Statistical Prosodic Modeling: From Corpus Design to Parameter Estimation

Jerome R. Bellegarda, *Senior Member, IEEE*, Kim E. A. Silverman, Kevin Lenzo, and Victoria Anderson

***Abstract*—The increasing availability of carefully designed and collected speech corpora opens up new possibilities for the statistical estimation of formal multivariate prosodic models. At Apple Computer, statistical prosodic modeling exploits the Victoria corpus, recently created to broadly support ongoing speech synthesis research and development. This corpus is composed of five constituent parts, each designed to cover a specific aspect of speech synthesis: polyphones, prosodic contexts, reiterant speech, function word sequences, and continuous speech. This paper focuses on the use of the Victoria corpus in the statistical estimation of duration and pitch models for Apple's next-generation text-to-speech system in Macintosh OS X. Duration modeling relies primarily on the subcorpus of prosodic contexts, which is instrumental to uncover empirical evidence in favor of a piecewise linear transformation in the well-known sums-of-products approach. Pitch modeling relies primarily on the subcorpus of reiterant speech, which makes possible the optimization of superpositional pitch models with more accurate underlying smooth contours. Experimental results illustrate the improved prosodic representation resulting from these new duration and pitch models.**

***Index Terms*—Intonation modeling, prosodic representation, prosody generation, speech database design and collection, text-to-speech systems.**

I. Introduction

In RECENT years, text-to-speech (TTS) systems have come to rely more and more on data-driven, statistical modeling. One reason has been the emergence of concatenative synthesis, which implies the existence of an automatic procedure to properly select candidate units from a recorded speech database. Another factor has been the steady shift from handwritten, hand-tuned pitch and duration rules to formal multivariate prosodic models, which requires the associated model parameters to be statistically derived from a training corpus. This has sparked interest in large scale, systematic data collection, of the kind carried out over the past decade in the field of speech recognition (see, e.g., [16], [23]). Whereas a limited number of systematically elicited exemplars might have been sufficient to, say, write a particular pitch or duration rule, many more observations are usually required to automatically uncover (and/or validate) the equivalent rule from a database. The more complex the underlying model, the more attention should be paid to the scale requirements of the collection.

In prosodic modeling, for instance, the number of model parameters is necessarily quite large. This is especially true when the prosodic phonological markup is used to control more than the traditional pitch and duration, such as the characteristics of the glottal excitation function and overall spectral slope. Other aspects of prosody, such as differences between citation forms and connected speech, issues associated with variable speaking rate, or the problem of paragraph-length prosody, entail the estimation of even more parameters.

The reliability of such estimation depends critically not just on the quantity, but also on the coverage and consistency of the data available. Thus, the design principles and collection procedures underlying the training corpus have a direct impact on the quality, communicative effectiveness, and naturalness of synthetic speech. Beyond the minimal requirement that statistical models be estimated on "enough data," there is no standard approach to corpus design and collection. Corpora vary from systematically generated nonsense words, through lists of discrete unrelated sentences, to news broadcasts. Some contain a single speaker, others span multiple speakers. Those corpora based on a single speaker are often recorded in disparate sessions over an extended period of time, with little consistency in recording conditions or speaking style. And there is typically little effort to systematically control the intonation. This variety can be traced to divergent goals in data collection. Is the purpose just to provide speech synthesis units? Or to also provide data for studies in methods of signal representation? Should the same speaker be used to construct both acoustic and prosodic models?

An effort to formally explore this process was recently undertaken at Apple Computer, with the aim to support the multiple facets of ongoing speech synthesis research and development. This entails different kinds of material and recording conditions than typically used to provide 100% coverage of small synthesis units, such as diphones or demi-syllables. The outcome of this broader outlook was a rich corpus of very large size, informally referred to as the *Victoria corpus*. This corpus is composed of five major subcorpora, each designed to cover a specific aspect of speech synthesis: polyphones, prosodic contexts, reiterant speech, function word sequences, and continuous speech. It was spoken in general U.S. English by one linguistically trained adult female. This corpus was instrumental in the statistical estimation of more accurate duration and pitch models for Apple's next-generation text-to-speech system in Macintosh OS X.

Manuscript received June 8, 2000; revised September 11, 2000. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Nick Campbell.

J. R. Bellegarda and K. E. A. Silverman are with the Spoken Language Group, Apple Computer, Inc., Cupertino, CA 95014 USA (e-mail: jerome@apple.com).

K. Lenzo was with Apple Computer, Inc., Cupertino, CA 95014 USA. He is now with Carnegie Mellon University, Pittsburgh, PA 15213 USA.

V. Anderson was with Apple Computer, Inc., Cupertino, CA 95014 USA. She is now with Fonix Corporation, Salt Lake City, UT 84111 USA.

Publisher Item Identifier S 1063-6676(01)00332-7.

Accordingly, this paper has a dual goal. First, it describes the design and collection of the Victoria corpus, and second, it examines how this corpus has been exploited so far for training new duration and pitch models. The paper is organized as follows. In Section II, we give an overview of the corpus and its various constituent parts. Section III identifies some of the important procedural issues to be addressed in collecting this kind of corpus. Section IV describes how this data was used for statistical duration modeling. Section V is the counterpart of Section IV for pitch modeling. Finally, Section VI reports on a series of experiments illustrating the resulting improvements in prosodic representation.

II. Victoria Corpus

The Victoria corpus is intended to

1) support research into high-quality signal representation;
2) provide source units for concatenative synthesis;
3) be of sufficient quality to support detailed pitch extraction, pitch epoch detection, and inverse filtering for estimation of glottal source parameters.

Following is an overview of each of its five constituent parts.

A. Polyphones

This segment of the corpus was designed to provide various types of units for concatenative synthesis, such as common syllables or polysyllabic strings, words, common word stems, and/or common inflectional morphemes. It uses the most common 28 000 words of U.S. English, as estimated from several text corpora available from the LDC [19]. These words were arranged into pairs, which were in turn placed in sets of two or three (separated by commas) to form "sentences," using a rich sampling of phoneme concatenations as the phonetic criteria for word groups. Half of these sentences ended with a period, and half with a question mark. A representative example is

rife undertaking, anyhow fanatic, grab disruptive? (1)

By design, each word was spoken with four distinct intonational patterns[1] (cf. [30])

1) $\mathbf{H}^*\,\mathbf{L}\,\mathbf{L}\%$, a sentence-final falling nuclear accent, as if followed by a period;
2) $\mathbf{H}^*\,\mathbf{L}\,\mathbf{H}\%$, a nuclear falling–rising pattern that typically occurs before a sentence-internal punctuation such as comma, colon, and quotation mark;
3) $\mathbf{H}^*\,\mathbf{H}\,\mathbf{H}\%$, a nuclear high-rise that is common in unmarked yes-no questions;
4) $\mathbf{H}^*$, a prenuclear high pitch accent.

The end result is that each of the 28 000 words was produced in four utterance positions

1) utterance-initial;
2) phrase-initial but utterance-medial;
3) phrase-final but utterance-medial;
4) utterance-final.

These by no means exhaustively represent the rich variation used in normal everyday conversation, but they cover the most common and perceptually-salient subset in the informative factual discourse to which synthetic speech is often applied.

The speaker for this corpus was linguistically trained, familiar with the transcription system and the intended intonation contours. Hence, although the texts were largely nonsensical, the intonation was produced consistently and in a natural way. Arguably, these are "citation form" contours, which only occur in real conversation at the most communicatively important places. One of the characteristics of normal human speech is that less information-carrying stretches of text are often under-articulated: we believe that this may also be true of the intonational contrasts during such passages. In order to model this under-articulation, however, it makes sense to begin with a parameterized model of the citation cases, and then investigate appropriate adjustments to the parameters such as fundamental frequency range and accent height during reduced articulation. Other sections of the Victoria corpus (described below) intentionally contain appropriate data to support such investigations.

One particular aspect of prosodic modeling which can be studied with the *Polyphones* subcorpus is the potential interaction between tonal structure and durational structure. A common approach to prosodic synthesis is to generate duration from the phrasal position, stress structure, and accent locations, and then to subsequently generate the associated pitch contour from the accent tones. However, this ordering assumes that the durational structure depends only on accent location, not on accent type. For instance, this approach would generate the same duration for an accented phrase-final word whether it contained a fall, a rise, or a fall–rise. Examples in [27] suggest that this is an inaccurate oversimplification. The *Polyphones* subcorpus is designed to allow research into whether such tonal independence really exists.

B. Prosodic Contexts

This segment of the corpus was designed to systematically cover all syllable shapes with rich phonetic variation. It complements the above segment, in that it comprises the most common prosodic boundaries in English which are not represented at all in the *Polyphones* subcorpus. Here we systematically vary the distance between pitch accents, and between accents and the next prosodic boundary.

Specifically, the *Prosodic Contexts* subcorpus focuses on

1) how pitch and duration of accented syllables vary with the distance to the next rightmost prosodic event [28];
2) utterance-internal prosodic boundaries [29];
3) final lengthening;

[1]The intonational notation used throughout this work comes from the tone and break indices (ToBI) transcription system [30]. Briefly, this system views intonation as a linear sequence of relatively local targets or gestures that are sparsely distributed over a text. These events are composed of just two primitive tones: high (**H**) and low (**L**). These are all linguistic choices by the speaker, in order to fulfill discourse roles and convey information not derivable from the words and syntax alone. Pitch targets or gestures that are associated with particular words, to give them a particular prominence or relationship with surrounding topics, are known as *pitch accents*. Pitch accents have a star in their transcription (e.g., $\mathbf{H}^*$, $\mathbf{L}+\mathbf{H}^*$), and are associated with a particular syllable of the accented word. Major phrase and utterances boundaries are marked by *boundary tones*, transcribed with $\%$ (e.g., $\mathbf{L}\%$) and attached to the rightmost edge of the phrase. Finally, the region between the last pitch accent in a phrase and its rightmost edge is governed by a *phrase tone* (**H** or **L**).

4) alignment of pitch and associated segmental structure in different syllable types.

Each utterance consists of two accented words, preceded by several unstressed function words. The words were chosen to systematically vary the number of syllables between the two accented syllables between 0 and 5, and to systematically vary the position of the word boundary. Thus, for example

$$as\ a\ \mathbf{frill}\ \mathbf{cheap}ly \tag{2}$$

has the two accents (marked in boldface on "*frill*" and "*cheap*") adjacent

$$as\ a\ \mathbf{swamp}\ custo\mathbf{mar}ily \tag{3}$$

has two unaccented syllables between the accents (on "*swamp*" and "*-mar-*") with the word boundary adjacent to the left, and

$$the\ te\mathbf{mer}ity\ \mathbf{throne} \tag{4}$$

also has two unaccented syllables between the accents (on "*-mer-*" and "*throne*") but this time with the word boundary adjacent to the right.

