Prosodic Modeling in Text-to-Speech Synthesis

Jan P. H. van Santen

Lucent Technologies – Bell Labs, 600 Mountain Ave., Murray Hill, NJ 07974, U.S.A.

ABSTRACT

This paper discusses three broad obstacles that must be overcome to improve prosodic quality in text-to-speech systems. First, direct and indirect limits set by the signal processing ("*synthesis*") components. Second, combinatorial and statistical constraints inherent in generalizing from training corpora to unrestricted domains, and that require the integration of content-specific knowledge and detailed mathematical modeling. Third, the nature of many empirical research issues that must be solved for prosodic modeling to improve: they are often too focused and model-dependent for academe, and too long-term for development organizations.

1. INTRODUCTION

The standard architecture of text-to-speech (*TTS*) systems involves components for *natural language processing* (*NLP*), *acoustic prosody*, and *synthesis*. Output from NLP components is symbolic, consisting of entities such as phoneme sequences and prosodic markers, e.g., for pitch accents and phrase boundaries. Acoustic-prosodic components compute timing and pitch contour; the term "prosodic model" is used to refer to the equations involved in these computations. Finally, synthesis components generate digital speech, either by concatenating stored speech fragments or – but less frequently so – by generating acoustic parameter trajectories by rule.

Currently, intelligibility of the best TTS systems is extremely good, and certainly good enough for many real applications. However, it rarely takes a listener more than 500 ms to decide that speech generated by TTS is not recorded natural speech, let alone speech generated by an actively communicating human. It is commonly assumed that lack of natural prosody is the main reason for this.

The question asked in this paper is how to improve the prosodic quality of TTS. The structure of the paper is as follows. It starts by providing some examples of prosodic modeling, and then discusses *domain coverage* – a concept critical for prosodic modeling and for TTS system construction in general. Next, the paper addresses the often neglected role played by synthesis components in the generation of prosody. After a discussion of the statistical/combinatorial challenges faced by prosodic modeling and making a case for the importance of *content-rich statistically tractable models*, the final section presents some examples of research issues that often fall in the cracks between, on the one hand, academic research not usually focused on TTS, and, on the other hand, short-term TTS-focused research. To avoid making the scope of this paper too broad and leave some room for concrete detail, the role of symbolic-prosodic NLP components is not discussed; it is certain, however, that their role is as critical as their task is daunting.

2. EXAMPLES OF PROSODIC MODELING

This section gives a brief sketch of what is meant by prosodic modeling, by example. First, consider the case of segmental duration. Input consists of descriptor vectors such as, for the /ϵ/ in "descriptor" in this very sentence,

$$\mathbf{d} = <\!/\epsilon/, \text{unstressed, word-initial syllable, ...},>.$$

Each element in this vector is a *level* on a *factor*, such as segmental identity, word stress, and intra-word location. Thus, a segmental duration model maps a *factorial space* on acoustic quantities. An elementary way of doing this is to first map each level of each factor on a numerical value (e.g.: $S_2(\text{unstressed}) = 0.75$; the subscript "2" refers to stress being the second factor listed in the descriptor vector), and then combine these numerical values by multiplying them:

$$\text{Duration}(\mathbf{d}) = \prod_i S_i(d_i), \tag{1}$$

where d_i is the i-th element of descriptor vector **d**. Equation (1) is an example of a *prosodic model*; the term *prosodic component* refers to an implementation of an algorithm for computing a prosodic model.

Yet another example of a prosodic model is an application of Classification and Regression Trees (*CART*) to segmental duration modeling [13, 10]. As in the multiplicative model, input is formed by a factorial space. During the training phase, a tree is formed by successively dichotomizing the factors (e.g., the stress factor is split into {1-stressed, 2-stressed} vs. {unstressed}) to minimize the variance of the durations under the two newly formed subsets of the speech corpus. For each node of the tree, the observed average duration of the associated subset of the speech corpus is listed. During synthesis, the tree is searched to find a match between a node and the incoming descriptor vector, and the corresponding average duration is retrieved.

Much more complicated models exist. For example, in Möbius's application of the Fujisaki model [4, 11], input to the intonation component consists of a sequence of syllable labels, their associated prominence levels and durations, and phrase boundary times. The syllables are combined into *accent groups* (i.e., an accented syllable followed by zero or more unaccented syllables), and for each accent group a rectangular *accent command* is computed. An accent command maps time onto the log frequency axis, and is non-zero in a region roughly aligned with the accent group. The accent commands in a phrase are smoothed to form *accent curves*, and are then added to another mapping from time onto the log frequency axis, the *phrase curve*.

What these models share is that they map symbolic input vectors provided by NLP components onto acoustic quantities (duration, fundamental frequency or F_0), which are then used by the synthesis component to generate speech with the desired acoustic-prosodic characteristics.

But these models do not merely map one thing onto another, they do so with the claim that their output mimics human speech. In other words, these models are claimed to produce, for any input from the target domain of the TTS system, *including "new" input never seen before*, an output that is reasonably close to what a human would produce given the same input. Thus, these models are concerned with *prediction*, and it is a reasonable question to ask what empirical assumptions they are based on.

3. TARGET DOMAIN COVERAGE

This section discusses in more detail what it means for a system to encounter new input. The issue that the text used in some fashion during TTS system construction often covers only an infinitesimal subset of the text a TTS system may encounter in actual applications, applies to all TTS components. In fact, *the central challenge of TTS is handling new input* – otherwise one might just as well use recorded speech.[1]

3.1. Coverage defined in terms of units

3.1.1. Training text. During TTS construction, some finite amount of *training text* is used. The word "training text" is meant here in the broadest sense, and includes not only the speech or text corpora used in what has come to be known as "statistical approaches", but also, for example, the sum total of all data reported in the segmental duration literature and that forms the basis for manually generated segmental duration rules.

