Speech Prosody 2004
Nara, Japan
March 23-26, 2004

Synthesis by Recombination of Segmental and Prosodic Information

Jan P. H. van Santen, Alexander Kain, & Esther Klabbers

Center for Spoken Language Understanding
OGI School of Science & Engineering at OHSU
{vansanten, kain, klabbers}@bme.ogi.edu

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 corpus based 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. A new method is proposed based on the following concepts. The set of phone sequences in a language can be partitioned in terms of the manner of production of their constituent phonemes. For each sub-class in this partition (e.g., vowel-nasal-unvoiced fricative), a representative sequence is chosen (e.g., [e]-[n]-[s]), and recorded in a wide variety of prosodic contexts. The remaining sequences in this subclass are recorded in a much smaller number of contexts, potentially only one context. The method describes a procedure for generating sequences in prosodic contexts in which they have not been recorded, by transplanting the prosodic contours of sequences in the same sub-class that have been recorded in these contexts. The method uses time warp algorithms in a superpositional framework.

10.21437/SpeechProsody.2004-93

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 [1], PSOLA [2], sinusoidal modeling [3], or MBROLA [4]. 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 method that combines elements from both. The method builds on earlier work described in [5], and forms an overall framework for the TTS research projects conducted in our lab.

We thank Xiaochuan Niu, Johan Wouters, Taniya Mishra, and Paul Hosom for insightful comments. We owe several of the key ideas to Mike Macon, who died in 2002. This material is based upon work supported by the National Science Foundation under Grants No. 0205731 ("ITR: Prosody Generation for Child Oriented Speech Synthesis") and 0082718 ("ITR: Modeling Degree of Articulation for Speech Synthesis").

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 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 proved 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., [6, 7, 8]) 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 Method

The proposed method attempts to address two issues. One is to increase the naturalness of target contours, and the other is to minimize the amount of signal modification. The key difference with Traditional Concatenative Synthesis is that the system uses natural target contours instead of target contours that are generated by rule. The key difference with Unit Selection Synthesis is that signal modification is performed.

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. Completeness of Incidence Matrices with Missing Data

Following [5], 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, $u_1 = [wij]$ and $c_1 = (phrase\ initial, unstressed, ...)$ 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* [9]) causes each cell in the matrix to contain 1:

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

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 $\mathbf{S}$ be the set of diphones in English and $\mathbf{C}$ a set of contexts known to affect prosody (e.g., combinations of word stress, sentence accent, within-phrase word location, ...; [10]), 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%.

3.2. Recombination Method

Consider (u_1, c_1), (u_2, c_1), and (u_1, c_2). For example let $u_1 = [wi]$, $u_2 = [jo]$, $c_2 = \mathrm{unstressed}$, and $c_1 = \mathrm{stressed}$. 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., [11, 12]) or about spectral balance [13].

3.2.1. Alignment of Equivalent Phone Sequences

Two sequences u_1 and u_2 are *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 equivalent.

Consider intervals of speech of the type

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

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 equivalent, this allows defining a piecewise linear time warp function, $W_{21\rightarrow 11}$, 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 $W_{ij\rightarrow km}$ 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 $W^{-1}_{ij\rightarrow km}$ exists and is equal to $W_{km\rightarrow ij}$.

3.2.2. Measurement of Context Effects on Timing

Similarly, for (u_1, c_1) and (u_1, c_2), we can establish a time warp function $W_{11\rightarrow 12}$ 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 [14, 15] 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

$$\mathrm{Slope}_{11\rightarrow 12}(t) = W_{11\rightarrow 12}(t+1) - W_{11\rightarrow 12}(t) \quad (1)$$

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

3.2.3. 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 [16, 17], uses *phrase curves* and *accent curves*. In other approaches (e.g., the Linear Alignment Model [18]) 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:

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

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 [18]. $C_{acc+seg}^{[i,j]}$ 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:

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

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* [19] followed by Gaussian smoothing (See Figure 2.).

3.2.4. 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 $W_{21\rightarrow 11}$ and $\mathrm{Slope}_{11\rightarrow 12}$ to create a time warp $W_{21\rightarrow 22}$, which is then applied to $\{\vec{s}_{21}(t)\}$ to create $\{\vec{s}_{22}(t)\}$. Let:

$$\mathrm{Slope}_{21\rightarrow 22}(\tau) \;=\; \mathrm{Slope}_{11\rightarrow 12}[W_{21\rightarrow 11}(\tau)] \qquad (4)$$

Then:

$$W_{21\rightarrow 22}(t) \;=\; \Sigma_{\tau \leq t}\mathrm{Slope}_{21\rightarrow 22}(\tau) \qquad (5)$$

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

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

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 $W_{21\rightarrow 11}$) time points in (u_1, c_1) to obtain (u_1, c_2).

