Synthesis of Prosody using Multi-level Unit Sequences [⋆]

Jan van Santen, Alexander Kain, Esther Klabbers, & Taniya Mishra

Center for Spoken Language Understanding
OGI School of Science & Engineering
Oregon Health & Science University

Abstract

Generating meaningful and natural sounding prosody is a central challenge in text-to-speech synthesis (TTS). In traditional synthesis, the challenge consists of how to generate natural target prosodic contours and how to impose these contours on recorded speech without causing audible distortions. In unit selection synthesis, the challenge is the sheer size of the speech corpus that is needed to cover all combinations of phone sequences and prosodic contexts that can occur in a given language. This paper describes new methods that are being explored, based on the principle of *superpositional prosody transplant.*

Both methods are based on the following procedure. In a recorded, prosodically and phonemically labeled corpus, the log pitch contours are additively decomposed into component curves according to a prosodic hierarchy, typically phrase curves (corresponding to phrases), accent curves (corresponding to feet), and segmental perturbation (or residuals) curves. During synthesis, the corpus is searched for multiple unit sequences: A unit sequence that covers the target phoneme string, and one or more unit sequences that cover the prosodic labels at a given phonological level (e.g., the foot or phrase) and are constrained by being matched to the phone match sequence in terms of the phonetic classes of the phonemes (or in terms of higher level entities, such as the number of feet and their sizes measured in syllables). The methods differ in terms of the level of detail of these constraints. A superpositional prosody transplant procedure generates a target pitch contour by extracting and recombining component curves from these sequences, and imposing this contour on the sequence that matches the phone string using standard speech modification methods. This process minimizes prosodic modification artifacts, optimizes the naturalness of the target pitch contour, yet avoids the combinatorial explosion of standard unit selection synthesis.

[⋆] We thank Xiaochuan Niu, Johan Wouters, and Paul Hosom for insightful comments. We owe several of the key ideas to Mike Macon, who died in 2002, and to whom this paper is dedicated. This material is based on work supported by the

1 Introduction

Generating meaningful and natural sounding prosody is a central challenge in text-to-speech synthesis (TTS). Two broad classes of methods are currently used. In both methods, natural language processing algorithms are used to generate a multi-layered symbolic, linguistic representation of the input text, or *Linguistic Data Structure*. In the *Traditional Concatenative Synthesis* method, target contours are computed by rule from the linguistic data structure, and these contours are then imposed on stored speech units using signal modification methods such as Linear Predictive Coding (2), PSOLA (3), sinusoidal modeling (4), or MBROLA (5). The *Unit Selection Synthesis* method uses neither target contours nor signal modification. Instead, a large, labeled speech corpus is searched for a sequence of speech intervals whose labels match the linguistic data structure. If a match can be found, then the resulting speech is simply a sequence of intervals of digitized natural speech and can be indistinguishable from natural speech.

This paper briefly discusses the problems inherent in the two methods, and then proposes a new approach that combines elements from both. This approach is based on the idea of *re-combining natural prosodic contours and phoneme sequences using a superpositional framework*; the latter refers to a general class of models that have been proposed by Ohman, Fujisaki, and others (6; 7; 8; 9; 10). Two instantiations of this approach are proposed, the second of which builds on earlier work described in (11) and in (12). The goal of this paper is primarily theoretical, with emphasis on (i) how this superpositional framework addresses problems in traditional methods and on (ii) which problems need to be addressed in order to build a complete system based on this framework.

2 Limitations of Current TTS Methods

2.1 Traditional Concatenative Synthesis

In this method, the quality of the generated speech prosody depends on two factors: the naturalness of the target contours and the absence of signal modification distortions. Although progress has been made on both fronts, the

National Science Foundation under Grants No. 0205731 ("ITR: Prosody Generation for Child Oriented Speech Synthesis"), jointly with Alan Black and Richard Sproat; 0313383 ("ITR: Objective Methods for Predicting and Optimizing Synthetic Speech Quality"); and 0082718 ("ITR: Modeling Degree of Articulation for Speech Synthesis").

current popularity of unit selection synthesis illustrates that neither problem is considered as having been fully solved. One fundamental problem is that prosodic control factors such as word stress and proximity to phrase boundaries affect multiple acoustic dimensions, including the fine temporal structure of the speech signal, pitch, spectral balance, and spectral dynamics. Both the task of computing target contours and the task of imposing these contours on speech have proven to be difficult.

2.2 Unit Selection Synthesis

The limiting factor in this method is the availability in the speech corpus of units that match any linguistic data structure that the system may be called upon to synthesize. It is well-known [e.g., (13; 14; 15)] that the number of distinct prosodic and phonemic contexts that a given phone sequence can occur in is extremely large in unrestricted domains, and even in restricted domains such as names and addresses. In fact, the probability is near-certainty that a given input text will require phone sequence / context combinations that the speech corpus does not have. These problems are ameliorated as a result of two factors. One is that the frequency distribution of phone sequence / context combinations is extremely uneven, so that frequency-optimized speech corpora can have much better coverage than corpora with randomly selected text. Second, not all contextual distinctions are associated with audible acoustic differences. Thus, the system may benefit from the presence of phone sequence / context combinations that are acoustically similar to combinations the system is searching for but that are absent.