3.1.2. Unit class. The input domains of most TTS components can be described as sets of discrete *units* (these sets are called *unit classes*). For example, the accenting component, which is an NLP component that assigns pitch accents to words, processes vectors describing the lexical identity of the word in question, its location in the phrase, and parts-of-speech tags of surrounding words. The intonation component uses vectors describing features associated with a syllable, including its segmental makeup, lexical stress, and pitch accent type. Of course, different components, as well as different approaches to the same component, differ in which unit classes are used.

3.1.3. Target domain. Finally, there is the *target domain* of a TTS system. Most systems are not constructed with a particular target application in mind, so that their target domain is simply the totality of all possible text in the language. However, it may be of interest in certain cases to construct TTS systems with specific applications and hence target domains in mind.

Now, whether a given training text covers a target domain depends on the target domain and the unit class. For example, when the unit class is the diphone, then it is quite easy to construct text in which each conceivable unit occurs at least once (most languages have fewer than 2,000 diphones); this training text covers any target domain. Also, when the units consist of the parts-of-speech of two successive words, then there are at most a few hundred possibilities, so that complete coverage again is easy. Finally, when the target domain consists of a small number of carrier phrases (containing names and numbers as variable fields), and the symbolic prosodic constellation of a given carrier phrase (defined, e.g., in terms of phrase boundaries, three prominence levels, and two accent types) does not depend on what the fields contain, then the prosodic space is quite limited.

In each of these three examples, complete coverage of the target domain is possible because the unit classes are small. However, for many unit classes this is not the case. For example, at least 70,000 distinct triphones occur in English, which means that even when the target domain consists of carrier phrases, complete coverage of all names via triphonic concatenative units is difficult. So one important conclusion is that *even a highly restricted target domain may be unrestricted with respect to some unit class.* Note that for training text to cover a target domain it is necessary that all unit classes pertaining to all components are covered.

3.2. Coverage and component structure

It seems intuitively obvious that complete coverage is a good thing, but when is it really necessary? Whether complete coverage is necessary depends on the structure of the TTS component whose units are the focus. For example, for the acoustic inventory it is obviously critical to have training data (here: the recorded speech corpus) that cover all units. But when the focus is on input feature vectors to a segmental duration component, and the multiplicative model is used, then few training data should be needed because the total number of parameters [the $S_i(d_i)$'s] is small. Certainly no coverage is required of all feature vectors, because construction of the duration component con-

[1] Large differences in TTS system performance on new vs. old text are a major reason why prepared demonstrations are much more deceptive in TTS than in other speech technologies, such as speech coding.

sists entirely of estimating these parameters. This shows that one has to take into account the structure of the component.

What is it about a component that allows it to be constructed without complete coverage? The key issue is how a TTS component represents the input units, and what it "learns" during training. Roughly speaking, systems that operate on the basis of rules or equations have fewer problems with incomplete unit coverage than systems that have a *list-like structure*. Examples of the former include pronunciation rules, equation-based duration and intonation models, rule-based synthesis, and morphological decomposition rules. Examples of the latter include concatenative synthesis, and dictionary based pronunciation.

Abstractly speaking, both lists and rule systems consist of *empirical assertions*. In the case of rules, assertions are made such as "in American English /t/ is aspirated only when it occurs at the head of a stressed syllable and is not preceded by a voiceless fricative", or Equation (1), or the many assertions in rule-based synthesis that certain acoustic parameters behave in some way in a particular context. In lists, a pronunciation dictionary asserts for each word how it is pronounced, and a concatenative system asserts that a given phone sequence in some context can be synthesized at some level of quality with a particular concatenative unit.

The difference is that lists make far more of these assertions than rule systems. The problem raised by unit coverage is that the number of assertions in lists is in many cases too large for individual construction and verification of each assertion. For example, the number of distinct word forms allowable in a language is large (in fact, in languages with sufficiently productive word and name formation processes, it is practically unrestricted), too large for manual verification of each individual word's pronunciation.

The key advantage general rules have is not only that they require few training data (they have better *generalization properties*), but also that they can capitalize on existing knowledge (e.g., the large body of research papers on voiceless stop aspiration, morphological decomposition, and speech timing), and can thus go beyond the training data in ways not available to list-like approaches.

3.3. Coverage and frequency

Thus far, occurrence probabilities of units have been ignored. The key question to be answered for list-like components is what the probability is that a *unit token* randomly sampled from the target domain belongs to a *unit type* that was represented by at least one token in the training text. If it turns out that this probability is close to certainty (with, say, one failure per 1,000 sentences), then one indeed may not have to worry about the training text not covering all unit types.

Elsewhere [19], results from analyses were reported where – for various unit classes – the probability was measured that all units occurring in a sentence randomly selected from the target domain had types that occurred in the training set (this probability was called the *coverage index* for a given training text / unit class / target domain triple). The logic here is that it generally takes only one glitch in a synthesized utterance to seriously downgrade our impression of the TTS system. Hence our interest in measuring the probability that *all* goes well in an arbitrary sentence.

In a nutshell, it was found that all large unit classes (such as the vocabulary, descriptor vectors relevant for predicting prosody, and triphones) have the unfortunate property that the number of unit types with very low frequencies is large enough that their combined frequencies add up to near-certainty. Thus, training text must cover a large number of these types to assure an acceptable coverage index, because apparently in language *the unusual is the rule*.