3.2.5. 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 $t \;=\; W_{22\rightarrow 21}(\tau)$, for $\tau \epsilon \mathbf{T_{22}}$, and define

$$C_{acc+seg}^{[2,2]}(\tau) \;=\; R_{11\rightarrow 12}(t) \;\times\; [C_{acc+seg}^{[2,1]}(t) + m_{21}] \;-\; m_{22} \qquad (7)$$

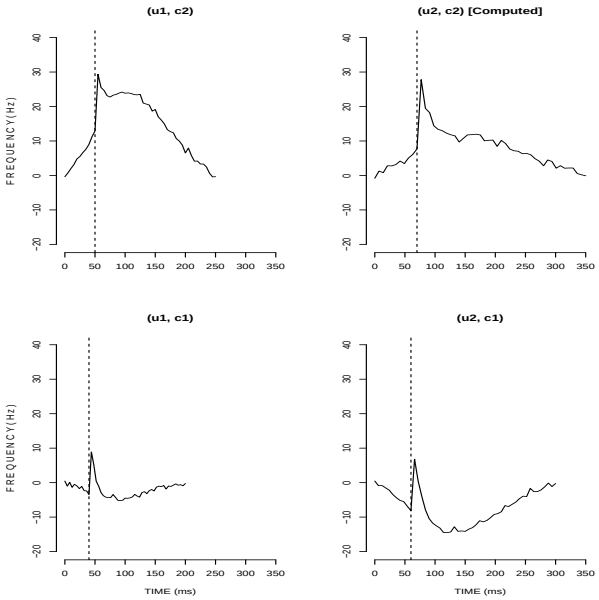

Figure 1: $C_{acc+seg}^{[i,j]}(t)$, for i, j = 1, 2. Vertical lines indicate vowel onset.

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

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 $C_{acc+seg}^{[2,1]}(t)$ and because the multiplication curve, $R_{11\rightarrow 12}(t)$, is smooth. There is evidence that certain segmental perturbations are independent of prosodic context, specifically accent status and proximity to phrase boundaries [18]. 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 [10].

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 [18], 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 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 [18]. 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.

4. Conclusions

A method was proposed that combines a scheme for speech corpus construction, based on complete incidence matrices with missing data, with a prosody recombination method, based on time warping and pitch target contour computation using smooth ratio curves within a superpositional framework. The method addresses key weaknesses in current approaches,

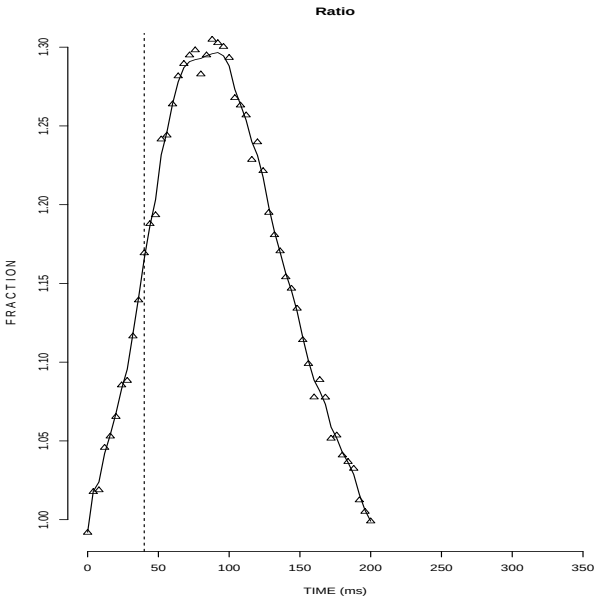

Figure 2: $R_{11\rightarrow12}(t)$.

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.

In order to create a full-scale implementation of the method, several problems still need to be addressed. First, determination of the phone sequences in a given domain, $\mathbf{S}$.

Second, determination the contexts in a given domain, $\mathbf{C}$. Elsewhere [20], it has been shown 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 thus needs to be extended.

Third, extending the method to prosodic features other than timing and pitch, such as spectral tilt and energy.

Fourth, non-supervised determination of the phrase curves. Although algorithms are available for this purpose in the context of the Fujisaki model [http://www.tfh-berlin.de/ mixdorff/fujisaki_analysis.htm], no such algorithms are available for the Linear Alignment Model.

5. References

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

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

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

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

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

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

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

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

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

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

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

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

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

[14] J. van Santen, J.C. Coleman, and M.A. Randolph, 1992, 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, p. 2444.

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

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

[17] H. Fujisaki, 1988, 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., p. 347–355. Raven, New York.

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

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

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