Nevertheless, Unit Selection Synthesis faces three profound problems. First, the sole avenue for quality improvement lies either in ever-larger speech corpora or in limiting the system to restricted domains. Second, there is an increasing interest in highly expressive speech. This poses problems for Unit Selection Synthesis because it increases the combinatorics and it creates larger pitch excursions that are more likely to cause prosodic discontinuities. Third, concept-to-speech and human-machine dialogue applications make mark-up language driven TTS increasingly more important. Mark-up tags make similar demands on the TTS engine as expressive speech.

3 Proposed Methods

The proposed methods attempt to address three issues: The naturalness of target contours, minimization of the amount of signal modification, and minimization of the amount of recordings needed to achieve coverage of a given domain.

The key idea is to use prosodic target contours that are extracted from the speech corpus itself, unlike Traditional Concatenative Synthesis where the target contours are synthetic. By thus de-coupling the prosodic and the phonemic requirements of the units, we also avoid the combinatorial problem in Unit Selection Synthesis where these requirements must be satisfied by one and the same unit sequence. Thus, the methods are based on *multiple unit sequences*. The first of these sequences consists of speech intervals whose phonemic labels match the input sentence (”*phoneme unit sequence*”) and the remaining sequences each consist of speech intervals whose prosodic labels match the input sequence at the level of some phonological entity such as a foot[1] or a phrase (”*prosodic unit sequences*”). The pitch contours from these sequences are then combined additively into a target pitch contour that is imposed on the first sequence using standard signal modification methods.

The key advantages of multiple unit sequence concept are, compared to traditional synthesis, that the pitch contours are natural instead of synthetic, and, compared to unit selection synthesis, that the combinatorics has been reduced from a quadratic to a linear problem because we are no longer searching for a unit sequence that simultaneously satisfies the phonemic and prosodic requirements.

The key, overall differences between the proposed methods and the work by Raux and Black (12) are the following. First, instead of constructing the target pitch contour by concatenating raw pitch contours, we use a *superpositional* approach: We construct the target pitch contour by addition of *component contours*. These component contours belong to various *contour classes* that, as is standard in superpositional approaches (7; 9) in turn are tied to different levels in a *phonological hierarchy* (segments, feet, phrases, etc.). These component contours are extracted from *multiple* prosodic unit sequences, again corresponding to these different phonological levels. This process has two advantages: It completely avoids pitch contour discontinuities, and it enables naturalness of pitch at multiple time scales.

Second, instead of generating segmental durations by a mechanism that does not make use of the information available in these multiple unit sequences (e.g., using traditional models for segmental duration prediction such as CART, or Sums-of-Products) we define duration *contours* and transplant these from the prosodic unit sequences to the phoneme unit sequence. This has the advantage of ensuring that the local temporal segmental structure and the time course of the pitch contours are tightly coordinated.

[1] We define a foot here in the classical Abercrombie sense and not the Jassem Narrow Rhythm Unit sense, as an accented syllable followed by zero or more unaccented syllables without regard to word boundaries; see (1).

The two methods proposed, which will be called the *Quadruples method* and the *Superpositional Unit Selection method*, differ in terms of the requirements on the speech corpus. Specifically, the corpus for the Quadruples method must have a specific structure while the corpus for the Superpositional Unit Selection method is far less restricted.

3.1 Superpositional Prosody Transplant Method I: Quadruples Method

The fundamental idea is to create a speech corpus consisting of phone sequence phonemic / prosodic context combinations that form a specially structured subset of the set of all such combinations, and then use a prosody transplantation method to generate the remaining combinations from this subset.

3.1.1 Completeness of Incidence Matrices with Missing Data

Following (11), we use the notation u_1, u_2, $\cdots$ to denote phone sequences, c_1, c_2, $\cdots$ to denote prosodic contexts, and (u_1, c_1), (u_1, c_2) for their combinations. For example, $u_1 = [wi : j]$ and $c_1 = (\textit{phrase initial}, \textit{unstressed}, ...)$ characterizes the sequence of phones and the corresponding prosodic / phonemic context for the initial part of the phrase "..., we use ..."