It was also found that the relative frequencies of unit types are quite variable across different text genres. In fact, these frequencies often differ by orders of magnitude. Thus, one may meticulously select training text to have an optimal coverage index with respect to some target domain, but the frequencies on which this optimization is based may be unreliable, thereby undermining the meaningfulness of the optimization.

3.4. Summary

There are three ways to address the coverage issue for a given TTS component. First, constructing the component with rule-based methods. Second, using very large, carefully constructed training corpora with very high levels of the coverage index; but this has the risk of irrelevance if the target domain statistics are either unknown or cannot be measured reliably. And third, and not to be ignored, restricting the target domain.

4. SYNTHESIS AND PROSODY

In discussions of prosody in TTS, the synthesis component is often left out. However, there are close connections between prosodic modeling and synthesis. The first link is that prosodic component output is meaningful only to the extent that the synthesis component is capable of imposing it on output speech. The second link is that certain new approaches to synthesis raise the question whether prosodic modeling is even needed.

Roughly speaking, there are two types of synthesis – rule based and concatenative. In rule based synthesis (e.g., MITalk [1]), a phone sequence is mapped onto acoustic parameter trajectories, which then drive a formant synthesizer. The trajectories are computed based on rules about how, e.g., formants, behave in specific segmental contexts. In concatenative synthesis, stored intervals of digital natural speech (usually coded in some form, e.g., linear predictive coefficients) are glued together and

stretched/compressed and otherwise altered to satisfy the requirements set by the preceding acoustic-prosodic components.

4.1. Concatenative Synthesis: Ingredients

There are two key ingredients in the synthesis component of a concatenative TTS system:

4.1.1. Acoustic unit inventory. These are intervals in digitized speech recordings, labeled by one or more of the following pieces of information:

(1) Phone sequence. For example, an interval that starts in the center of an /n/ and ends in the following /o/ is labeled by $<$ n - o $>$.

(2) Prosodic descriptors. The $<$ n - o $>$ unit in stressed and utterance-medial context is labeled $<$ n - o , *stressed*, *medial* $>$.

(3) Surrounding phones. The $<$ n - o $>$ unit followed by consonants that have the effect of increasing values of F_2 in preceding vowels is labeled $<$ n - o , *high* F_2 $>$.

4.1.2. Concatenation operation. This operation involves at least three sub-operations:

(1) Attach successive acoustic units to each other. This may involve various forms of interpolation and smoothing, but may also be straight abutment, with an instantaneous change from the last byte of one interval to the first byte of the next interval.

(2) Stretch/compress acoustic units. The temporal structure (e.g., segmental durations) computed by the temporal component in general does not match the temporal structure of the stored intervals. Signal processing techniques are used to alter the structure of the latter.

(3) Impose intonation contour. The F_0 contour computed by the intonation component in general does not match the sequence of local F_0 contours in the stored intervals; again, signal processing techniques are needed for making the appropriate alterations.

4.2. Concatenative Synthesis: Examples

These concepts are illustrated with three examples that, while inspired by actually existing systems, are not intended to be significantly similar to any particular existing system.

Example 1: Basic diphone synthesis. In "basic diphone synthesis", speech intervals span the regions between midpoints of successive phonetic segment pairs. These intervals are labeled in terms of the identities of their phonetic segments, but no reference is made to prosodic context or surrounding phones.

It goes without saying that this scheme is extremely limited in which coarticulatory phenomena can be captured. This may result in audible spectral discontinuities, because some coarticulatory effects reach much farther than the second (first) half of the preceding (following) segment. For example, in a study where I measured the effects of postvocalic consonants on F_2 in preceding vowels at 40 ms from the start, restricted to vowel tokens with durations of at least 100 ms (average: 120 ms), the difference between /l/ vs. /k/ was more than 150 Hz. This explains why an $< s-I >$ unit excised from "six" and an $< I-l >$ unit cut from "million" often produce an audible spectral discontinuity when synthesizing "sill". Figure 1 shows the large differences in formant values between /I/ in "six" (137 tokens) vs. "million" (90 tokens).

Next, because the acoustic units are not prosodically annotated, this approach may require drastic alterations to be performed on the acoustic units to match the temporal and F_0 targets specified by the prosodic modules. As a rule, the more extreme the ratios are between the original F_0 and duration of the acoustic unit and the target F_0 and duration, the more trouble signal processing methods have avoiding audible distortions.

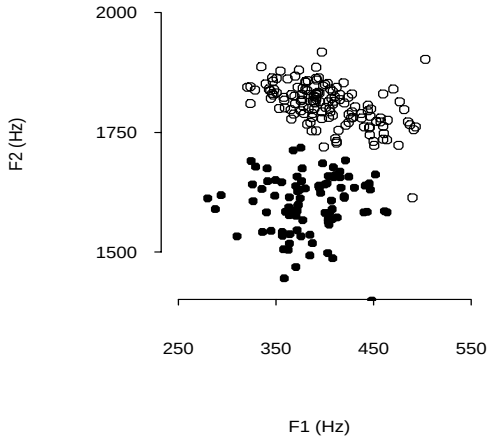

Figure 1: Formant values of /I/ at midpoints in "six" (open circles) vs. "million" (closed circles).

Example 2: Sophisticated *N*-phone synthesis. In this scheme, several improvements are made over the basic diphone scheme. First, to capture some of the longer-range forms of coarticulation, acoustic units can be longer than diphones, and some acoustic units are annotated in terms of their phonemic context – such as in the $<$ n - o , *high* F_2 $>$ unit. Second, considerable attention is paid to selecting the optimal *token* from a set of several recorded tokens of each acoustic unit *type*. Figure 1 explains why this is important, by showing how variable formant values can be even inside exactly the

same word (“six” and “million”) produced by the same speaker in prosodically highly similar sentences (numbers). Third, instead of cutting acoustic units at phonetic segment midpoints, they are cut to optimize two criteria: (1) maximal spectral proximity to spectral “ideal points”, and (2) shortness to make room for interpolation and enhanced smoothing between successive concatenated acoustic units. Fourth, acoustic units are prosodically annotated in terms of some factors known to have major spectral effects, such as utterance-finality.