The design decision was made to only use real words (from the PRONLEX dictionary; cf. [19]), rather than to construct nonsense words to achieve the desired items. This was because we could not be confident that phrases of nonsense words would be spoken sufficiently naturally to represent normal English prosody. A grammar was constructed of English syllables, with phonemes collapsed into major classes. Words were chosen by a greedy algorithm to ensure that every possible syllable type occurred in each of the two relevant accent positions, and within each syllable type there was systematic variation of the instances of each of the phonemes in each of the classes. Thus, for example, in the syllable class "voiceless fricative, sonorant, high vowel, sonorant," if the word "*frill*" were used for one item, then the next time that syllable class was required a different word such as "*swill*" would be used instead.

The list generation procedure was completely automated to conform to these interacting sets of constraints, so that each generated list of phrases depended only on an initial condition (the first word chosen). Obviously, not all possibilities could be filled: sometimes there was no word available for a particular set of constraints. By using different initial conditions, different instances of the list of phrases were obtained. We recorded two different such lists. One contained 733 phrases, the other 631. Together these yielded 50 797 phonemes.

Each utterance was spoken in two ways: 1) a $\mathbf{H}^*$ pitch accent on both words (usually the accent on the second word was downstepped or reflected final lowering) and 2) a $\mathbf{H}^*$ on each word, but with an intervening $\mathbf{L}$ tone associated with an utterance-internal intermediate phrase boundary.

To also support studies in speaking rate variation, we have recorded one pass through this subcorpus at the speaker's fastest possible speaking rate. Note that we did not attempt to repeat this exercise at the speaker's slowest possible (!) speaking rate. This was for two reasons: 1) there is little use for slow speech in speech synthesis applications and 2) we believe that most instances of slow speech, where speakers and listeners perceive that the speaking rate has slowed down to increase clarity, actually consists of more frequent and more major prosodic boundaries.

C. Reiterant Speech

This segment of the corpus was designed to support detailed modeling of pitch contours for the intonations in the above two segments, but uncontaminated by segmental perturbations. Reiterant speech [18] is commonly used to elicit smooth and natural-sounding fundamental frequency contours, in order to understand and model the relationship between fundamental frequency, which is an acoustically defined parameter, and intonation, which is a linguistic abstraction and a perceptual phenomenon (see, e.g., [11]). It is based on the observation that fundamental frequency contours typically show steep fluctuations and discontinuities in the regions of obstruent consonants such stops and fricatives, whereas they are much smoother and continuous in vowels and sonorant consonants such as nasals.

Semantically-coherent utterances were constructed to cover a superset of the intonational melodies in the above subcorpora, on a subset of the syllable structures. Specifically, the three tunes from the *Polyphones* subcorpus ($\mathbf{H}^*\,\mathbf{LL}\%$, $\mathbf{H}^*\,\mathbf{L}\,\mathbf{H}\%$, and $\mathbf{H}^*\,\mathbf{HH}\%$), with the accent on the first of the two content words in each phrase, were intended to model the way these common and communicately important tunes vary when stretched over texts of different length; the two tunes from the *Prosodic Contexts* subcorpus ($\mathbf{H}^*\,\mathbf{H}^*\,\mathbf{L}\,\mathbf{L}\%$ and $\mathbf{H}^*\,\mathbf{L}\,\mathbf{H}^*\,\mathbf{L}\,\mathbf{L}\%$) were intended to model the influence of an utterance-internal phrase boundary (marked with a low phrase accent) on the preceding and following accents; and the additional tune $\mathbf{H}+\mathbf{L}^*\,\mathbf{H}^*\,\mathbf{L}\,\mathbf{L}\%$ was intended to model the $\mathbf{L}^*$ accent in a variety of syllabic structures and polysyllabic contexts.

Each utterance was spoken with the intended intonation, then immediately followed by a version where every open syllable was replaced with "*ma*" and every closed syllable replaced with "*mom*" (cf. [18]). To illustrate, Fig. 1 shows the pitch contour resulting from the recording of

$$but\ it's a\ \mathbf{gra}nular\ concate\mathbf{na}tion \tag{5}$$

while Fig. 2 corresponds to the same pitch contour for the associated recording

$$mom\ mom\ ma\ \mathbf{ma}\ ma\ mom\ mom\ ma\ ma\ \mathbf{ma}\ mom \tag{6}$$

in which the only consonant is /m/. The contour of Fig. 2 is clearly smoother, in part because 1) all the unvoiced gaps have been eliminated and 2) the choice of the continuant /m/ generates comparatively small segmental perturbations. Note that it is much easier to see the (relatively invariant) local shape of a $\mathbf{H}^*$ pitch accent on Fig. 2 (from 1.65 s to the peak). Also much clearer is the interpolation from the preceding accent target up to the beginning of the $\mathbf{H}^*$ laryngeal gesture (from 0.96 to 1.65 s on Fig. 2).

Pitch modeling typically has a dual aim: 1) to fit a parametric model to the shape of the fundamental frequency contours with minimal contamination by segmental effects and 2) to model the alignment of the contours with the segmental

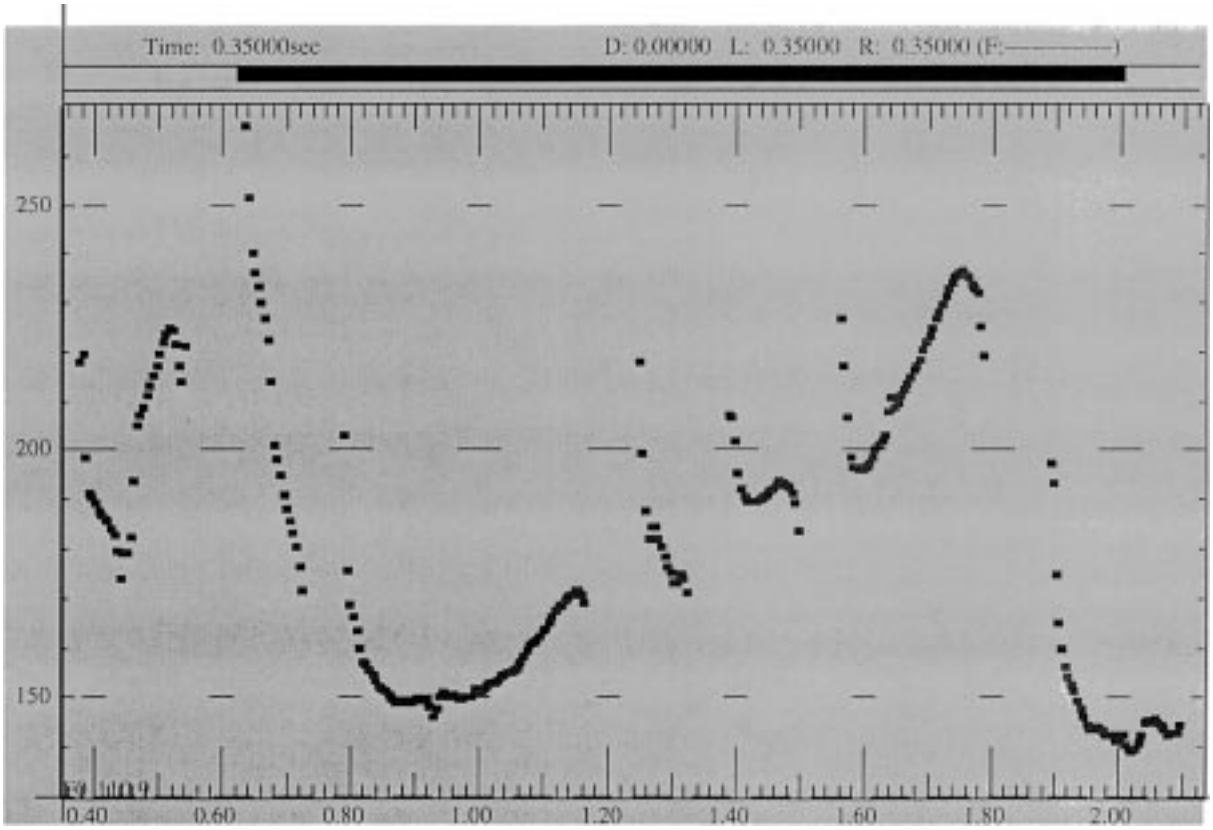


Fig. 1. Original pitch contour.

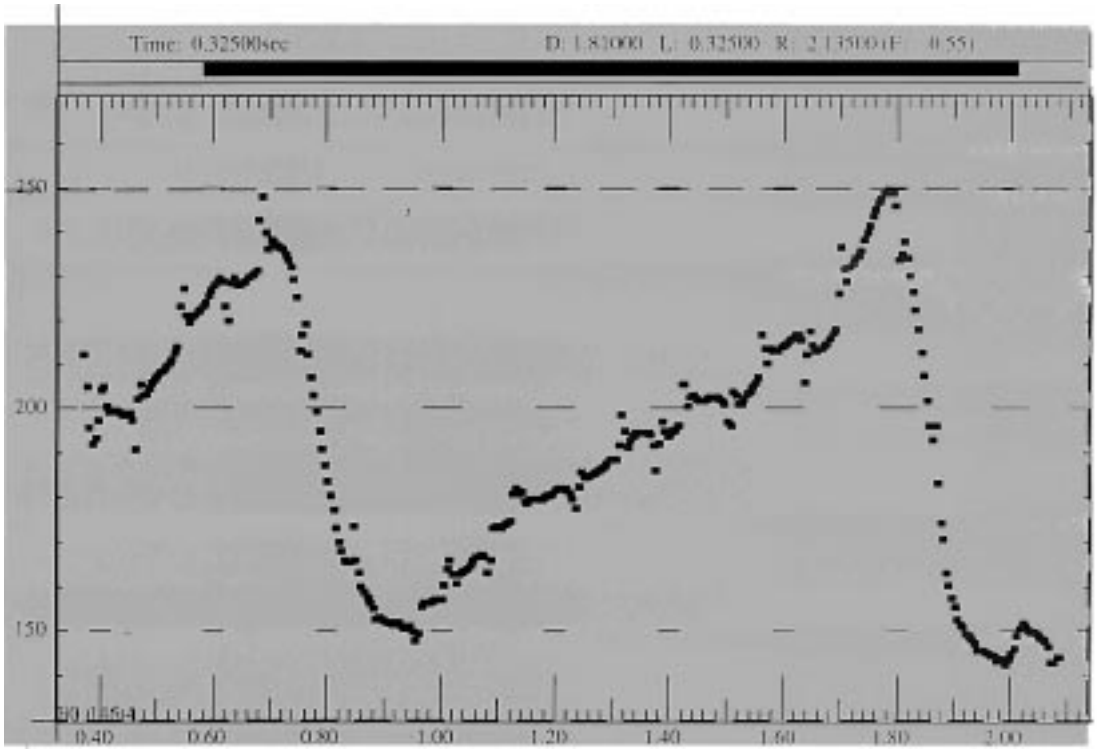


Fig. 2. Pitch contour of reiterant speech.

structure. Silverman [27] found that when stretches of reiterant speech were embedded in frame sentences spoken with neutral declarative intonation, the reiterant speech exhibited normal intrinsic vowel and consonant durations, and showed the same well-known durational interactions with adjacent phonemes as for normal speech. In addition, Steele (reported in [28]) and

others (e.g., [9], [21]) found stress-related lengthening and phrase-final lengthening on reiterant open and closed syllables. On the basis of such phonetic research, we believe that the reiterant speech collected in this subcorpus (i.e., produced by a phonetically trained speaker, with an intended intonation contour, spoken immediately after a real sentence with the same syllable structure and intonation) is sufficiently realistic to support a first pass of a model of segmental/suprasegmental alignment (cf. Section V). For example, we can model the locations of accent-related peaks, relative to the start of the vowel in the associated syllable, as a function of the distance in syllables to the next intonational phrase boundary.