Let $\mathbf{S}$ and $\mathbf{C}$ be the sets of phone sequences and contexts in a given domain. If one has a *recombination method* for generating (u_k, c_m) from (u_i, c_j), (u_k, c_j), and (u_i, c_m), then one can generate any (u_p, c_q) if the following is true. First construct the binary $\#\mathbf{S} \times \#\mathbf{C}$ incidence matrix $\mathbf{M}$, in which cell (i, j) contains 1 whenever (u_i, c_j) is present in the speech corpus, and 0 otherwise. Matrix $\mathbf{M}$ is said to be *complete* if iterative application of the following rule (known as the *R-method* (16)) causes each cell in the matrix to contain 1:

$$\begin{array}{l} \text{If } \mathbf{M}_{\mathrm{ij}} \ = \ 1, \ \mathbf{M}_{\mathrm{im}} = 1, \ \text{and } \mathbf{M}_{\mathrm{kj}} = 1 \\ \text{then } \mathbf{M}_{\mathrm{km}} \ \rightarrow \ 1 \end{array}$$

In other words, if (i) a combination method is available, (ii) the incidence matrix of a given corpus is complete, and (iii) the sets $\mathbf{S}$ and $\mathbf{C}$ cover all phone sequences and contexts in the target domain, then all combinations needed for the target domain can be generated. An example of a complete matrix is a matrix in which $\mathbf{M}_{1j} = 1$ for all j and $\mathbf{M}_{i1} = 1$ for all i.

Because, as this example suggests, only $\#\mathbf{S} + \#\mathbf{C} - \mathbf{1}$ cells need to be occupied for $\mathbf{M}$ to be complete, the amount of recordings necessary for coverage could be reduced by orders of magnitude compared to unit selection based synthesis. To illustrate, if we let **S** be the set of diphones in English and **C** a set of contexts known to affect prosody (e.g., combinations of word stress, sentence accent, within-phrase word location, ...; (17)), having 2,000 and 20 elements respectively, then the number of recordings is reduced from 40,000 diphone tokens to 2,019, or by 95%.

3.1.2 Recombination Method

Consider (u_1, c_1), (u_2, c_1), and (u_1, c_2). For example let $u_1 = [wi]$, $u_2 = [jo]$, $c_2 =$ unstressed, and $c_1 =$ stressed. The proposed recombination method measures the *difference between* (u_1, c_1) and (u_1, c_2) and applies this difference to (u_2, c_1) in order to obtain (u_2, c_2).

A central assumption in the proposed methods is the same assumption that underlies traditional synthesis, which is that the speech signal can be *decomposed* into segmental and prosodic information. For an example, consider Linear Predictive Coding (LPC) based synthesis where segmental information is contained in a vector (*"segmental vector"*) comprising *filter parameters* and a *voicing flag*, and the prosodic information is represented by a vector comprising the fundamental frequency and the duration of the frames (*"prosodic vector"*). Many other examples of segmental/prosodic decomposition exist, including representations in which the segmental vector contains all information about the raw speech wave and the prosodic vector information is used to modify the segmental vector at run time; the prosodic vector may contain information not only about fundamental frequency and loudness but also about the rate of spectral change in order to mimic reduction phenomena (e.g., (18; 19)) or about spectral balance (20).

Alignment of Phonemically Equivalent Phone Sequences. Two sequences u_1 and u_2 are *phonemically equivalent* if they contain the same number of phones and if in both sequences the k-th phone has the same manner of production for all k. Thus, the phone sequences in the words "medal" and "neighbor" are phonemically equivalent.

Consider intervals of speech of the type

$$\mathbf{T_{ij}} \;=\; \{t | T_{ij,\text{start}} \leq t \leq T_{ij,\text{end}}\}$$

where the first subscript corresponds to phone sequence u_i and the second subscript to context c_j. When u_1 and u_2 are phonemically equivalent, this allows defining a piecewise linear time warp function, $W_{21 \to 11}$, that relates (u_1, c_1) and (u_2, c_1) and maps $\mathbf{T_{21}}$ onto $\mathbf{T_{11}}$ by extending the correspondence between the phone boundaries in the two intervals.

Note: Throughout, the notation $W_{ij\to km}$ will be used for a time warp that maps $\mathbf{T_{ij}}$ onto $\mathbf{T_{km}}$. It will be assumed that the time warps are strictly increasing, so that $W^{-1}_{ij\to km}$ exists and is equal to $W_{km\to ij}$.

Measurement of Context Effects on Timing. Similarly, for (u_1, c_1) and (u_1, c_2), we can establish a time warp function $W_{11\to 12}$ that maps $\mathbf{T_{11}}$ onto $\mathbf{T_{12}}$. This time warp characterizes the temporal effects of the contextual change from c_1 to c_2. Because the same phone sequence is involved, this time warp does not have to rely on the piecewise linear extension of the correspondence between the phone boundaries in the two intervals, but instead can use dynamic time warping based on a frame-to-frame distance measure between the frames in the two speech intervals. It has been shown (21; 22) that certain contextual effect are far from uniform within phone intervals, and that these non-uniformities can be captured with dynamic time warping. For example, phrase-final lengthening affects primarily the final part of the vowel.

An equivalent characterization of context effects on timing is in terms of the slope of the time warp, or

$$\mathrm{Slope}_{11\to 12}(t) \;=\; W_{11\to 12}(t+1) - \; W_{11\to 12}(t) \tag{1}$$

$\mathrm{Slope}_{11\to 12}$ measures the amount of stretching or compression at time t as a result of the contextual change from c_1 to c_2.