This approach can be faulted for still requiring serious alterations of the acoustic units because the coverage of the prosodic space is necessarily limited by the low unit/utterance ratio of this approach. By this the following is meant. The scheme attempts to minimize spectral discrepancy by careful selection of acoustic units based on proximity to spectral ideals, and hence rejects many recorded units. It also uses a corpus made of carrier phrases that each contain at most three or four target acoustic units to guarantee evenness of the prosodic context. These two restrictions conspire to reduce the number of actually usable acoustic units per recorded carrier phrase. One has to take into account here that few speakers are capable of producing such carrier phrases with adequate levels of constancy, and that per day no more than at most a few thousand carrier phrases can be recorded.[2]

Example 3: Corpus based synthesis. In this proposal, there is no pre-excision, there are no alterations of pitch or timing during synthesis, and there are no acoustic-prosodic modules. Instead, at run time the original speech corpus is accessed directly. The corpus is annotated with phone boundaries, prosodic descriptors, and surrounding-phone descriptors. The corpus thus implicitly defines a set of acoustic units, corresponding to the (extremely large) set of all intervals that can be excised. For example, a ten-hour corpus defines more than half a billion acoustic units.[3]

At run time, the system attempts to find a sequence of multi-segment intervals in the corpus that simultaneously optimizes three criteria: (1) Phone labels must match the target phone sequence, and each individual segment must match some spectral “ideal” (measured, e.g., by the average of the same segments in similar contexts); (2) Prosodic labels must match the (symbolically specified, because there are no prosodic components in this proposal) target prosody; (3) Spectral match between successive acoustic units must be small, to avoid audible discontinuities; note that this includes pitch, because the concatenation operation is completely confined to abutting the units – F_0 or timing are not altered.

[2] Some discussions of acoustic inventory size focus on computer memory and disk space, which have indeed increased astronomically in recent years; but the real limits are set by the speaker.

[3] Ten hours contain 36,000 100-ms phones, which corresponds to 648,018,000 phone sequences or units.

With a limited speech corpus, these three criteria may clash severely: the system may face a choice between a phonemically correct but prosodically incorrect acoustic unit vs. a prosodically correct but spectrally discontinuous acoustic unit sequence. The central claim on which this proposal rests is that with a sufficiently large yet practically viable speech corpus, acoustic units can be found in the corpus that score sufficiently high on each of these three criteria that no audible discontinuities or outright phonemic/prosodic errors occur.

However, our analyses of coverage issues [19], strongly suggest that, although half a billion seems like a large number, the combinatorial possibilities of the language at large are so vast that the prosodically annotated phone sequences contained in even ten hours of speech are infinitesimal by comparison. I have not heard system demonstrations that disprove this belief.

It would seem, then, that the key challenge for corpus based synthesis is to study what type of domain restrictions are necessary to satisfy criteria (1)-(3) to acceptable degrees.

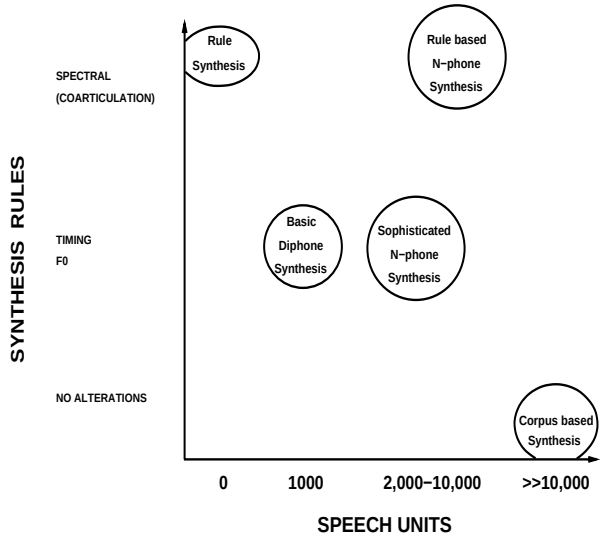

Figure 2: Concatenative synthesizers in Rules by Units space.

4.3. Concatenative Synthesis: Spanning the domain

A conclusion from the preceding section is that, certainly for unrestricted domain TTS, prosodic modeling is indeed necessary. However, besides settling this issue, concatenative synthesis was discussed in some detail because of the role it plays in setting limits on prosodic quality. Here, this role is discussed in more detail.

TTS systems cover input domains by combining two dimensions (Figure 2). One is the acoustic inventory, the other is the set of alterations performed by the prosodic and synthesis modules to cover the range of (phonemic and prosodic) contexts in which acoustic units can occur. A TTS system can have few (basic diphone system), many

(sophisticated system), or extremely many (corpus based system) acoustic units, and perform little (corpus based system), intrusive (sophisticated system), or highly intrusive (basic diphone system) alterations. For completeness, the Figure includes pure rule synthesis as a special case, where the TTS input domain is covered by altering an empty acoustic inventory.

Regardless of the deep differences between the three hypothetical TTS systems, they share a key assumption about natural speech: *the range of contexts in which a given acoustic unit can occur in the target domain only alters the acoustic unit's temporal structure, pitch, and amplitude*. This is so for the simple reason that this is all that concatenation algorithms currently do. This assumption is called the *concatenative assumption*. Note that the concatenative assumption is an assumption about natural speech, and that its truth depends on the specifics of the synthesis system such as the set of contexts in which an acoustic unit can occur, how richly the acoustic units are labeled, and how restricted the target domain of the TTS system is.