Arguably, a model of the alignment of intonational events with the phonetic segments that carry those events requires natural speech and concurrent modeling of the underlying smooth contours and the segmental effects. The reiterant speech alone is insufficient for such a model: for example we suspect that word boundary locations, ambisyllabicity, and distinctions between intervocalic singletons and geminates, may be less well represented in reiterant speech than in natural speech. The Victoria corpus contains other material suitable to perform this concurrent modeling (see Section II-E), once quantitative models of the component influences on fundamental frequency contours have been developed. The advantage of using reiterant speech is that the smooth contours illustrated in Fig. 2 enable accurate fitting of parameterized pitch models using gradient descent and multiple regression. The method and results obtained so far will be discussed in Section V, along with some implications for prosodic phonology.

D. Function Word Sequences

This segment of the corpus was designed to cover frequent sequences of unstressed function words (such as "*and he has*," "*in that*," and "*that we have*") and common clitic groups (such as "*couldn't've*"). These heavily co-articulated sequences are notoriously difficult to synthesize, extremely common in connected speech, yet totally lacking in the above subcorpora. The list of function word sequences was automatically derived from the *Wall Street Journal* corpus [16].

Each unstressed sequence was spoken between two accented content words. A representative example is

$$\textbf{clone}\ \textit{those in the}\ \textbf{fund} \tag{7}$$

for the unstressed sequence "*those in the*." The content words were chosen such that the consonants adjacent to the function words were 1) easily segmentable phonetically and 2) systematically varied in their place and manner of articulation. This was in order to provide polysyllabic concatenative units which would be fit in a variety of articulatory contexts. In addition, this subcorpus was also intended to give an extra set of contexts for duration modeling.

E. Continuous Speech

In all prior segments, utterances are in citation form, which runs the danger of producing over-articulated synthetic speech. Besides, citation forms tend to comprise largely unconnected short sentences. This is inherently inadequate to model larger-span effects, such as prosodic behavior in very long sentences, paragraph-length prosody [27], and the variations in speaking rate for communicative connected speech.

This segment of the corpus was designed to address this issue. Both read and spontaneous speech were captured. The read speech segment comprises short stories chosen by the speaker for their literary style being easy to read out loud. The speaker familiarized herself with the content before reading them, in order to minimize speech errors and to produce prosody and articulation appropriate to the content. The spontaneous speech segment was produced by having the speaker describe some properties of images. Two examples include 1) looking at a map and describing directions to travel between certain points and 2) looking at an grid of the "faces" method of multivariate data display and describing inferable data patterns.

III. Collection Procedures

Speech synthesis corpora demand much higher signal quality than typically acceptable for other speech technologies. This requires minimization of background noise, phase distortion, and spectral distortion. Dual recording of acoustic and glottal signals is highly desirable for accurate pitch extraction and identification of glottal closure. In addition, speaking style and consistency are two difficult and related issues which need to be addressed.

A. Noise Compensation

Standard professional recording procedures, such as high-end audio equipment and a structurally isolated double-walled acoustic studio, were found to be insufficient to control the background noise. We iteratively traced the sources of this noise by spectrally analyzing it, identifying the most prominent partials and resonances, and then isolating the relevant causes in the recording setup and circuitry. We found, for example, that placing the pre-amplifier in the recording booth, and running it off batteries instead of grid power, reduced the 60 Hz harmonics and decreased the susceptibility to radio-frequency interference from computers in a nearby lab.

One persistent source of noise was the wideband hiss generated internally within the microphone itself. This was found to be true for a wide range of professional broadcast-quality microphones. We successfully reduced it by 2–3 dB by simultaneously recording from two such microphones next to each other and adding their signals, thereby neutralizing the two associated (uncorrelated) background hisses. The two microphones were suspended in rubber shock absorbers to reduce vibrations from the floor, at a constant distance (12 to 13 cm) from the speaker's mouth.

Another source of noise in digital recordings is nonlinearities in the low-order bits of a digital-to-analog converter. To address that, we recorded direct-to-disk at 20 bits/sample and subsequently rounded to 16 bits. The signal-to-noise ratio (computed spectral signal peak to spectral noise peak) was in the range 52–55 dB for all measured recordings.

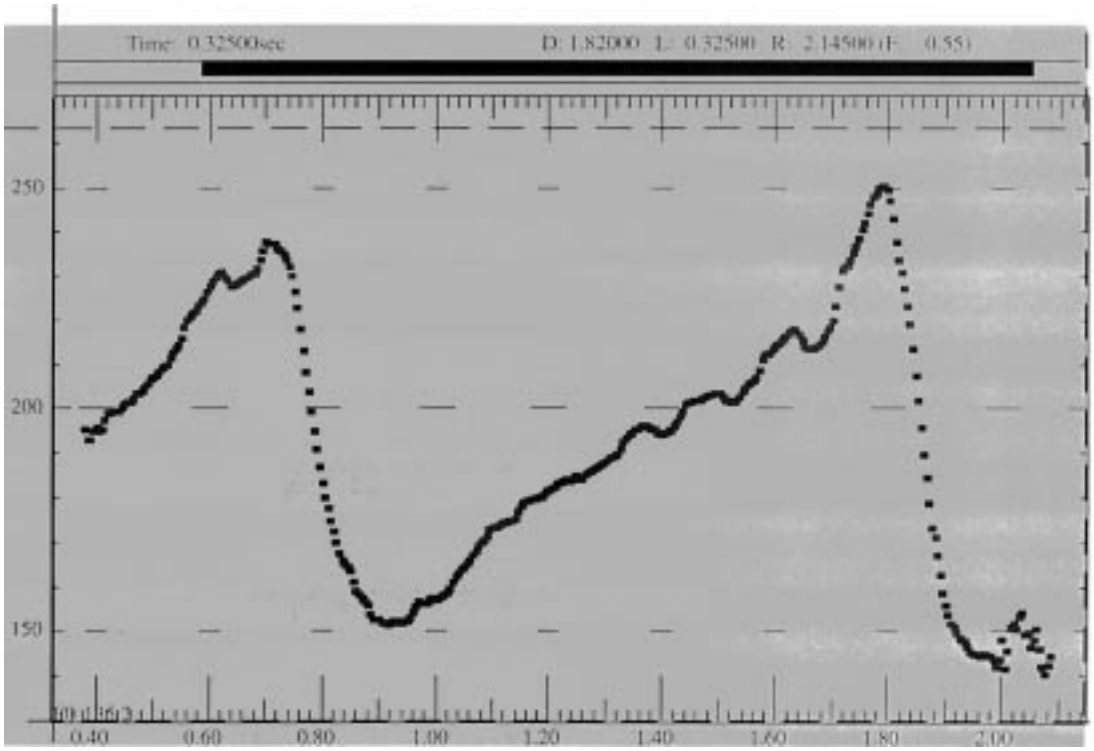


Fig. 3. Pitch contour of simultaneous EGG signal.

B. Dual Recording

An electroglottograph (EGG) signal was recorded simultaneously with the acoustic signal. The capture of dual acoustic and EGG channels is extremely useful for extracting very smooth and reliable fundamental frequency contours, and provides an extra source of information to help determine the glottal closure.

To illustrate, Fig. 3 shows the pitch contour associated with the example of Figs. 1 and 2, but this time extracted from the corresponding EGG signal, i.e., measured near the excitation of the vocal tract. A comparison of Figs. 2 and 3 shows that even reiterant speech using the consonant /m/ still exhibits a consistent pattern of segmental perturbations. In Fig. 2, fundamental frequency appears to be raised by 5–10 Hz for about 1 to 2 glottal periods at the release of each /m/ closure into the next vowel, and depressed by about the same amount at the transition from the preceding vowel into each closure of the /m/. These consistent transient effects in the acoustic signal are not present in the excitation signal, as measured by the EGG. We hypothesize that these are therefore caused by a change in the length of the signal path through the vocal tract.

During an open vowel the majority of the acoustical energy is transmitted through the mouth. When the oral cavity is occluded by the lips closing for /m/, the acoustical energy is necessarily transmitted through the nasal cavity, entailing a slightly longer path. So at the point in time when the lips close there is a Doppler-like effect: the onset of the next glottal period is delayed because of a longer travel time through the nose, which means that the glottal epoch spanning the time of lip closure appears longer, and therefore of lower pitch. Similarly, at the release of /m/ into the next vowel, the glottal period appears shortened and therefore of higher pitch. The size of the change in fundamental frequency (5–10 Hz) indeed corresponds to what we would predict by calculating the delay from the length of the oral and nasal tracts, based on the speed of sound.

Further evidence in favor of this hypothesis is that the magnitude of the effect seems to depend on the speed of consonant closure and release gestures. We notice, for example, that a tautosyllabic post-vocalic /n/ or /ŋ/, as in "*sin*" or "*sing*," typically lacks any apparent transient depressed pitch. These nasals tend to have slow closing gestures, so the gradual change in vocal tract length is spread out across much of the preceding vowel.

This remark notwithstanding, note that the EGG signal is not completely free of segmental perturbations. Fig. 3 does show a (nontransient) tendency for the fundamental frequency to be depressed in the glottal signal throughout the closure of /m/ (cf. for example, from 0.62 to 0.69 s, and 1.63 to 1.71 s). The presence of this segmental perturbation in the EGG suggests that it is at least partly due to the change in vocal tract impedance (looking forward from the glottis). During closure there is some back pressure which decreases the trans-glottal pressure gradient and therefore causes a decrease in the rate of glottal oscillation.

All EGG signals were measured using a single-electrode EGG device. In retrospect, this was adequate but not completely successful. Subsequent analysis of the recordings showed that occasionally the speaker's larynx would move above or below the electrodes, and this would cause the signal to disappear momentarily. This occurred most often during the pitch peaks of nuclear accented syllables. We were able to minimize these dropouts by providing an oscilloscope display of the EGG signal in the speaker's line of sight. The speaker was thereby able to monitor the signal for potential dropouts, often adjusting the electrodes and re-recording as necessary. Nevertheless this

increased the cognitive load on the speaker and the overall duration of the collection. For this type of recordings, we recommend using multiple electrodes, or investigating sonar or radar techniques to capture laryngeal activity.