Measurement of Context Effects on Fundamental Frequency. The procedure followed is based on *superpositional modeling.* According to this approach, F_0 contours are viewed as resulting from the additive (typically in the log frequency domain) combination of underlying curves having different temporal scopes and tied to different phonological entities. The best known of these, the Fujisaki Model (23; 7), uses *phrase curves* and *accent curves.* In other approaches (e.g., the Linear Alignment Model (9))

also segmental perturbation curves are included, representing the systematic effects of certain segmental classes on the pitch contour (e.g., F_0 is shifted upward in vowel regions during the first 50-100 ms after the offset of an obstruent.)

Denoting the F_0 contour in (u_i, c_j) as $F_0^{[i,j]}$, we decompose $F_0^{[i,j]}$ into two underlying curves, a phrase curve and a *combined* accent and segmental perturbation curve:

$$F_0^{[i,j]}(t) \;=\; C^{[i,j]}_{phr}(t) \;+\; C^{[i,j]}_{acc+seg}(t) \tag{2}$$

The phrase curves *occurring in the speech corpus* (i.e., for (i, j) = (1,1), (1,2), and (2,1)) are currently estimated manually using a graphical speech display, while the phrase curves *that are to be computed* (i.e., for (i, j) = (2,2)) are generated by rule using the Linear Alignment Model (9). $C^{[i,j]}_{acc+seg}$ is computed by subtracting $C_{phr}^{[i,j]}$ from $F_0^{[i,j]}$. Figure 1 shows examples.

The method chosen for measuring the relationship between $C_{acc+seg}^{[1,1]}$ and $C_{acc+seg}^{[1,2]}$ proceeds as follows. Letting m_{ij} denote the mean of the section of the phrase curve corresponding to the time interval spanned by (u_i, c_j), define the curve:

$$R_{11\to 12}(t) = \frac{C_{acc+seg}^{[1,2]}[W_{11\to 12}(t)] + m_{12}}{C_{acc+seg}^{[1,1]}[t] + m_{11}} \tag{3}$$

This curve describes the relationship between $C_{acc+seg}^{[1,1]}$ and $C_{acc+seg}^{[1,2]}$ as a ratio curve; the values between which the ratios are computed are taken from corresponding points in the segmental vector stream. This ratio curve is not smooth, and is subjected to smoothing using *isotonic smoothing* (24) followed by Gaussian smoothing (See Figure 2.).

Computing the Segmental Vector Sequence. The segmental vector sequence in (u_2, c_1) consists of the sequence $\{\vec{s}_{21}(t)\}$, where t ranges over the interval $\mathbf{T_{21}}$. We now use $W_{21\to 11}$ and $\text{Slope}_{11\to 12}$ to create a time warp $W_{21\to 22}$, which is then applied to $\{\vec{s}_{21}(t)\}$ to create $\{\vec{s}_{22}(t)\}$. Let:

$$\text{Slope}_{21\to 22}(\tau) \;=\; \text{Slope}_{11\to 12}[W_{21\to 11}(\tau)] \tag{4}$$

Then:

$$W_{21\to 22}(t) \;=\; \Sigma_{\tau\leq t}\text{Slope}_{21\to 22}(\tau) \tag{5}$$

Finally, denoting for a given combination (u_i, c_j) at (discrete) time t the segmental vector as $\vec{s}_{ij}(t)$:

$$\vec{s}_{22}(\tau) \;=\; \vec{s}_{21}[W_{22\to 21}(\tau)] \tag{6}$$

In words, to generate (u_2, c_2) from (u_2, c_1), we apply the same local stretch or compression factor to the time points in (u_2, c_1) as are applied to the corresponding (via $W_{21\to 11}$) time points in (u_1, c_1) to obtain (u_1, c_2).

Computing the Prosodic Vector Sequence. The generation of $F_0^{[2,2]}(t)$ proceeds as follows. First, a phrase curve, $C_{phr}^{[2,2]}$, is computed by rule, via the Linear Alignment Model. Let $t \;=\; W_{22\to 21}(\tau)$, for $\tau\epsilon\mathbf{T_{22}}$, and define

$$\begin{aligned} C_{acc+seg}^{[2,2]}(\tau) \;&=\; R_{11\to 12}(t) \;\times \\ &\quad [C_{acc+seg}^{[2,1]}(t) + m_{21}] \;-\; m_{22} \end{aligned} \tag{7}$$

Finally, let $F_0^{[2,2]}(\tau) = C_{acc+seg}^{[2,2]}(\tau) + C_{phr}^{[2,2]}(\tau)$ (Figure 1, upper right panel).