At first glance, it seems obvious that the concatenative assumption must be wrong for the sophisticated and the basic schemes. Speech production involves extremely rapid actions by hundreds of muscles whose coordination cannot possibly be that tight that no room is left for some degree of independence. It is a priori likely that this articulatory independence leads to de-coupling of acoustic variables, so that the changes that a given acoustic unit undergoes in different contexts cannot me mimicked merely by stretching and compressing it – the changes are *spectral*.

There is little doubt that many long-range coarticulatory phenomena exist with measurable acoustic consequences that go well beyond affecting only timing and pitch. Examples include anticipatory lip rounding due to a rounded vowel that may affect the acoustics of a schwa as far back as one or two syllables, and mutual effects between vowels separated by voiced /h/.

Likewise, prosodic factors have spectral effects that go beyond pitch, timing and amplitude. For example, stress has effects on spectral tilt [14], and utterance boundaries have effects on various aspects of the glottal wave form [12].

If all these prosodic and long-range coarticulatory spectral effects have to be captured via special-purpose acoustic units, then the size of the acoustic inventory may have to increase by orders of magnitude (e.g., merely annotating acoustic units in terms of two rounding levels, three stress levels, and two location levels, already would increase the acoustic unit inventory by one order of magnitude). That renders the sophisticated concatenative synthesis scheme impractical, because of the low unit/utterance ratio.

4.4. Rule based N-phone synthesis

Synthesis limits prosodic quality by the extent to which it has the ability to mimic prosodic spectral effects. If no spectral alterations can be made, then a very large acoustic inventory may be needed to span the prosodic space. The three criteria discussed in the context of corpus-based synthesis are obviously relevant for any synthesis scheme; thus, any attempts to increase prosodic coverage while keeping the acoustic inventory at a practical size will only result in increased spectral discontinuities or phone label mismatches. I conjecture that ultimately the only options will be either to use corpus based synthesis for appropriately restricted target domains (if non-trivial domain restrictions can indeed be found, without intruding on the territory of stored speech), or to develop new signal processing methods for altering speech units to mimic these spectral prosodic and coarticulatory effects – "rule based N-phone synthesis" (Figure 2).

When acoustic units can be spectrally altered during synthesis, this has the added benefit that it may also be possible to reduce spectral discontinuities between units, so that the token selection process can be less selective, resulting in being able to afford larger acoustic inventories while keeping the amount of recording constant. This is why in the Figure, "rule based N-phone synthesis" is slightly to the right of "sophisticated N-phone synthesis". But it is hard to overestimate the difficulties that must be overcome here, given the problems that current signal processing methods already have with the comparatively simple F_0 and timing alterations.

5. CONTENT-SPECIFIC MODELING

The discussion now turns to prosodic modeling proper. In this section, combinatorial and statistical arguments are presented for why a particular type of model is required – a type called "content-specific". The next section discusses specific empirical issues.

5.1. Sparsity and generalization

The descriptor vectors that form the input domain to prosodic modules have to be rich to capture all aspects of text affecting timing and intonation, and there is little doubt that as a result the input domains are extremely large [19]. For example, segmental duration depends on at least seven factors: segmental identity, some characterization of the identities of the surrounding segments, word stress, accent, position in the syllable, position of the syllable in the word, and position of the word in the phrase and utterance [7, 16]. For intonation, at least that many factors are relevant.

As a result, coverage indices of even large training corpora are very low, which means that for prosodic modeling to work models are needed with strong generalization properties, while list-like approaches must be avoided altogether.

5.2. Interactions

As pointed out in Section 3.2, even if many factors play a role, this does not necessarily make generalization impossible. This is illustrated by the multiplicative model in Eq. (1), in which only a few parameters have to be estimated; their number is roughly equal to the sum of the numbers of levels on each factor. Even with as many as 12 factors each having 10 levels, this number (120) is still a very small fraction of the number of descriptor vectors that can occur in the language because the latter number is related to the product of the numbers of levels on each factor. As a result, training data can be used with a very low coverage index.

Unfortunately, the multiplicative model is at best an extremely coarse description of how factors jointly affect duration. For example, in English, the proportional effects of postvocalic voicing on preceding vowel duration are amplified dramatically by phrase boundaries [5, 16]. And also in the case of intonation, there is little doubt that such aspects as F_0 peak height or the timing of F_0 rises depends in very complicated ways on factors such as sentence mode, pitch accent type, and the local segmental context.

This poses a problem: Good generalization requires models with few parameters, but realism requires complicated models. What is needed are models that are *elaborate enough to capture key aspects of inherently complicated phenomena, yet use few parameters.* It seems obvious that such models have to incorporate content-specific knowledge in the way they are structured. Somewhat harder arguments for this conjecture are provided now, by contrasting a general-purpose statistical technique (CART) with a content based model.

5.3. Content-specific modeling

Trees generated by CART often look quite elaborate. The question is whether they have the right type of elaborateness. Maghbouleh [10] trained two segmental duration models, and then tested their generalizability on new data, both "somewhat new" data taken from portions held out from the training corpus and "very new" data taken from entirely different corpora. Generalizability was measured by the correlation between observed and predicted segmental durations. One model was CART, the other model a "sums-of-product model" developed by the author [17]. Results were straightforward: a training corpus of a few hundred data points was sufficient for the sums-of-product model to reach an asymptote at generalizability levels higher than reached by CART, even after training the latter on as many as 10,000 data points. Moreover, the difference was more pronounced for the "very new" data than for the "somewhat new".

