, D.R. (1991). Manipulating synthetic intonation for speaker characterisation. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Toronto, Canada, pp. 453–456.

Murray, I.R. and Arnott, J.L. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *Journal of the Acoustical Society of America 93*:1097–1108.

NIST. (2000). DARPA communicator travel reservation corpus—June 2000 evaluation. Technical report, National Institute of Standards and Technology, Gaithersburg, MD. Speech Data published on CD-ROM.

Olive, J. (1998). The talking computer: Text to speech synthesis. In D.G. Stork (Ed.), *HAL's Legacy: 2001's Computer as Dream and Reality*, Cambridge, MA: MIT Press, Chap. 6, pp. 101–130.

Pierrehumbert, J. (1979). The perception of fundamental frequency declination. *Journal of the Acoustical Society of America*, *66*(2):363–369.

Schroder, M. (2001). Emotional speech synthesis—a review. *Proceedings of Eurospeech*, Aalborg, Denmark, pp. 561–564.

Shih, C. and Kochanski, G.P. (2000). Chinese tone modeling with Stem-ML. *Proceedings of the International Conference on Spoken Language Processing*, Beijing, China, vol. 2, pp. 67–70.

Shih, C. and Kochanski, G.P. (2001). Synthesis of prosodic styles. *4th ISCA Workshop on Speech Synthesis*, Perthshire, Scotland, pp. 229–234.

Shih, C., Kochanski, G.P., Fosler-Lussier, E., Chan, M., and Yuan, J.-H. (2001). Implications of prosody modeling for prosody recognition. In M. Bacchiani, J. Hirschberg, D. Litman, and M. Ostendorf (Eds.), *Proceedings of the ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding*. International Speech Communication Association. Red Bank, NJ, pp. 133–138.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., and Hirschberg, J. (1992). ToBI: A standard for labeling English prosody. *Proceedings of the International Conference on Spoken Language Processing*, Banff, Canada, vol. 2, pp. 867–870.

Strik, H. and Boves, L. (1992). Control of fundamental frequency, intensity and voice quality in speech. *Journal of Phonetics*, *20*(1):15–25.

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

Taylor, P.A. (2000). Analysis and synthesis of intonation using the tilt model. *Journal of the Acoustical Society of America*, *107*(3):1697–1714.

van Santen, J.P.H. and Möbius, B. (2000). A quantitative model of f_0 generation and alignment. In A. Botinis (Ed.), *Intonation: Analysis, Modelling and Technology*. Dordrecht, The Netherlands: Kluwer Academic Publishers, pp. 269–288.