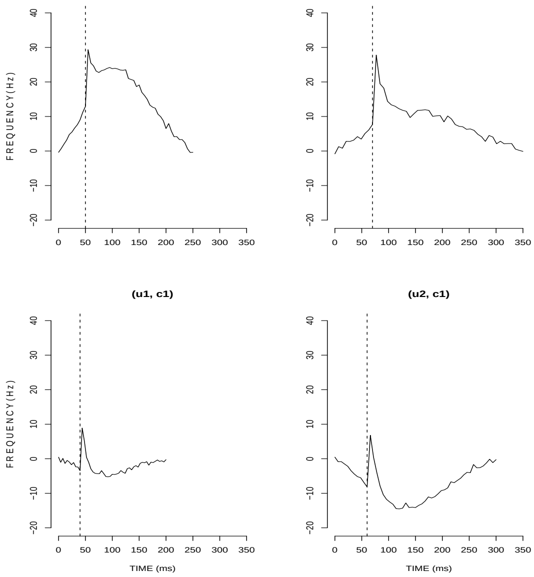

Fig. 1. $C^{[i,j]}_{acc+seg}(t)$, for i, $j = 1, 2$. Vertical lines indicate vowel onset.

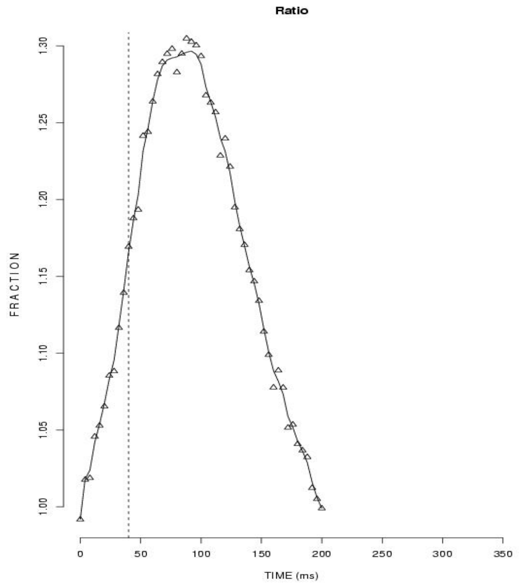

Fig. 2. $R_{11\to12}(t)$.

This operation has three important properties. First, it preserves the synchrony between local segmental perturbations of the F_0 contour and the segmental frames, because these perturbations are represented in $C^{[2,1]}_{acc+seg}(t)$ and because the multiplication curve, $R_{11\to12}(t)$, is smooth. There is evidence that certain segmental perturbations are independent of prosodic context, specifically accent status and proximity to phrase boundaries (9). An additional benefit or preserving this synchrony is that it has been shown that signal processing artifacts can be predicted by comparing the original and target pitch contours in terms of pitch values and pitch derivatives (17).

Second, the alignment of the F_0 contour, for example as measured by peak location relative to syllable boundary locations, is known to vary as a function of the manner of production of the segments associated with a pitch accent (9), specifically with the segments in the coda of the accented syllable. This fact, in combination with the need for time warps between different phone sequences, forms the primary reason for focusing on phonemically equivalent phone sequences.

Third, peak location has been shown to vary non-uniformly with the durations of the segments making up the accented (and post-accented) syllables (9). For example, a change in the duration of the onset brings about a much larger change in peak location than the same change in the duration of the nucleus or coda. The non-uniform time warping procedures reflect this result.

3.2 Superpositional Prosody Transplant Method II: Superpositional Unit Selection method

This method differs from the Quadruples method in that, instead of obtaining a target contour by applying the result of a *differencing operation* – i.e., applying the difference between (u_2, c_1) and (u_2, c_2) to (u_1, c_1) to obtain (u_1, c_2) – the method does not require the presence of such quadruples in the corpus. Instead, it involves searching for multiple prosodic unit sequences, extracting pitch component curves (or parameters characterizing these curves) from these sequences, and additively recombining them to form a target pitch contour; a roughly analogous process is described for duration.

3.2.1 Pre-processing of speech corpus

Prosodic Labeling. Following Klabbers (17), the corpus is labeled in terms of a hierarchical scheme with as key phonological entities, for example, the phoneme, syllable, word, foot, phrase, sentence, and paragraph. Each of these entities in turn is marked with tags such as for stress and location within the next-larger entity; in addition, entities are marked in terms of their internal constituents. Thus, a foot may be characterized as *medium stress, phrase final,sentence medial, consisting of 3 syllables.*

Additive Decomposition of pitch contours. As in the Quadruples Method, the pitch contours are decomposed into component curves. The curves are the segmental perturbation or residuals curve, accent curve, and the phrase curve. In addition, parameters characterizing phrase curves are extracted and used to compute patterns of phrase curve parameters characterizing entities larger than the phrase, such as the sentence and the paragraph. For example, a sentence consisting of two phrases is characterized by three time-frequency pairs for each of the phrases, where the three time points correspond to the start and end of the phrase and the start of the syllable carrying the nuclear pitch accent.