What could explain this? According to sums-of-products models, the duration for a phoneme/context combination described by the feature vector $\mathbf{d}$ is given by:

$$\mathrm{DUR}(\mathbf{d}) \;=\; \sum_{i \in K} \prod_{j \in I_i} S_{i,j}(d_j). \qquad (2)$$

Here, K is a set of indices, each corresponding to a *product term*. I_i is the set of indices of factors occurring in the i-th product term. Product terms can contain just one parameter. To illustrate, consider the well-known duration model by Klatt [1], according to which:

$$\mathrm{DUR}(v,c,p) \;=\; S_{1,1}(v)S_{1,2}(c)S_{1,3}(p) + S_{2,1}(v) \quad (3)$$

Here, v denotes the vowel identity factor, c the class of the postvocalic consonant (*voiced* vs. *voiceless*) and p the phrasal position factor (*phrase-medial* vs. *phrase-final*). In the usual formulation, $S_{2,1}(v)$ is the minimum duration of vowel v, $S_{1,1}(v)$ is the *net duration* (defined as the difference between the inherent duration and the minimum duration), $S_{1,2}(c) = K_c$, and $S_{1,3}(p) = K_p$; the latter two are constants tied to the postvocalic consonant and to phrasal position. This model has two product terms, with index sets I_1 = {1,2,3} and I_2 = {1}; K = {1,2}. The multiplicative model has one product term with index set I_1 = {1,2,3}, K = {1}. The additive model has three terms with index sets I_1 = {1}, I_2 = {2}, and I_3 = {3}; K = {1,2,3}.

Thus, sums-of-products models subsume several existing segmental duration models, which suggests that this class formalizes some important idea. This idea is in all likelihood the following. Sums-of-products models capture an important phenomenon often observed in segmental duration: *directional invariance*. This refers to the property that, holding all else constant, the effects of a factor have always the same direction. To illustrate, in English, when the durations of the vowels /i:/ and /ϵ/ are compared in two otherwise identical contexts, one can be certain that with enough data the average duration of /i:/ will be longer than that of /ϵ/. The same holds true for effects of word stress, voicing of the postvocalic constant, and intra-utterance position.

When all parameters in a sums-of-product model are positive, then the model automatically predicts directional invariance. In other words, directional invariance is not an accidental feature, it is a "structural" property of the model class. Yet, these models can obviously capture interactions such as the amplification of the effects of postvocalic voicing that takes place at utterance boundaries.

Directional invariance is not an intrinsic property of CART. Also, CART has no known mechanism for abstracting from training data general principles such as directional invariance. I strongly suspect that capitalizing on the broad empirical property of directional invariance is a major factor explaining Magbouleh's results.

The sums-of-product model exemplifies a model that

manages to capture fairly intricate interaction phenomena but still has a small number of parameters. The most important point here is that the structure of the model is based on *content-specific empirical considerations.* That is, both the exact choice of factors and the directional invariance property of sums-of-product models are based on our understanding of segmental duration.

To further illustrate the concept of content-specificity, suppose that one encodes intra-word position by *syllable number* (first syllable in "syllable" has position 1, second 2, third 3) and the *number of syllables*, so that the first syllable in "syllable" is encoded as (1,3), the second as (2,3) and the third as (3,3). This seems like a reasonable coding scheme, *but it is not directionally invariant*: holding word stress and other factors constant, the second syllable is longer than the first syllable in 2-syllable words (because it is word-final; see below), but shorter in 3-syllable words (because it is neither word-initial nor word-final). An alternative scheme would be *word-initiality* and *word-finality* [resulting in coding these syllables as (0,1), (1,1), and (1,0)]. It is well-established that final syllables are longer than penultimate syllables (both stressed or both unstressed) in words with three or more syllables [17], perhaps because many word boundaries coincide with some type of syntactic boundary [6]. Being word-initial has also a lengthening effect (comparing syllables in first vs. second position in words having at least three syllables [17]). I do not know of any cases where either factor is reversed by some other factor, and see no obvious reasons why this would occur.

Sums-of-products models and selection of directionally invariant factorial coding schemes are content-specific, but only in a minimalist sense, because nothing is being said about underlying processes and explanations of the phenomena; all one has is a well-argued surface representation of the data. Yet, these models make the point that when content-specific knowledge is incorporated in a model, significant gains can be made in the ability to generalize from training data to new data.

Modeling F_0 contours is at least as complicated as modeling segmental duration, because it requires shaping an entire curve and because more factors are involved. This implies that, even more than in segmental duration modeling, content-specific modeling is required. Although some work has been done with content-less modeling (e.g., [15]), there actually are many content-specific models here. However, the primary problem seems to be that, most likely reflecting the complexity of the phenomena, the models are still a long distance away from having a tight, quantitative connection with data. For example, models by Fujisaki [4] and by the author [21, 22] are able to accurately predict some quantitative aspects of F_0 contours, but are not really able to make accurate predictions for complete F_0 contours for new, unrestricted text. Other intonation models describe data at a more qualitative level, and hence are not able to provide the level of quantitative detail required for TTS.

6. OPEN ISSUES IN PROSODY RESEARCH

There are difficult research issues that must be resolved in order to improve TTS naturalness, yet are ignored. This is probably because, on the one hand, they are too long-term and complicated for development organizations, and, on the other hand, too focused on TTS and too model-dependent to be of interest to the academic community. This section gives some examples of these issues. The examples were chosen primarily because our TTS group happened to stumble into them, and not because they are in any way representative for this class of issues in general.

6.1. Concatenation

Section 4 already mentioned the challenges that have to be addressed in signal processing if one is to consider making spectral alterations in concatenative units.