C. Speaking Style

Previous corpora have generally either left the speaking style up to the speaker, or have requested a professional news-reading style. Although the latter does tend to reduce variation in the signal amplitude, it implies a sustained level of vocal effort across all syllables of all words, which is not typical in normal conversation. (In regular day-to-day communicative speech, this degree of laryngeal tension and subglottal pressure is reserved for just the most salient syllables of the few most important words.) Since the spectral correlates have more high-frequency energy during voiced speech, we have found that the news-reading style results in a perceptually unpleasant, somewhat strident voice quality.

After exploring a few different styles we adopted a more relaxed, slightly more breathy speaking style. This style is used in the acting community to produce an impression of more intimacy. It is usually produced very close to the microphone, and is used, for example, when an actor is "talking to himself" or narrating his internal thoughts for the audience's benefit. We have found this to produce a softer, more pleasant voice quality in synthetic speech. Note that this relaxed laryngeal mode produces a larger open quotient in the larynx, and hence increases 1) the proportion of subglottal coupling and 2) the proportion of nonharmonic noise (breathiness). Consequently the signal is less well-modeled by such frequency-domain representations as formant analysis or linear prediction. We use a time-domain signal representation, which is more robust to the corresponding assumption violations.

D. Consistency

A common problem in concatenative synthesis is that the units do not have consistent glottal slope, vocal effort, or perceived intensity. This produces discontinuities at the concatenation points which are not evident in the formant structure *per se*. It is difficult but crucially important to ensure that the speaker maintains a consistent speaking style, articulation rate, and vocal effort across the whole corpus. For reference, each session started with playing example recordings of the "target" voice quality, intonation, vocal effort, and pitch range. These examples were available all the time and often referred to by the speaker. In addition, for the initial recording sessions, an independent listener also performed close, live monitoring of the intended production, correcting the speaker as necessary.

IV. Duration Modeling

In natural speech, durations of phonetic segments strongly depend on contextual factors such as the identities of surrounding segments, stress, accent, and phrase boundaries (cf., e.g., [35]). For synthetic speech to sound natural, these duration patterns must be closely reproduced. Among the various methods that have been proposed for duration prediction, a "sums-of-products" (SoP) approach has a number of useful advantages [20].

A. SoP Modeling

Assume there exist N contextual factors influencing duration, and denote by $\mathcal{S}_i$ the scale vector quantifying the duration effects associated with the ith factor $\mathcal{F}_i$, $1 \leq i \leq N$. For example, if $\mathcal{F}_1$ corresponds to the stress factor, it might comprise two levels, "*stressed*" and "*unstressed*." In the simplest case, the effect of this factor on duration can be captured by a vector comprising two elements, say $\mathcal{S}_1 = [S_{1,\,stressed}\ S_{1,\,unstressed}]$, with appropriate values of $S_{1,\,stressed} = \mathcal{S}_1(stressed)$ and $S_{1,\,unstressed} = \mathcal{S}_1(unstressed)$ estimated from the data. Now let a given phonetic segment be characterized by an input vector $[f_1\ f_2 \cdots f_N]$, where each f_i represents the observed level of the associated factor $\mathcal{F}_i$, $1 \leq i \leq N$. Generically, the duration D of this phonetic segment can be described as

$$F(D(f_1, f_2, \cdots, f_N)) = G(\mathcal{S}_1(f_1), \mathcal{S}_2(f_2), \cdots, \mathcal{S}_N(f_N)) \tag{8}$$

where $F(\cdot)$ and $G(\cdot)$ could be, in principle, two arbitrarily complex functions of the various durations involved.

The SoP model assumes that F is a monotonically increasing transformation, and that G can be decomposed as a sum of products of single factor parameters. This amounts to postulating the following form for (8)

$$F(D(f_1, f_2, \cdots, f_N)) = \sum_{i \in K} \prod_{j \in I_i} S_{i,j}(f_j) \tag{9}$$

where

- F *unknown* but strictly increasing;
- K some collection of indices associated with subsets of the set of factors;
- I_i collection of indices of factors occurring in the ith subset and the scale function;
- $S_{i,j}$ simply a mapping from discrete to numerical values.

In this expression, choosing $K = \{1, 2, \cdots, N\}$, $I_i = \{i\}$, and $F(x) = x$ leads to various derivatives of the additive model originally proposed by Klatt [14]. Alternatively, choosing $K = \{1\}$, $I_1 = \{1, 2, \cdots, N\}$, and $F(x) = \log(x)$ leads to multiplicative models such as described by van Santen [35]. The evidence appears to indicate that the latter perform better than the former. Two reasons why this might be the case are 1) the distributions tend to be less skewed after the log transformation and 2) the fractional approach underlying multiplicative models is better suited for more extreme durations. Thus, the latter set up is normally used. There is, however, no evidence that the log transformation is optimal. Rather than eliminating skewness in the data, it tends to merely reduce (and reverse) it. And while it is true that contexts such as phrase-final position are likely to lengthen long phonemes more than short phonemes, there is no *a priori* reason for all factors to be strictly multiplicative across all durations. We will revisit this point shortly.

Note that the right-hand side of (9) is linear in some appropriate parameterization(s) of the scales $S_{i,j}$. In practice, SoP methods are therefore closely related to multiple linear regression analysis in either linear or log domain, depending on the transformation F selected [34].

B. Theoretical Observations

The origin of the SoP approach can be traced to the "axiomatic measurement" theorem [15], as applied to duration data. Briefly, for (9) to hold, it is *necessary* that the duration function $D(f_1, f_2, \cdots, f_N)$ be *generalized additive*. This is the case if D has a decomposable structure, i.e., if all components $f_1, f_2, \cdots, f_N$ "contribute their effects independently to the attribute in question" [15]. Furthermore, for ordinal regularity to hold, D must exhibit *monotone decomposability*, meaning not only the function F but also each per-component mapping is monotonically increasing. Strictly speaking, there are therefore two conditions that must be satisfied for the SoP description to apply to duration modeling: joint independence of the variables, and monotonicity of the transformation and scale functions.

Diagnosis of an N-variable function on the basis of joint independence is a matter of testing each $(N-1)$-tuple of variables for independence of the Nth. This is generally a complex process: to illustrate, for a simple polynomial function with $N=3$, it requires following all the steps mentioned in the flowchart in [15, p. 345]. In the case of duration data, such diagnosis is not going to be successful, since joint independence clearly will not hold for *all subsets* of $2, 3, \cdots, N$ factors. For example, accent and phrasal position interact in a complex way in their influence on vowel duration, i.e., these factors do not contribute their effects independently. More generally, any form of simultaneously additive and multiplicative interaction violates joint independence, and thus decomposability.

It has been argued that most interactions follow the principle of directional invariance [32], meaning that their effects are amplificatory, rather than reversed or otherwise permuted [35]. This smooth general behavior often offers a justification for applying (9) anyway. As pointed out in [35], the "regular patterns of amplificatory interactions" make it "quite plausible that *some* sums-of-products model will fit the [appropriately transformed] durations" (emphasis ours). Reversal interactions do exist, however. For example, "when we compare words such as *butter*, *return*, *finer*, and *beneath*, we find that the /t/ burst in *return* is longer than the /n/ in *beneath*, while the opposite holds contrasting the /t/ burst in *butter* with the /n/ in *finer* …. This is a reversal of the effects of the segmental identity factor brought about by a change in the stress levels of the surrounding vowels" [32]. In such situations, it would appear that $D(f_1, f_2, \cdots, f_N)$ is not generalized additive, and the choice of the usual log function for F is probably not optimal.

In fact, given the joint independence violation, the monotonicity condition on the transformation F may no longer apply at all. Losing this theoretical guideline may substantially complicate the search for a suitable F. As the optimal F may no longer be strictly increasing, this opens up the possibility of inflection regions, or even discontinuities. Transformations other than the log function, in particular, may result in better models, as we first showed in [4], and further explored in [31]. Continuing the same line of investigation, we will argue in favor of a piecewise linear formalism, which appears to be robust against all interactions, amplificatory or otherwise. This reasoning is grounded in the following empirical observations.

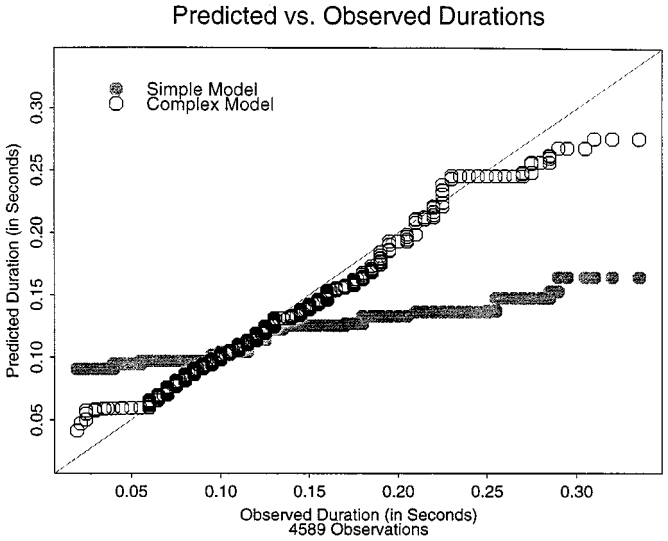


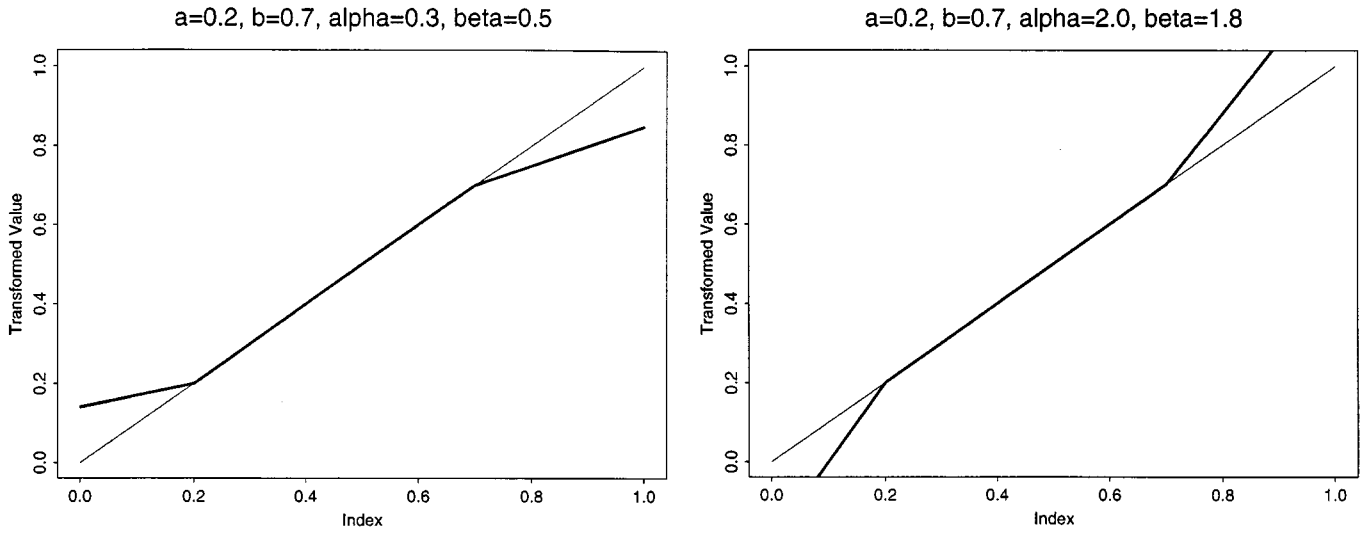
Fig. 4. Effects of adding more regression variables.