A major challenge in this step is the decomposition of pitch curves. A key reason why this is difficult is that, in contrast to the standard analysis using the Fujisaki model, we do not want to make assumptions about the shapes of these component curves. Recently, an approach has been explored in which wavelets are used to reveal the phrase curve, with promising initial results (25). It was found that this method was able to reliably recover phrase curves from pitch curves that were generated either by the Fujisaki model or from the Linear Alignment Model. The method performs a wavelet transform of the input pitch curve, removes transform components via level-dependent tresholding to remove components corresponding to the accent and residuals curves, and then performs the inverse wavelet transform to produce an estimate of the phrase curve. While these results are limited because the input curves were synthetic rather than natural, and were smooth rather than containing gaps due to non-sonorant regions, these results are nevertheless promising. Other work relevant for additive decomposition is by Sakai (26), who was able to extract *average* component curves for various prosodic contexts from a corpus *without* making assumptions about their shapes. We are currently working on methods that in addition to combining these two approaches, will also make use of the special privilege stemming from the fact that in speech synthesis, one can assume that the speech corpus has been labeled and segmented. Nevertheless, for now, we use semi-manual methods.

Relative Duration contours For each phoneme class, *intrinsic durations* are computed across the entire data base, using the multiplicative model (27). Formally, for a phonetic segment with identity p in a context characterized by levels $f_1 \cdots f_n$ on prosodic control factors $F_1 \cdots F_n$:

$$\begin{aligned} \mathrm{DUR}(\mathbf{p}, \mathbf{f_1}, \mathbf{f_2}, ..., \mathbf{f_n}) \approx \mathrm{D\hat{U}R}(\mathbf{p}, \mathbf{f_1}, \mathbf{f_2}, ..., \mathbf{f_n}) = \\ I(p) \times S_1(f_1) \times \cdots S_n(f_n) \end{aligned} \tag{8}$$

Here, $I(p)$ is an estimate of the intrinsic duration, and $S_n(f_n)$ is a parameter representing the impact of factor level f_n on factor F_n, normed such that $\prod_{f_n \in F_n} \; S_n(f_n) \; = \; 1$. Thus defined, $I(p)$ can be interpreted as an estimate of the true average of the (log) durations if the data set was perfectly balanced, i.e., if the data set contained all combinations of the levels on the control factors equally often (which is, of course, impossible.)

Assuming the multiplicative model, these intrinsic durations are used to produce relative durations, by dividing actual durations by intrinsic durations. I.e.:

$$\mathrm{DUR}_{\mathrm{rel}}(\mathbf{p}, \mathbf{f_1}, \mathbf{f_2}, ..., \mathbf{f_n}) =_{def} \mathrm{DUR}(\mathbf{p}, \mathbf{f_1}, \mathbf{f_2}, ..., \mathbf{f_n})/\mathbf{I}(\mathbf{p}) \tag{9}$$

Under this model, the resulting relative durations are equal to

$$\mathrm{DUR}_{\mathrm{rel}}(\mathbf{p}, \mathbf{f_1}, \mathbf{f_2}, ..., \mathbf{f_n}) = S_1(f_1) \times \cdots S_n(f_n) \tag{10}$$

and hence completely independent of phonemic identity; they solely reflect prosodic factors.

We note that instead of the intrinsic duration estimated via the multiplicative model we also could have used the simple overall medians or means, computed over all occurrences of a phoneme in a corpus. However, this has severe drawbacks because of correlations between phoneme identity and prosodic context (28; 27). For example, the /e/ sound (as in "bed") can never be phrase-final in English because except for loan words this vowel does not occur in open syllables. As a result, the difference in average duration between /i:/ (as in "bead") and /e/ is much larger than the difference in intrinsic duration; the latter roughly corresponds to the duration difference between the two vowels in carefully matched contexts.

Once these relative durations are obtained, the *relative duration contours* (i.e., curves depicting relative duration as a function of the position of the phonetic segments in an utterance) can be optionally smoothed. We note, however, that relative durations are not necessarily smooth because contextual effects can be highly localized. For example, we have found that the effects of strong emphasis on intervocalic /n/ (as in "banner") are zero, whereas the preceding vowel and consonant are substantially lengthened.

Using the results from the multiplicative model analysis, the relative duration contours in turn can be decomposed into parameters tied to the different levels in the phonological hierarchy.

3.2.2 Synthesis

Unit sequences. The input to the search system is a fairly conventional hierarchical phonological structure, containing target phonemes, target syllables, target feet, target phrases, etc. The corpus is searched for the following unit sequences:

Phoneme match sequence: Matches target phoneme sequence.

Foot match sequence: Matches the target foot sequence, containing the same number of phonetic segments as the target phoneme sequence, with the further constraint (the "phoneme-class constraint") that these segments belong to the same broad phonemic classes, defined as *vowel or diphthong*, *sonorant consonant*, and *other*.

Phrase match sequence: Matches the target phrase in terms of number and lengths (in syllables) of feet.

Sentence match sequence: Matches the target sentence in terms of number and lengths (in feet) of phrases.

Etc.

The search for these unit sequences jointly optimizes multiple acoustic and symbolic costs, including:

Concatenation cost: Measure of acoustic mismatch between successive units.