Before turning to issues directly relevant for prosodic modeling, one more issue concerning concatenation should be mentioned. The measurement of spectral discrepancies and discontinuities, which is essential for concatenative unit selection and excision methods, is currently done in ways that ignore perception entirely. For example, some methods use cepstral distance and others use formants, but there is no evidence that these representations capture all aspects that perceptually matter, let alone that the particular ways of assigning weights to dimensions in the distance measures make any sense. Thus, perceptual studies are needed to fill in these gaps. These studies can be done at the surface level (e.g., with some psychophysical weight optimization procedure), or – which is our suspicion – may have to dig deeply into auditory processing.

6.2. Timing

Most TTS systems control timing via segmental duration. While conceptually and computationally convenient, it is far from clear whether segmental duration captures how the temporal structure of speech is affected by various factors. This section addresses issues concerning the temporal unit.

Sub-segmental timing In a study on the effects of postvocalic voicing on the spectral time course of vowels [20], a time warping technique adapted from [9] was used that maps spectral frames in, say, the sonorant region of "melt" on the frames in the sonorant region of "meld".

When the spectral frames are plotted in formant space, the formant trajectories of "melt" and "meld" appeared to travel on a common path, but traversed this path with different timing. Moreover, the temporal difference was non-uniform, with the latter half of /ϵ/ and the /l/ stretched out far more than the first half of /ϵ/ and the /m/. This has two implications. First, this *path invariance* under changes of post-vocalic voicing lends some credibility to the as-

sumptions underlying concatenative synthesis: at least the effects of postvocalic voicing on /m-e-l/ units can be handled purely by temporal, non-spectral alterations.

Second, the non-uniformity of the effects of postvocalic voicing shows that controlling timing via segmental duration is inadequate. It seems likely that the same holds true for the effects of other factors. Hence, unless there are convincing reasons for the perceptual irrelevance of these phenomena, a time-warp based approach to timing control is needed.

But this raises questions about the joint effects of multiple factors. Suppose that for some prosodic factors and some acoustic unit the path invariance assumption is accurate, so that it makes sense to generate appropriately warped versions of this unit for all contextual constellations defined by these factors. In the case of segmental duration modeling via the multiplicative model, the effects of multiple factors were combined simply by multiplying them. However, in a time warp based approach it is unclear how to combine the per-factor warps. For example, if (1) the stored unit is from a stressed, utterance-medial context, (2) the stress warp mostly elongates the center of a vowel, and (3) the utterance-finality warp mostly elongates the final part, what does the warp required for the unstressed/utterance-final context look like? How does it relate to the two per-factor warps?

Timing of supra-segmental units The idea that units larger than the phonetic segment play a role in timing has been around for a long time and has come in many forms, including the isochrony hypothesis (which holds that speakers attempt to keep durations of stress feet as constant as they can) and the syllabic timing hypothesis (which holds that durations of syllables in some sense are "computed" prior to segmental durations).

It is important to draw a distinction between phonological entities vs. durations of speech intervals. There is no doubt that stress feet and syllables play important roles as phonological entities in accounting for various prosodic phenomena. For example, consonant duration is critically affected by intra-syllabic location (onset vs. coda vs. ambisyllabic), and pitch accent curves are often aligned with feet – not words or individual syllables [21, 22].

But the evidence for their importance as time intervals is unconvincing. For example, I measured durations of syllables in a fixed prosodic context (e.g., first syllable in polysyllabic word in utterance-medial position, etc.), and found that these durations could be predicted quite accurately from the intrinsic durations of their constituent segments as estimated from other parts of the speech corpus. E.g., the first syllable in "sitting" is 65 ms longer than the first syllable in "knitting", almost exactly matching the difference in average duration between /s/ and /n/ measured in other contexts (61 ms). Thus, far from segmental durations being computed as an afterthought after syllable durations have been established, it appears that the duration of a syllable depends on the exact segments it contains and their intrinsic durations.

In an analysis of stress foot intervals, no effects were found of stress foot length (measured by the number of syllables) on segmental duration, once the effects of word boundaries were taken into account with partial correlation methods [16].

These speech interval hypotheses can be viewed as attempts to understand the role played by supra-segmental phonological units. If it turns out that these hypotheses are indeed flawed, then the question still has to be answered of why a segment being part of some larger phonological unit appears to affect (sub-)segmental timing. Are segments in longer words compressed because they are more redundant than the same segment in shorter words? If so, perhaps discourse structure or word frequencies have to be taken into account in new ways. Or are these effects all boundary phenomena, where some type of non-zero boundary occurs after most words and where these effects spread leftward inside words? If so, our concept of boundary has to be broadened and made more precise.

6.3. Intonation

Even more so than in timing, in F_0 modeling many issues are the focus of intense controversy, including how to describe a pitch contour (as a sequence of tones vs. superposition of underlying curves). There is little doubt that for improved intonation in TTS, this issue as well as many others the current phonetics literature worries about must be addressed. However, there are also some issues that, because they involve quantitative detail, at first glance do not appear to be of sweeping theoretical import. Yet, they have to be resolved for improving TTS prosody. One such issue is discussed next.

When are two pitch accent curves the same? The precise timing of local pitch excursions ("accent curves") associated with accented syllables can have important effects on how listeners interpret an utterance. In a well-known series of perceptual experiments, Kohler [8] showed that relatively small (100 ms) changes in location of F_0 peaks or rises relative to segmental anchors (such as stressed syllable start or vowel start) are not only audible, but in fact alter the intentional meaning of the utterance. Similar results were obtained by d'Imperio and House [3]. Yet, even when sentence context and meaning are kept completely the same, merely changing the syllables with which the local pitch excursion is associated (e.g., "Now I know Sheila" vs. "Now I know Mitch", with single pitch accents on the final words) can shift peaks by at least 150 ms [21].