C. Empirical Observations

All empirical evidence regarding duration modeling was gathered from the *Prosodic Contexts* subcorpus mentioned in Section II. After collection, phoneme boundaries were automatically aligned using a speaker-dependent version of the Apple large vocabulary continuous speech recognition system. A SoP algorithm was implemented via weighted least-squares multiple regression, as implemented in the S+ v3.2 software package. One distinct model was computed for each of 15 classes of phonemes, across which, for simplicity, we used a common set of factors. These included accent, preceding and following phoneme identity, and similarly well-known factors reported in the literature. In all cases, the standard log transformation was used. The overall fit obtained was comparable to published results.

However, close analysis of the residuals showed that they were not spread evenly throughout the data range. Specifically, long durations tended to be underestimated and short durations overestimated. This is, of course, a common modeling phenomenon, which typically becomes less and less severe as the models acquire more independent variables representing higher-order interactions between contexts.

Fig. 4 illustrates this error reduction for a subset of the above data (consisting of the four unvoiced fricatives). The predicted and observed values have each been sorted in ascending order, and the two distributions plotted against each other. If the predictions were perfect, all the points would lie on the dotted grey diagonal line. Instead, overestimated durations are above the line and underestimated durations are below it. The grey filled circles represent the predictions from a simple, four-factor SoP model (comprising a total of about 20 regression coefficients), and the black hollow circles represent a more complex, 40-factor model (comprising a total of about 200 regression coefficients). The two models account for 32.6% and 87.2% of the total standard deviation, respectively. Clearly, the additional parameters allow the complex model to more closely predict the more extreme observations in the data. Nevertheless, the overall shape of



(a) (b)

Fig. 5. Transformation with (a) compression at both ends and (b) expansion at both ends.

both sets of predictions suggests that the overestimation of short durations and underestimation of long durations is a structural pattern over a wide range of regression equations. Moreover, this observation is consistent across the entire dataset.

Two solutions could be considered. The traditional approach (cf. Fig. 4) is to add more independent variables, representing orthogonal interaction terms, to the regression equation. However, each variable added represents only one particular higher-order interaction between factors, and thus only one specific subset of the data. As more interaction terms are added, they are trained on fewer and fewer points and account for smaller and smaller particular subsets of the outliers. At the extreme, this raises the issue of parameter reliability, as well as generalization to new combinations of context.

The other approach is to first apply an appropriate transformation to the raw durations, to compensate as much as possible for the structural nature of the pattern observed in the residuals. This fits naturally within the framework of (9), which leads us to search for a class of functions F compatible with the above observations.

D. New Transformation

In fact, the data of Fig. 4 suggests that some interactions are only amplificatory for long durations: when durations are short, these interactions seem to exert the opposite influence. This general pattern seems to support the need for compression at both ends of the range, which suggests the presence of at least one inflection point in F. This observation first led us to consider a sinusoidal function [4]. But the parameters in this function turned out to be somewhat nonintuitive, which called for an alternative formulation [17]. We then focused on a more conventional sigmoid function, of the type widely used in neural networks, and showed that this function yielded better performance, as measured by the proportion of variance left unexplained by the regression model [31].

What ultimately matters, however, is not the variance left unexplained in the transform domain, but the variance left unexplained in the original domain, where the model is eventually applied. Compressing the durations at the two extremes of the range clearly helps improve performance in the transform domain, in part by reducing the influence of extreme durations on the linear regression. By the same token, however, it may not be as effective in the original domain. This is because, essentially, the extreme durations are now acting even more as outliers (with respect to a model which was trained to downplay their importance). To ensure good performance in the original domain, it may actually be more appropriate to *expand* the durations at the two ends of the range, to force the model to give them more weight. To make this possible, we need a more flexible transformation framework than that of either [4] or [31]. This is achieved using the following piecewise linear formulation.

Let A and B denote the minimum and maximum duration observed in the training data for the particular phoneme (or class of phoneme) under study. For each duration D observed, the associated variable

$$x = \frac{D - A}{B - A} \tag{10}$$

takes on normalized values in the interval $[0, 1]$. The piecewise linear transformation is then defined by

$$F(x) = \begin{cases} \alpha x + a(1-\alpha), & \text{if } 0 \le x < a; \\ x, & \text{if } a \le x \le b; \\ \beta x + b(1-\beta), & \text{if } b < x \le 1 \end{cases} \tag{11}$$

where the parameters a, b, α, and β (all nonnegative) help control the shape of the function. Specifically, the interval $[a, b]$ defines the identity portion of the transformation, while α and β control the amount of compression/expansion which happens in the intervals $[0, a]$ and $[b, 1]$, respectively. The values $\alpha, \beta < 1$ correspond to a compression, and the values $\alpha, \beta > 1$ correspond to an expansion. The further α and β deviates from 1, the further the extrema deviate from their original values. Fig. 5 depicts the shape of the function (11) for the two sets of values: 1) $a = 0.2$, $b = 0.7$, $\alpha = 0.3$, and $\beta = 0.5$ and 2) $a = 0.2$,

$b = 0.7$, $\alpha = 2.0$, and $\beta = 1.8$. Since the four parameters can be set independently, the function (11) is able to cover a wide range of behavior.

Referring back to Fig. 4, it seems that, regardless of model complexity, the residuals for unvoiced fricatives are disproportionately greater in long durations than in short durations. Thus, we would expect the associated transformation to impact long durations more than short durations. From the above, this implies that β deviates from 1 more than α. This was confirmed experimentally: for this phoneme class, we found an optimal value of $\alpha = 1.4$ and $\beta = 1.6$. Note that, in general, the optimal values of the parameters a, b, α, and β depend on the phoneme (class) identity, since the shape of the function is tied to the way contextual factors influence the durations of particular phonemes.

V. Pitch Modeling

Like duration modeling, pitch modeling involves both phonological and phonetic considerations. Among the various methods that have been proposed for fundamental frequency prediction, it is attractive to model these influences separately, along the lines of the superpositional approach [5]. Accordingly, we decompose contours into a relatively smooth underlying pitch contour, and a separate contribution from the influence of the phonetic segments.

A. Superpositional Modeling

Overall pitch variation is determined by the prosodic intonational structure, and thereby conveys information about the semantic role and dialogue function of sentences—information that cannot be derived from the words themselves. This is the *suprasegmental structure* of pitch variation. At the same time, however, the phonemes that make up the words strongly influence the more local behavior of pitch variation. These effects are called *segmental perturbations* of the underlying smooth variation. In many intonation synthesis models (e.g., [1], [7], [11], [33]), modeling smooth pitch contours is considered to be sufficient for intonation synthesis: segmental perturbations are ignored. This belief is often reinforced by referring to the segmental contributions as "microprosody," implying that they are microscopic relative to the linguistic influences on pitch contours (in fact, it has often been claimed that they are not perceptible at all!). Nevertheless, data from speech production and perception argue that both of these influences must be captured for synthetic speech to sound natural.

Suprasegmental structures (such as pitch accents, phrasal tones, and the overall pitch range in which these occur) jointly produce a relatively smooth underlying pitch contour, which can be thought of as commands sent to the larynx. The phonetic segments of the utterance which carries this suprasegmental information perturb the fundamental frequency values away from this otherwise smooth contour. These segmental perturbations are not negligible: often their magnitude matches or exceeds that of tonal events such as downstep or different final boundary tones (cf. Fig. 1). Segmental perturbations are not random, but rather are systematically related to the identity of the associated phonetic segments. Listeners indeed expect these perturbations, and use them to help identify the phonemes [26]. Even more importantly, listeners expect and factor out these effects when recovering the underlying intonation. It is thus necessary for them to be correctly modeled in the synthetic signal for listeners to correctly perceive both the intended intonation and the segments themselves [27].

We therefore model fundamental frequency contours as a superposition of relatively local segmental perturbations and a smooth underlying intonation contour (cf. [5], [27]). This approach is different from other superpositional models (e.g., [11], [12]) in that the superposition here is strictly limited to the segmental level. We do not decompose the underlying contour into accentual tunes riding on top of phrasal tunes. Also, in contrast with decision tree methods (e.g., [8]), we characterize the underlying contour using the ToBI transcription system [3], [30]. This leverages agreement across the major traditions of intonational analysis, and allows us to directly relate the transcription to typical text processing performed in the front-ends of synthesizers and dialogue systems.

The parameterization in our version of the superpositional pitch model is amenable to statistical estimation using a suitable corpus of data. It assumes negligible interactions between levels outside of those specifically modeled. For example, the rate and magnitude of segmental perturbations is possibly related to oral articulation. In cases where our knowledge of such interdependencies is sufficient, they can be modeled by an articulatory synthesizer (cf. [25]). Otherwise, in order to minimize violations in this assumption, it is beneficial to estimate all the parameters from a single speaker, speaking in a single and consistent style, within a single recording setup. Hence, by design, the Victoria corpus provides us with exactly the right kind of data to refine and optimize such a superpositional pitch model. The following carries out this optimization within the framework originally proposed in [27], enhancing the approach for a more accurate smooth contour estimation.