Acoustic phonemic target cost: Measure of acoustic distance to acoustic template of a phonetic class.

Foot/phoneme cost: Violations of the same-class constraint.

Phrase/foot cost: Mismatches between the target and the phrase match sequence in terms of the number and lengths of the feet.

Sentence/phrase cost: Mismatches between the target and the phrase match sequence in terms of the number and lengths of the feet.

Target Pitch Contour. The target pitch contour is generated by superposition of retrieved curves, with optional adjustment of their shapes. Specifically, the following steps are taken:

(1) *Residuals curve:* The residuals curve of Phoneme match sequence is time-warped synchronously with the spectral contents of this sequence, as dictated by the target duration contour.
(2) *Accent curves:* Accent curves are retrieved from Foot match sequence, timewarped to align phoneme sequences in corresponding feet.
(3) *Phrase curves:* Phrase curves are retrieved from Phrase match sequence, timewarp to align feet. Shift or tilt phrase curves as specified by Sentence match sequence.
(4) Add residual curve, accent curves, and phrase curves.

Target Duration Contour. The target durations contour is generated using the following steps:

(1) Consider a foot in the Foot match sequence consisting of phonemes $q_1, \cdots, q_m$, and the corresponding sequence of phonemes in the Phoneme match sequence, $p_1, \cdots, p_m$. Compute the following *preliminary target duration contour*:

$$\mathrm{DUR}_{\mathrm{target}}(\mathbf{p_i}) = \mathbf{I}(\mathbf{p_i}) \times \mathrm{DUR}_{\mathrm{rel}}(\mathbf{q_i})/\mathrm{DUR}_{\mathrm{rel}}(\mathbf{p_i}) \tag{11}$$

(2) Multiply the preliminary target contours with estimated parameters that reflect the contributions from factors at the phrase- end sentence-match

levels.

4 Conclusions

We believe that the proposed methods address key weaknesses in current approaches, namely the reliance on extraordinary amounts of data in Unit Selection based Synthesis and the reliance on artificial target contours and signal modification methods in Traditional Concatenative Synthesis. They do so by taking advantage of two key ideas: Recombination of natural spectral and prosodic information, and using decomposition of pitch curves (and duration curves) into re-combinable component curves.

We briefly contrast here the proposed approach with the Superposition of Functional Contours (SFC) model proposed by Bailly and his colleagues (e.g., (29)). There are two key differences. First, the SFC generates synthetic component curves using neural nets instead of extracting and recombining natural component curves. Second, the SFC model handles a prosodic hierarchy (foot, phrase, sentence, ...) by positing a component curve class for each of these levels, whereas the proposed method extracts phrase curve *parameters* to model the contributions of these levels. It is an open empirical question whether, as in the SFC model, the additivity assumption can be used to model the contributions of all levels of a prosodic hierarchy, or whether, as in the proposed approach, some of these contributions must be modeled non-additively.

In order to create a full-scale implementation of the method, several problems still need to be addressed. First, determination of the contexts **C** (see 3.1.1) in a given domain. It has been shown by Klabbers and van Santen (30; 17; 31) that “foot based tagging” provides a concise characterization of the joint factors of word stress, sentence accent, and within-phase location. However, this tagging scheme leaves out other prosodic factors such as contrastive stress and sentence mode, and certainly emotional factors, and thus needs to be extended.

Second, extending the method to prosodic features other than timing and pitch, such as spectral tilt and energy. van Santen and Niu (20) generated synthetic spectral balance trajectories, using the same methods as used in the Bell Labs system for generating synthetic target duration (32). Methods need to be generated for *transplanting* spectral balance trajectories.

Third, as remarked earlier, although significant progress has been made (25), non-supervised determination of the phrase curves is still an unsolved problem.

References

[1] Caroline Bouzon and Daniel Hirst, "Isochrony and prosodic structure in British English," in *Proc. Speech Prosody 2004*, Nara, Japan, 2004.

[2] J. Olive and M.Y. Liberman, "Text to speech – an overview," *Journal of the Acoustic Society of America, Suppl. 1*, vol. 78, no. Fall, pp. s6, 1985.

[3] F. Charpentier and E. Moulines, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," in *Proc. of Eurospeech-1989*, Paris, 1989, pp. 13–19.

[4] M. W. Macon, *Speech synthesis based on sinusoidal modeling*, Ph.D. thesis, Georgia Tech., October 1996.

[5] Th. Dutoit, *An Introduction to Text-to-Speech Synthesis*, Kluwer, Dordrecht, the Netherlands, 1997.

[6] Sven E. G. Öhman and J. Lindqvist, "Analysis-by-synthesis of prosodic pitch contours," *Speech Transmission Laboratory—Quarterly Progress and Status Report*, vol. 4, pp. 1–6, 1966.

[7] H. Fujisaki, "A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour," in *Vocal physiology: Voice production, mechanisms and functions*, Osamu Fujimura, Ed., pp. 347–355. Raven, New York, 1988.