For TTS, this means that an *alignment model* is needed that describes how accent curves *within* a given phonological/perceptual equivalence class vary with the structure of syllables, and how curves in distinct equivalence classes

differ. The model currently used by the Bell Labs system is based on non-linearly time-warped accent curve templates within a superpositional framework [21]. However, I have only started exploring how the model can account for the myriad of intonational phenomena that have been documented. In addition, the model raises many new issues (or re-phrases old issues), such as quantifying stress clash, interactions between tones in Mandarin Chinese, the behavior of plateau-like F_0 regions, and whether continuation rises involve one or two quasi-independent gestures.

7. SUMMARY

This paper set out by asking what can be done to improve the prosodic quality of TTS systems, and tried to answer this question from different angles.

First, synthesis components play an important role in setting limits on prosodic quality. Overcoming these limits involves some very serious obstacles. Corpus based approaches may appear to avoid these obstacles, but, for combinatorial reasons, most likely can only work for limited target domains. For larger domains, new signal processing techniques must be developed that alter stored speech units much more drastically than is currently the case, *and* do so while causing less audible distortions than current methods.

Second, improving prosody requires a special type of modeling, which counters the combinatorial complexity of a language with content-specific models that are mathematically tractable and have good statistical properties. Some examples from segmental duration modeling were shown to demonstrate what is meant here.

Third, there are many difficult empirical questions that must be resolved, but the TTS community is not large enough to make quick progress with these problems. These questions, while quite exciting for researchers directly involved with TTS, may not appear that interesting from a wider phonetics and linguistics standpoint because they seem specific and overly concerned with quantitative detail. Generating broader interest in these types of questions is an important challenge for the TTS community.

8. ACKNOWLEDGMENTS

The ideas in this paper were inspired by discussions with members of the Bell Labs TTS team, in particular Bernd Möbius, Joseph Olive, Chilin Shih, and Richard Sproat. I also thank Joseph Olive and Bernd Möbius for editorial comments. Of course, I take complete responsibility for the speculative remarks made in this paper.

9. REFERENCES

1. J. Allen, S. Hunnicut, and D.H. Klatt. *From text to speech: The MITalk System*. Cambridge University Press, Cambridge, U.K., 1987.
2. M. d'Imperio and D. House. Perception of questions and statements in Neapolitan Italian. In *Proceedings ESCA Tutorial and Research Workshop on Intonation: Theory, Models and Applications*, Athens, September 1997.
3. H. Fujisaki. Dynamic characteristics of voice fundamental frequency in speech and singing. In Peter F. MacNeilage, editor, *The production of speech*, pages 39–55. Springer, New York, 1983.
4. D.H. Klatt. Interaction between two factors that influence vowel duration. *Journal of the Acoustical Society of America*, 54:1102–1104, 1973.
5. D.H. Klatt. Vowel lengthening is syntactically determined in connected discourse. *Journal of Phonetics*, 3:129–140, 1975.
6. D.H. Klatt. Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America*, 82(3):737–793, 1987.
7. K.J. Kohler. Macro and micro F0 in the synthesis of intonation. In J. Kingston and M.E. Beckman, editors, *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, pages 115–138. Cambridge: Cambridge University Press, 1990.
8. M.J. Macchi. Using dynamic time warping to formulate duration rules for speech synthesis. *Journal of the Acoustical Society of America*, 85:S1(U49), 1989.
9. A. Maghbouleh. An empirical comparison of automatic decision tree and hand-configured linear models for vowel durations. In *Proc. of SIGPHON-96*, pages 1–7, Santa Cruz, 1996.
10. B. Möbius, M. Pätzold, and W. Hess. Analysis and synthesis of German F0 contours by means of Fujisaki's model. *Speech Communication*, 13, 1993.
11. G. Richard and C.R. d'Allessandro. Modification of the aperiodic component of speech signals for synthesis. In J.P.H. van Santen, R.W. Sproat, J. Olive, and J. Hirschberg, editors, *Progress in speech synthesis*, pages 41–56. Springer, 1996.
12. M.D. Riley. Tree-based modeling for speech synthesis. In G. Bailly and C. Benoit, editors, *Talking Machines: Theories, Models, and Designs*, pages 265–273. Elsevier, 1992.
13. Agaath Sluijter. *Phonetic correlates of stress and accent*. Holland Institute of Generative Linguistics, 1995.
14. C. Traber. F0 generation with a database of natural F0 patterns and with a neural network. In G. Bailly and C. Benoit, editors, *Talking Machines: Theories, Models, and Designs*. North Holland, Amsterdam, 1992.
15. J. P. H. van Santen. Contextual effects on vowel duration. *Speech Communication*, 11:513–546, 1992.
16. J. P. H. van Santen. Analyzing N-way tables with sums-of-products models. *Journal of Mathematical Psychology*, 37(3):327–371, 1993.
17. J. P. H. van Santen. Combinatorial issues in text-to-speech synthesis. In *Proceedings of Eurospeech-97*, Rhodes, September 1997.
18. J. P. H. 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*, 92(4, Pt. 2):2444, October 1992.
19. J.P.H. van Santen and B. Möbius. Modeling pitch accent curves. In *Proceedings ESCA Tutorial and Research Workshop on Intonation: Theory, Models and Applications*, Athens, September 1997.
20. J.P.H. van Santen, C. Shih, and B. Möbius. Intonation. In R.W. Sproat, editor, *Multilingual Text-to-Speech Synthesis*. Kluwer, Dordrecht, the Netherlands, 1997.