B. Suprasegmental Structure

There are a number of reasons for characterizing suprasegmental structure using the ToBI transcription system [30]. While it may not capture all of the linguistically significant variation in English intonation, it represents agreement across many different approaches to intonational analysis concerning the major and most common phenomena to be modeled, and does so in a way that is relatively free of the strong theoretical differences between the different approaches. It is less abstract then the work of [24], while inheriting many of the advantages of that work. This makes it easy to learn, and produces good agreement between different transcribers. Most importantly for speech synthesis, the ToBI system was developed to relate discourse analysis to fundamental frequency contours: the same text will carry different intonation when spoken in different contexts and with different discourse roles. Therefore the ability to change the intonation in any text according its discourse role is a requirement for successful text-to-speech synthesis. This in turn requires that the intonation model does not merely match the shape of a corpus of contours, but supports generation of appropriate new

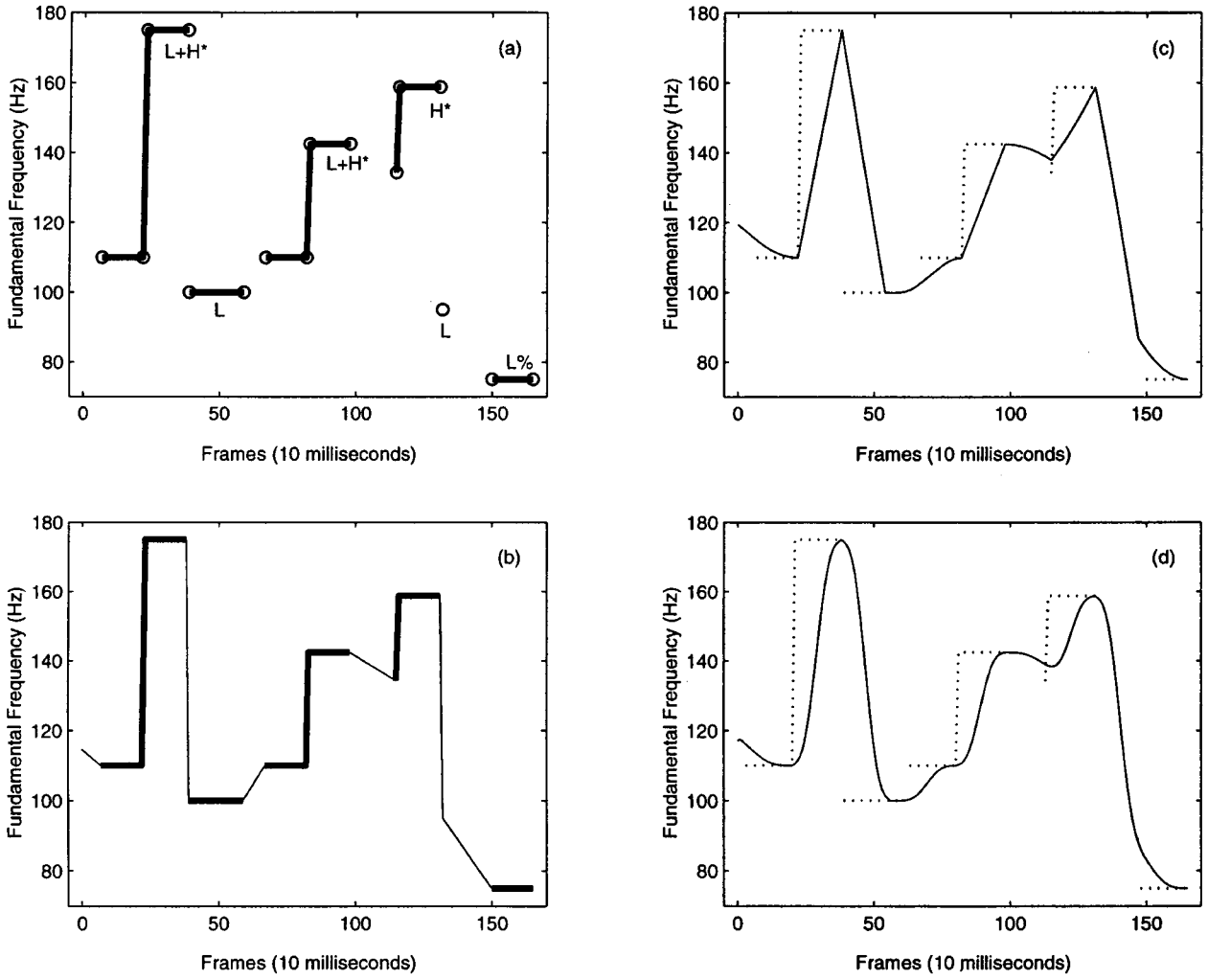


Fig. 6. Generating pitch contours from ToBI-like transcription.

contours on the basis of linguistic and pragmatic information obtained from the front-ends of synthesizers and dialogue systems. In the ToBI system, these contours are specified by the choice and locations of the pitch accents, phrase accents and boundary tones.

A common approach to generating pitch contours from a ToBI-like transcription was first presented by [1], and is illustrated in Fig. 6. Each component of the accent inventory was modeled as an abstract shape: **H*** accents for example, consist of a plateau preceded by a vertical "leg," resembling an upside-down letter "L." Final boundary tones consist merely of a plateau. On a time/frequency plot, these shapes are aligned with their associated syllables [Fig. 6(a)], and then any gaps between them are linearly interpolated [Fig. 6(b)]. Then the contour is causally smoothed by convolution with a rectangular window whose duration is equivalent to the length of the plateaux [Fig. 6(c)]. This equivalence is necessary in order to ensure that the smoothed contour reaches the target values of the accents. If the window is wider than the plateaux then the targets will be undershot.

Silverman [27] modified this approach in a number of ways. The relative heights of fundamental frequency events were scaled exponentially, rather than linearly. This is psychoacoustically more defensible, it allows **H** and **L** tones to be treated symmetrically, and does away with the need for phrase-final raising in some question contours. Another modification concerned the window shape. The rectangular window often produces very sharp peaks and valleys, and sudden changes in the direction of pitch contours, as illustrated in Fig. 6(c). Natural pitch contours do not exhibit such sudden changes in direction, except when these are induced by segmental perturbations. Production of such discontinuities in pitch or its derivatives from a human larynx would require brief moments of extremely high muscular force, and sudden extreme changes in muscular force. Studies of human motor control often characterize muscular movement in terms of peaks of acceleration and rate of change of acceleration (known as "jerk"). Nelson [22] found that human speakers economize on effort by minimizing jerk, and that the minimum-jerk velocity profile for muscular movement is almost indistinguishable from simple harmonic motion. Therefore the rectangular window was replaced with a Hamming window. This minimizes the higher derivatives, approaching the steps of pitch accents in a way that reduces jerk and hence speaker effort. It maintains the model property of guaranteeing that targets are reached, but in a smoother way [Fig. 6(d)].

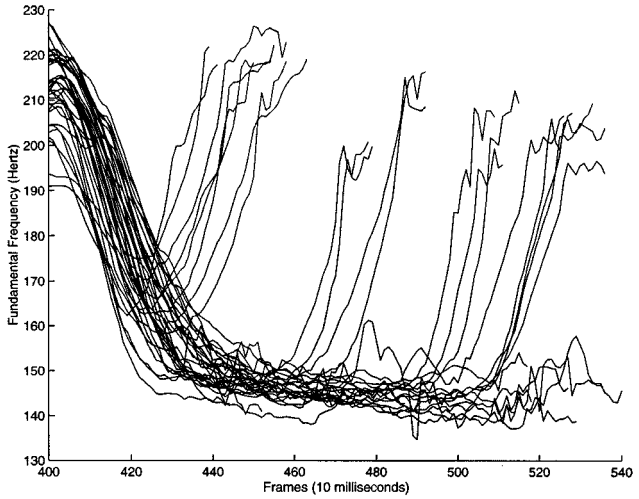


Fig. 7. Nuclear falling and falling–rising accents, aligned by $\mathbf{H}^*$ peak.

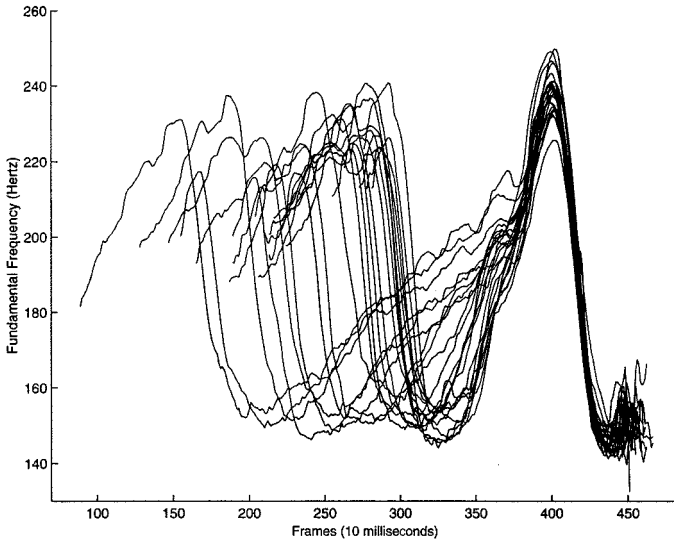


Fig. 8. Two-accent contours $\mathbf{H} + \mathbf{L}^* \mathbf{H}^* \mathbf{L}\,\mathbf{L}\% \mathbf{H}^*$–$\mathbf{L}$ transition.

C. Smooth Contour Estimation

Still, there are a number of problems associated with the use of a fixed-duration window to generate the relatively invariant local shape of pitch accents and the overall global interpolation between pitch events. At best, this can only be partially successful, because the window cannot satisfactorily handle both the local shapes and the necessary smoothness of the contours. The issue is illustrated in Figs. 7 and 8. As mentioned earlier, the suprasegmental behavior of pitch is characterized from the EGG signal as opposed to the original signal, to minimize segmental contamination.

Fig. 7 shows EGG-derived pitch contours for two of the tunes previously discussed: 1) a nuclear fall $\mathbf{H}^*\ \mathbf{LL}\%$ and 2) a nuclear fall–rise $\mathbf{H}^*\ \mathbf{L}\,\mathbf{H}\%$. Since these tunes are associated with the most perceptually salient and most semantically important words, it is important to synthesize them well. This is difficult, however, when they are spread over different lengths of segmental material and different number of words. In Fig. 7 it is clear that these two different tunes have much in common. Both have a pitch peak on the accented word, followed by a steep fall, no matter how far to the end of the utterance. (In the figure, all of the contours are aligned by this peak.) The primary difference between the tunes is the presence or absence of a final rise corresponding to $\mathbf{H}\%$. This is amenable to being modeled by a sequence of a low and high target aligned with the end of the utterance, independently of the nuclear $\mathbf{H}^*$ to $\mathbf{L}$ transition. While this argues in favor of a locally invariant characteristic shape for the pitch accent, it also transpires that pitch movements have a variable degree of steepness that cannot be rendered well by a fixed-duration window. For example, rises tend to be less steep than falls, and also less steep in prenuclear than nuclear position.

Fig. 8 shows EGG-derived pitch contours for the tone sequence $\mathbf{H} + \mathbf{L}^*\ \mathbf{H}^*\ \mathbf{L}\,\mathbf{L}\%$ aligned by the $\mathbf{H}^*$ to $\mathbf{L}$ transition. The utterances differ in the number of unaccented syllables separating the $\mathbf{H} + \mathbf{L}^*$ from the following $\mathbf{H}^*$. Clearly, the final nuclear sequence can be modeled independently of the preceding material. (Similar plots aligned by the $\mathbf{H} + \mathbf{L}^*$ accent show that it, too, is contextually invariant.) Also strong in Fig. 8 is the evidence that the pitch follows an almost linear interpolation between the two accents, with all the contours converging at the start of the small local rising gesture that is the beginning of the final $\mathbf{H}^*$ (cf. frame 378). Again, this argues in favor of a global interpolation between pitch events, but a fixed-duration window tuned for fast utterance-final falls runs the risk of approaching prenuclear accent peaks too steeply.