[8] Nina Thorsen, "A study of the perception of sentence intonation—evidence from Danish," *Journal of the Acoustical Society of America*, vol. 67, pp. 1014–1030, 1980.

[9] J. van Santen and B. Möbius, "A model of fundamental frequency contour alignment," in *Intonation: Analysis, Modelling and Technology*, A. Botinis, Ed. Cambridge University Press, 1999.

[10] Y. Morlec, G. Bailly, and V. Auberg, "Generating intonation by superposing gestures," in *Proc. ICSLP '96*, Philadelphia, PA, 1996, vol. supplement.

[11] J. van Santen, L. Black, G. Cohen, A. Kain, E. Klabbers, T. Mishra, J. de Villiers, and X. Niu, "Applications of computer generated expressive speech for communication disorders," in *Proceedings of Eurospeech-2003*, Geneve, Switzerland, September 2003.

[12] A. Raux and A. Black, "A unit selection approach to f0 modeling and its application to emphasis," in *Proceedings of ASRU 2003, St Thomas, US Virgin Is*, December 2003.

[13] J. van Santen, "Combinatorial issues in text-to-speech synthesis," in *Proceedings of Eurospeech-1997*, Rhodes, Greece, September 1997.

[14] B. Moebius, "Rare events and closed domains: Two delicate concepts in speech synthesis," in *4th ISCA Tutorial amd Research Workshop on Speech Synthesis*, Pilochry, Scotland, 2001, ESCA.

[15] D.R. Baaijen, *Word frequency distributions*, Kluwer, Dordrecht, The Netherlands, 2000.

[16] Y. Dodge, *Analysis of experiments with missing data*, Wiley, New York NY, 1981.

[17] E. Klabbers and J.P.H. van Santen, "Control and prediction of the impact of pitch modification on synthetic speech quality," in *Proceedings of Eurospeech-2003*, Geneve, Switzerland, September 2003.

[18] J. Wouters and M. Macon, "Effects of prosodic factors on spectral dynamics. I. Analysis," *Journal of the Acoustical Society of America*, vol. 111, no. 1, pp. 417–427, 2002.

[19] J. Wouters and M. Macon, "Effects of prosodic factors on spectral dynamics. II. Synthesis," *Journal of the Acoustical Society of America*, vol. 111, no. 1, pp. 428–438, 2002.

[20] J. van Santen and X. Niu, "Prediction and synthesis of prosodic effects on spectral balance," in *IEEE Workshop on Speech Synthesis*, Santa Monica, California, 2002, IEEE.

[21] J. van Santen, J.C. Coleman, and M.A. Randolph, "Effects of postvocalic voicing on the time course of vowels and diphthongs," *Journal of the Acoustical Society of America*, vol. 92, no. 4, Pt. 2, pp. 2444, October 1992.

[22] J. van Santen, "Segmental duration and speech timing," in *Computing Prosody*, Y. Sagisaka, W.N. Campbell, and N. Higuchi, Eds. Springer-Verlag, New York, 1996.

[23] H. Fujisaki, "Dynamic characteristics of voice fundamental frequency in speech and singing," in *The production of speech*, Peter F. MacNeilage, Ed., pp. 39–55. Springer, New York, 1983.

[24] J. van Santen and R.W. Sproat, "High-accuracy automatic segmentation," in *Proceedings of Eurospeech-1999*, Budapest, Hungary, September 1999.

[25] J. van Santen and T. Mishra, "Estimating phrase curves in the general superpositional intonation model," in *Proc. of the 5th ISCA Speech Synthesis Workshop (SSW5), Pittsburgh*, June 2004.

[26] S. Sakai, "F0 modeling with multi-layer additive modeling based on a statistical learning technique," in *Proc. of the 5th ISCA Speech Synthesis Workshop (SSW5), Pittsburgh*, June 2004.

[27] J. van Santen, "Analyzing N-way tables with sums-of-products models," *Journal of Mathematical Psychology*, vol. 37, no. 3, pp. 327–371, 1993.

[28] J. van Santen, "Contextual effects on vowel duration," *Speech Communication*, vol. 11, pp. 513–546, 1992.

[29] S. Raidt, G. Bailly, B. Holm, and H. Mixdorff, "Automatic generation of prosody: comparing two superpositional systems," in *Proc. Speech Prosody 2004*, Nara, Japan, 2004.

[30] E. Klabbers and van Santen, "Prosodic factors for predicting local pitch shape," in *Workshop on Speech Synthesis*, Santa Monica, California, 2002, IEEE.

[31] E. Klabbers and J. van Santen, "Clustering of foot-based pitch contours in expressive speech," in *Proc. of the 5th ISCA Speech Synthesis Workshop (SSW5), Pittsburgh*, June 2004.

[32] J. van Santen, "Assignment of segmental duration in text-to-speech syn-

thesis," *Computer Speech and Language*, vol. 8, pp. 95–128, April 1994.