Probabilistic Model of Acoustic/ Prosody/ Concept Relationships for Speech Synthesis

Nanette M. Veilleux
Boston University

Abstract

This paper describes the formalism for incorporating emerging linguistic theory in a joint model of the acoustic/prosody/concept relationships. It makes use of binary decision trees to estimate model parameters, the conditional probabilities. In doing so, the model remains general, and can accommodate the results of our evolving understanding of the interaction between factors that determine prosody. While this model has been successful in both speech synthesis and analysis applications, it has made use of syntactic and pragmatic information alone. Extension of this model to map prosodic structure to other higher order linguistic structures that more fully describe the meaning that an utterance is straightforward. As hypotheses are developed in the ranking of competing constraints, including focus structure, and in the role of discourse history, they can be integrated into the model as features in the binary decision tree.

1 Introduction

Prosody, particularly the placement of phrasing and relatively prominent syllables within an utterance, is important in human understanding of speech. (Boogaartand Silverman, 1992; Price et.al., 1991) While great improvements in the prosody of synthetic speech have been made over the past decade, naturalness itself has proved elusive (Boogaartand Silverman, 1992; Veilleux, 1994). One reason for the remaining differences between synthetic and human speech is an incomplete understanding of the mapping between the speaker's intended meaning and that meaning's acoustic consequences, which are, in part, encoded in the prosodic structure (Price et.al., 1991). Prosody, therefore, is an important part of the route from meaning to speech. In order to see how prosody can be improved in automatic speech synthesis systems, it is useful to examine what is known about the relationship between prosody and the acoustic speech signal on the one hand and between prosody and the meaning embedded in that speech on the other.

Clearly, prosody is related to the acoustic speech signal. In human speech, prosodic phrases and prominence are cued by acoustic features such as f0 contour and duration. For example, final syllable lengthening and a descending f0 pattern on the word *think* (as well as *know*) will lead the listener to the perceive a phrase break between *know* and *I* in the underlined utterance:

> Don't you think Michael Jordan is great?
> I don't think; I know.

Furthermore, this sentence might reasonably be produced with a pitch accent (e.g. an H*, or rise/fall f0 pattern) on the words *think* and *know*, lending the perception that these two words are more prominent than other words in the utterance and that they are being contrasted (Prevost, 1996).

Prosody is also related to higher order linguistic structures such as the syntax of an utterance. In the above example, the main clauses [S I don't think] and [S I know] align with major prosodic phrase breaks. Other researchers, as well as myself, have gainfully used this prosody/syntax relationship in both speech synthesis and speech analysis (automatic recognition and understanding) applications. (e.g. (Veilleux, 1996; Wang and Hirschberg, 1992)). Also note that the same word string could be used to convey the opposite meaning, as in

> What's the capital of Sri Lanka?
> I don't think I know.

However, this sentence does not have the same syntactic structure and one would not expect the same prosodic structure (e.g. there would be less of a "break" after *think* in the second example).

However, just as syntax is not fully determined by word choice or order, syntax is not the only factor that determines the prosodic structure. For example (from (Steedman, 1991)), the sentence

[S [NP Mary] [VP prefers [NP corduroy]]]

can be naturally produced with a major phrase break bisecting the verb phrase:

(Mary prefers) (corduroy).

It is reasonable to conclude that prosody is constrained by factors in addition to (and possibly in conflict with) syntax. The following example shows that semantic issues also play a role in determining prosody.

Did she come with Bill's wife?
* No, she came with BILL's sister.

Bill is inappropriately produced with prosodic prominence (denoted by the use of capital letters) given the intention of the speaker to directly answer the question. Notice that the following pair is perceived as appropriate:

Did she come with John's sister?
No, she came with BILL's sister.

Considerations of this kind suggest that prosody is therefore linked to the syntax and semantics of an utterance, which in turn are related to the speaker's intentions, or the concept that the speaker wishes to convey. Collectively, syntax and semantic structure will be referred to here as the information structure of an utterance. However, we purposely leave *information structure* as a loosely defined term that can be expanded as our understanding of the relationship between prosody and the meaning embedded in speech evolves.

Therefore, we find that prosody is related to the information structure on one hand and to the acoustic signal on the other. Consequently, prosody can be useful as a bridge, or intermediary representation between concept and speech.

This work presents a general methodology for deriving a formalism to describe the acoustic/ prosody/ concept relationships for use in automatic spoken language systems, by generating a mapping between the acoustics of speech on one hand, and a syntactic/semantic representation of the speaker's intentions on the other. This computational mapping serves to create an acoustic/ prosody/ concept model. Most recent applications of this model have been in the area of automatic speech recognition where the acoustic/ prosody/ syntax mapping[1] was used to decrease word error (Veilleux, 1996). However, the mapping is bi-directional, and this model can be used for speech synthesis as well as for speech analysis (recognition and understanding). The method for doing so will be presented after the model is described in more detail.