In the present approach, we keep the basic framework of parameterizing each EGG-derived pitch contour by a number of characteristic shapes related to ToBI symbols, and using linear interpolation to fill gaps between characteristic shapes. However, the control points associated with the pitch events are placed more carefully, for example by relaxing the constraint that the "leg" preceding a tone plateau be vertical. With the abstract accent shapes thus taking more responsibility for the local accent shape, the smoothing window is required to do less work and so the smoothing window can contract. Since each characteristic shape can be described by a relatively small number of local targets (typically two or three), we can use gradient descent to *jointly* optimize the abstract shapes of the underlying accents and the window duration. Thus, with the data from the *Reiterant Speech* corpus, a single set of parameters is trained across the entire corpus.

D. Segmental Perturbations

Segmental perturbations comprise such phoneme-level phenomena as phonetic voicing or vowel height. Again using the superpositional paradigm within the framework of [27], a separate model is therefore estimated to capture perturbations induced by consonants and vowels. The vowel-induced perturbations are characterized, relative to the underlying smooth contour, by rule-based targets linked by piecewise parabolic segments. The consonant-induced perturbations are similarly characterized by cubic and parabolic decays in the preceding and following vowels, respectively. In contrast with [27], the segmental effects are calculated given the (more accurate) smooth contours empirically estimated as above.

The final pitch model is the superposition of the smooth model and the two (consonant- and vowel-induced) perturbation

models. Unlike some other methods, this approach is particularly well-suited to the analysis of the different components influencing pitch behavior. For example, it is straightforward from this model to quantify the effect of voicing on pitch. This makes it easy to relate the values predicted by the model to the large body of studies done in various languages over the past decades.

Note that this approach does not address the issue of interactions between segmental and suprasegmental components of the model. In order to model all the systematic and perceptually-relevant variation in natural contours, it would be necessary to jointly optimize both of these components. The Victoria corpus does support such joint estimation of parameters, and this optimization will be documented in a forthcoming publication.

VI. Experimental Results

This section illustrates the improved prosodic representation resulting from the new duration and pitch models. In all cases, objective assessment measures were used to automatically evaluate the models on held-out portions of the Victoria corpus.

A. Duration Modeling

For duration modeling the evaluation criterion was taken to be the proportion of standard deviation left unexplained *in the original domain* (as opposed to the transform domain). We first used a gradient descent algorithm to iteratively adjust the four parameters for each phoneme class, using the goodness of fit of the subsequent regression (in the original domain) as the criterion. This produced the following set of parameters, with the mean indicated in parentheses: $0.05 \leq a \leq 0.30$ ($\overline{a} = 0.17$); $0.65 \leq b \leq 0.95$ ($\overline{b} = 0.86$); $1.3 \leq \alpha \leq 4.2$ ($\overline{\alpha} = 1.66$); and $1.3 \leq \beta \leq 4.4$ ($\overline{\beta} = 1.61$). Thus, none of the resulting shapes showed any compression at either end of the range, and in a few cases the expansion was substantial. This underscores the suboptimality of approximating such shapes by a logarithm curve.

A 40-factor regression model along with the standard logarithmic transformation of the raw durations left 14.4% of the standard deviation unexplained. The same independent variables were then regressed against the piecewise linearly transformed durations, using the same weighted least squares implementation. This left only 12.3% unexplained, which corresponds to a reduction of 14.6% in the proportion not accounted for. This improves slightly on the results of [31], when the evaluation criterion is applied in the original domain. To put things in perspective, the root sinusoidal transformation described in [4] only achieved a reduction of 10.4% with this criterion.

The above experiments were then repeated with a range of different numbers of equation coefficients, representing different choices of factors and interaction terms, to eliminate the possibility that the above result might somehow be linked to the particular regression model selected. Fig. 9 reports the outcome, in terms of the percentage of standard deviation left unexplained as a function of the total number of parameters in the modeling (i.e., regression coefficients as well as parameters required for the transformation). It can be seen that the piecewise linear transformation (filled triangles) is consistently

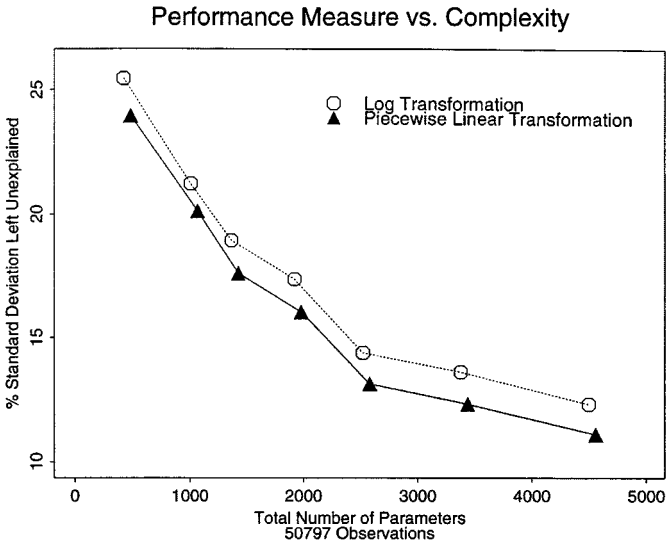


Fig. 9. Duration modeling performance.

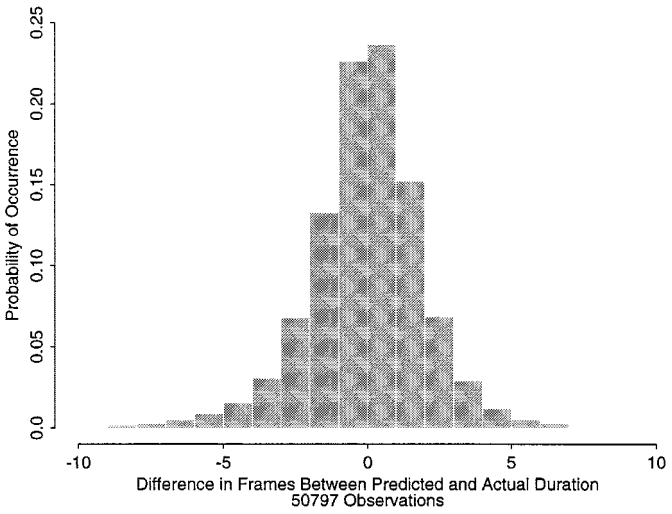


Fig. 10. Duration error probability distribution.

superior to the log transformation (hollow circles) across the entire range of parameters considered. Only when very simple models are involved is the improvement brought about by the piecewise linear framework somewhat mitigated by the extra parameters required by the transformation.

A consequence of Fig. 9 is that the piecewise linear transformation generally provides for a more parsimonious representation of the regular patterns in the observed data. In other words, for a given level of performance, the piecewise linear approach allows the underlying SoP expression to comprise less coefficients. For example, to leave 12.5% of the standard deviation in the original durations unexplained would require approximately 4500 parameters with the log transformation, but only about 3400 parameters with the piecewise linear transformation. This entails a 25% reduction in the number of parameters to estimate.

Fig. 10 shows the distribution of errors between predicted and actual durations, in terms of the number of corresponding frames. It can be seen that approximately 80% of duration errors involved a difference of one to two frames, either positive or negative, while only about 1% of the errors involved differences

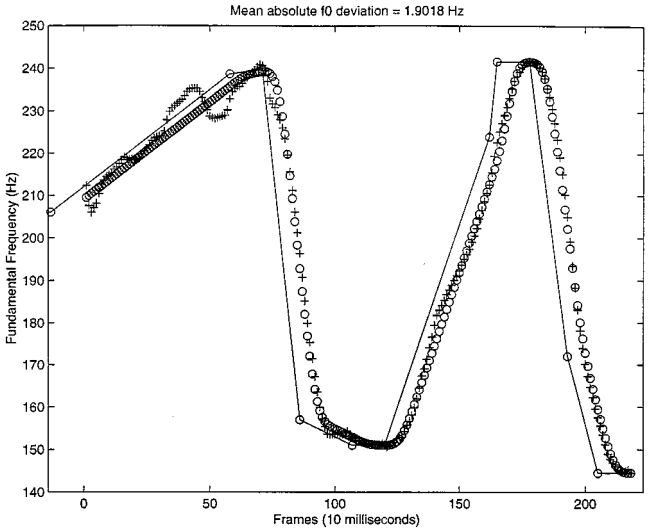


Fig. 11. Fitted (o) and actual ($+$) fundamental frequency values and derived model targets, for an utterance from Fig. 8.

greater than five frames either way. This implies that the vast majority of predictions are in the vicinity of the typical threshold for just noticeable difference (cf., e.g., [10]), with only a few "howlers" occurring from time to time.

This improved modeling has implications for the voice generation in a speech synthesizer, because it will generate a greater quantity of longer and shorter phonemes than previous approaches. Short phonemes are difficult to synthesize because they are typically associated with undershoot of articulatory targets [2]. Mere warping (in the time domain) of units that sound appropriate with longer durations is likely to result in unnaturally sudden spectral transitions. Similarly, the longer durations produced by this model will require careful voice processing to avoid unnaturally salient steady states. Consequently, we believe that as duration models improve, there will be greater need for articulatory approaches to voice generation (cf. [25]).

B. Pitch Modeling

For pitch modeling the evaluation criterion was taken to be the mean absolute frequency deviation across each utterance considered. Note that this is different from usual criteria like the root-mean-square (rms) error and the correlation (R^2) between original and synthetic contours. While such measures have been used in a number of intonation evaluations (e.g., [6]), and were shown by Hermes [13] to be well motivated similarity metrics, we believe the mean absolute frequency deviation may be even better suited for determining how well the new pitch models capture the qualities of the underlying data. This is because rms and R^2 are both based on the L_2 norm, which tends to reduce the effect of rare outliers. In pitch modeling, rare outliers can be perceptually egregious, and it might therefore be beneficial to avoid downweighting their influence.

Fig. 11 shows the result of gradient descent optimization on one of the pitch contours shown in Fig. 8. The solid line interpolates between the targets, and then undergoes causal smoothing with a Hamming window. Fitted fundamental frequency values are generated in 5-ms frames. In this example, the mean absolute deviation between fitted and original values is about 1.9 Hz. Over the entire *Reiterant Speech* subcorpus, this figure typically varies between 2–3 Hz.

Fig. 11 illustrates several characteristics of the approach. One is that most of the model error arises from the segmental perturbations remaining in the EGG signal (see, for example, frames 50–90). This confirms that the residual from the fitted smooth contours will provide good data for systematically modeling the perturbations. Another is that the optimal smoothing window (60 ms long) turns out to be considerably shorter than in prior work (e.g., 180 ms in [27]). Consequently, the local accent targets bear more responsibility for generating the smooth curves. For example, an extra target (at frame 98) was needed between the high plateau at the start of $\mathbf{H}+\mathbf{L}^*$ (at frame 67) and the low plateau corresponding to $\mathbf{L}^*$ (at frame 118). This extra target became part of the abstract characteristic shapes for all $\mathbf{H}+\mathbf{L}^*$ accents. Finally, note that the first target (at frame 0) occurs 40 ms before the start of the utterance. The data showed that when modeled this way, the pitch contour across unaccented utterance-initial syllables could be modeled with one initial value across the entire corpus. This holds regardless of whether it subsequently rises to a $\mathbf{H}$ tone or drops to a $\mathbf{L}$ tone on the first accent, and regardless of the distance to that accent.

VII. Conclusion

The Victoria corpus collection effort has underscored a number of interesting lessons. First, despite superficial similarity, different criteria should be used for the design of speech synthesis and speech recognition corpora. In the latter, background noise and variability are desired, and lack of coverage of rare cases is often of little consequence. In the former, distortions, noise, and variability that characterize most real-world conditions will result in poorer synthesis quality, and coverage is necessary because most combinations of phonetic and prosodic contexts are rare. Second, collecting such a large speech synthesis corpus presents very real challenges. Consistency of vocal effort, pitch range, voice quality, and speaking rate across multiple months of recording sessions is as crucial as it is difficult. It helped to have spoken examples of the desired speaking style and intonational tunes available, and to make frequent reference to them.

The *Prosodic Contexts* portion of the Victoria corpus was instrumental to uncover empirical evidence for the use of a piecewise linear transformation in the well-known sums-of-products approach to duration modeling. Compared to the standard log transformation, the piecewise linear function reduced the proportion of the standard deviation left unexplained in the original domain by about 15%. Alternatively, at a given operating point, it reduced the number of parameters required by about 25% at usual levels of complexity.

The Victoria corpus was also instrumental for pitch modeling. The smooth and reliable contours extracted from the EGG signals of the *Reiterant Speech* subcorpus enabled the estimation of more accurate characteristic shapes, as objectively illustrated by a low mean absolute frequency deviation (between 2 and 3 Hz) between original and synthetic fundamental frequency variations. This in turn supports a better (both more complete and more realistic) model of pitch behavior. We are currently

gathering large-scale subjective evidence to confirm that the improved prosodic representation resulting from these models leads to more natural-sounding synthetic speech.

ACKNOWLEDGMENT

The authors would like to thank their colleague D. Naik for optimizing a phoneme aligner based on the Apple large vocabulary continuous speech recognition system, and generating the phoneme boundaries.

REFERENCES

[1] M. Anderson, J. Pierrehumbert, and M. Liberman, "Synthesis by rule of English intonation patterns," *Proc. IEEE*, vol. 72, pp. 2.8.1–4, 1984.
[2] G. Bailly, "Learning to speak: Sensory–motor control of speech movements," *Speech Commun.*, vol. 22, no. 2–3, pp. 251–267, 1998.
[3] M. Beckman, "Local shapes and global trends," in *Proc. Int. Conf. Phon. Sci.*, 1995, pp. 100–107.
[4] J. R. Bellegarda and K. E. A. Silverman, "Improved duration modeling of English phonemes using a root sinusoidal transformation," in *Proc. Int. Conf. Spoken Language Proc.*, Sydney, Australia, Dec. 1998, pp. 21–24.
[5] W. N. Campbell, *Synthesizing Spontaneous Speech*, Y. Sagisaka, W. N. Campbell, and H. Norio, Eds. Berlin, Germany: Springer-Verlag, 1997, pp. 165–186.
[6] R. A. J. Clark and K. E. Dusterhoff, "Objective methods for evaluating synthetic intonation," in *Proc. 6th Eur. Conf. Speech Commun. Technol.*, vol. 4, Budapest, Hungary, Sept. 1999, pp. 1623–1626.
[7] F. Cooper, "Some reflections on speech research," in *The Production of Speech*, P. MacNeilage, Ed. New York: Springer-Verlag, 1993, pp. 275–290.
[8] K. E. Dusterhoff, A. W. Black, and P. Taylor, "Using decision trees within the tilt intonation model to predict F0 contours," in *Proc. 6th Eur. Conf. Speech Commun. Technol.*, vol. 4, Budapest, Hungary, Sept. 1999, pp. 1627–1630.
[9] C. Fougeron and P. Keating, "Articulatory strengthening at edges of prosodic domains," *J. Acoust. Soc. Amer.*, vol. 101, pp. 3728–3740, 1997.
[10] A. Friberg, "A quantitative rule system for musical performance," Ph.D. dissertation, Dept. Speech, Music, Hearing, Royal Inst. Technol., Stockholm, Sweden, 1995.
[11] H. Fujisaki, "Dynamic characteristics of voice fundamental frequency in speech and singing," in *The Production of Speech*, P. MacNeilage, Ed. New York: Springer-Verlag, 1993.
[12] N. Grønnum, "Superpositional and subordination in intonation: A nonlinear approach," in *Proc. Int. Conf. Phon. Science*, 1995, pp. 124–131.
[13] D. J. Hermes, "Measuring the perceptual similarity of pitch contours," *J. Speech Lang. Hear. Res.*, vol. 41, pp. 73–82, Feb. 1998.
[14] D. H. Klatt, "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence," *J. Acoust. Soc. Amer.*, vol. 59, pp. 1209–1221, 1976.
[15] D. H. Krantz, R. D. Luce, P. Suppes, and A. Tversky, *Foundations of Measurement, Vol. I*. New York: Wiley, 1971, ch. 6–7, pp. 245–368.
[16] F. Kubala, J. R. Bellegarda, J. R. Cohen, D. Pallett, D. B. Paul, M. Phillips, R. Rajasekaran, F. Richardson, M. Riley, R. Rosenfeld, R. Roth, and M. Weintraub, "The hub and spoke paradigm for CSR evaluation," in *Proc. ARPA Speech Natural Language Workshop*, Mar. 1994, pp. 40–44.
[17] K. Lenzo, "personal communication," unpublished, Aug. 1998.
[18] M. Y. Liberman and L. A. Streeter, "Use of nonsense–syllable mimicry in the study of prosodic phenomena," *J. Acoust. Soc. Amer.*, vol. 63, no. 1, pp. 231–233, 1978.
[19] "Linguistic Data Consortium pronouncing Lexicon," [Online] Available: http://www.ldc.upenn.edu/Catalog/LDC97L20.html.
[20] A. Magbouleh, "An empirical comparison of automatic decision tree and linear regression models for vowel durations," in *Proc. Annu. Meeting ACM*, Santa Cruz, CA, 1996.
[21] L. H. Nakatani, K. D. O'Connor, and C. H. Aston, "Prosodic aspects of American English rhythm," *Phonetica*, vol. 38, pp. 84–106, 1981.
[22] W. L. Nelson, "Physical principles for economies of skilled movements," *Biol. Cybern.*, vol. 46, pp. 135–147, 1983.
[23] D. B. Paul and J. Baker, "The design for the Wall Street Journal–based CSR corpus," in *Proc. 5th DARPA Speech Natural Language Workshop*, Feb. 1992, pp. 357–362.
[24] J. Pierrehumbert, "The phonology and phonetics of English intonation," Ph.D. dissertation, Mass. Inst. Technol., Cambridge, MA, 1980.
[25] J. Schroeter and M. M. Sondhi, "Speech coding based on physiological models of speech production," in *Advances in Speech Signal Proc.*, S. Furui and M. M. Sondhi, Eds. New York: Marcel Dekker, 1992, pp. 231–268.
[26] K. E. A. Silverman, "Segmental perturbations depend on intonation: The case of the rise after voiced stops," *Phonetica*, 1986.
[27] ——, "The structure and processing of fundamental frequency contours," Ph.D. dissertation, Cambridge Univ., Cambridge, U.K., 1987.
[28] K. E. A. Silverman and J. B. Pierrehumbert, "The timing of prenuclear high accents in English," *J. Acoust. Soc. Amer.*, vol. 82, 1987.
[29] K. E. A. Silverman, "Utterance–Internal prosodic boundaries," in *Proc. Australian Conf. Speech Science Technology*, 1988.
[30] K. E. A. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. B. Pierrehumbert, and J. Hirschberg, "ToBI: A standard for labeling English prosody," in *Proc. Int. Conf. Spoken Language Processing*, vol. 2, Banff, AB, Canada, August 1992, pp. 867–870.
[31] K. E. A. Silverman and J. R. Bellegarda, "Using a sigmoid transformation for improved modeling of phoneme duration," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, Phoenix, AZ, March 1999, pp. 385–388.
[32] R. Sproat, Ed., *Multilingual Text-to-Speech Synthesis: The Bell Lab Approach*. Norwell, MA: Kluwer, 1998, ch. 2 and 5.
[33] J. Hart, R. Collier, and A. Cohen, *A Perceptual Study of Intonation—An Experimental–Phonetic Approach to Speech Melody*. Cambridge, U.K.: Cambridge Univ. Press, 1990.
[34] J. P. H. van Santen, "Contextual effects on vowel duration," *Speech Commun.*, vol. 11, pp. 513–546, 1992.
[35] J. P. H. van Santen, "Assignment of segmental duration in text-to-speech synthesis," *Comput. Speech Lang.*, 1994.

Jerome R. Bellegarda (M'87–SM'98) received the Dipl. Ing. degree (summa cum laude) from the Ecole Nationale Supérieure d'Electricité et de Mécanique, Nancy, France, in 1984, and the M.S. and Ph.D. degrees in electrical engineering from the University of Rochester, Rochester, NY, in 1984 and 1987, respectively.

In 1987, he was a Research Associate with the Department of Electrical Engineering, University of Rochester, developing multiple access coding techniques. From 1988 to 1994, he was a Research Staff Member with the IBM T. J. Watson Research Center, Yorktown Heights, NY, working on speech and handwriting recognition, particularly acoustic and chirographic modeling. In 1994, he joined Apple Computer, Inc., Cupertino, CA, where he is currently Principal Scientist in Speech Recognition in the Spoken Language Group. At Apple, he has worked on speaker adaptation, Asian dictation, statistical language modeling, advanced dialog interactions, and voice authentication. He has written more than 70 journal and conference papers and holds 15 patents. He has contributed chapters to several books including *Advances in Handwriting and Drawing: A Multidisciplinary Approach* (Paris, France: Europia, 1994), *Automatic Speech and Speaker Recognition: Advanced Topics* (Norwell, MA: Kluwer, 1996), and *Robustness in Language and Speech Technology* (Dordrecht, The Netherlands: Kluwer, in press). His research interests include voice-driven man–machine communications, multiple input/output modalities, and multimedia knowledge management.

Dr. Bellegarda was a Member of the ARPA CSR Corpus Coordination Committee from 1992 to 1994. He is currently a Member of the Speech Technical Committee of the IEEE Signal Processing Society. He serves as Associate Editor in the area of language modeling for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING .

Kim E. A. Silverman, photograph and biography not available at time of publication.

Kevin Lenzo, photograph and biography not available at time of publication.

Victoria Anderson, photograph and biography not available at time of publication.