The consequences of the analysis of the relationships between information structure, prosody and speech described above show that the use of prosody in concept-to-speech synthesis (CTS) can not be achieved merely as plug-in extension of prosody models that have been developed for text-to-speech (TTS) systems (e.g. (Veilleux, 1994; Wang and Hirschberg, 1992)). Although prosody models could be applied after the generation of the word string, as in TTS applications, prosodic structure must be determined not only by the word string, but also by the meaning-specific syntax and semantic structure. The lack of mapping between this information structure and prosody is partly responsible for the persistent unnaturalness in synthetic speech.

Clearly, there are many aspects of the factors that determine prosodic structure that are still to be understood. While the acoustic correlates of prosody(Pierrehumbert80, 1980; t'Hart et.al., 1990) and the relationships between prosody and syntax (e.g. (Selkirk, 1984; BachenkoFitzpatrick, 1990; Ostendorf and Veilleux, 1993; Gee and Grosjean, 1983; Wang and Hirschberg, 1992; Terken and Hirshberg, 1994)), have been investigated in some length, the mapping and interactions between these domains has not been completely quantified. Moreover, the effect of other factors such as focus and discourse structure on prosody, have not been studied as extensively. A strength of the model presented here is that it is general, and therefore adaptable as our understanding of the prosody/ concept and prosody/speech relationships improves. This generality in the formalism will be pointed out as the model is described.

One more general comment about this model is in order before we turn to the details. The formalism presented here is probabilistic. It is data-driven (labeled training data is used to derive model parameters, i.e. the probabilities). It therefore has

[1]The concept domain was primarily represented by syntax, because syntactic structure was readily available and its role in constraining prosodic structure is better understood.

the strengths of corpus-driven approaches in that it is informed by a large body of example. It also has the characteristic weakness of probabilistic models in that it tends to generalize unless otherwise directed (e.g. by manipulating misclassification costs, or providing constraints to distinguish different events.) One other feature of the model presented here is that it explicitly incorporates linguistic knowledge in the design of decision trees used as probability estimators. This serves to cluster data according to linguistic context, decreasing the bias towards generalizing the most prevalent.

2 Probabilistic Mapping of Acoustic/Prosody/Concept Relationships

2.1 Model Formalism

If the task of automatic speech synthesis can be framed as that of selecting the most probable acoustic production associated with a text string annotated for intended meaning, it can be achieved by finding

$$argmax_{\mathbf{x}} p(\mathbf{x}|meaning). \quad (1)$$

That is, by finding the acoustic sequence $\mathbf{x}$ that maximizes the joint probability of the acoustic sequence given the meaning or concept the speaker wishes to convey. For this work, the word string itself is assumed to have been generated in a prior step. Therefore, the acoustic sequence would be a sequence of prosody related supra-segmental features. The term *meaning* is quite open to interpretation. For the purposes of this work, *meaning* will be assumed to be straightforwardly encoded in the information structure, that is, the syntax and semantic structure of the utterance. For the present, information structure (denoted by $\mathcal{I}$ in the equations below) represents some aspects of the underlying concept that are covered by theory (such as syntax) or description (such as an instantiation of a focussed constituent). This structure will be represented as annotations on the text string, which can serve as input to a speech synthesis system. What features, beyond syntax, are relevant and can be reliably annotated is, of course, an open research question, that will probably be answered over time. Some suggestions for features and a method for incorporating them in this probabilistic model are described in Section 3. So, returning to the model derivation, we see that we wish to find

$$argmax_{\mathbf{x}} p(\mathbf{x}|\mathcal{I}), \quad (2)$$

that is, the acoustic sequence $\mathbf{x}$ that maximizes the probability of the acoustics given the informational annotation. Inserting prosody as an intermediary representation, we can re-write this conditional probability of the acoustics given the information structure in terms of the sequence of abstract prosodic labels $\mathbf{a}$.

$$p(\mathbf{x}|\mathcal{I}) = \sum_{\mathbf{a}} p(\mathbf{x}, \mathbf{a}|\mathcal{I}) \quad (3)$$

This sequence of abstract prosodic labels describe an ordered set of prosodic events, such as prominence and phrasing.

In order to capture these prosodic events in a computational model, this work uses prosodic labels based on the ToBI transcription convention (Tones and Break Indices, see (Silverman et.al., 1992)). The ToBI system labels prosodic prominence with a pitch accent type (Tone), from a subset of Pierrehumbert's inventory (Pierrehumbert80, 1980). In previous work, data with this level of detail was not available so the simple label of $\pm$ prominence on each syllable was used. Prosodic phrasing is captured by placing a break index (BI: 0 to 4) at each word juncture, to indicate the level of de-coupling between the words. For example, the juncture between two words in a clitic group would be labeled with a 0 break index. At the other end of the spectrum, the junction between two words separated by a major prosodic phrase would be labeled with a 4 break index. Moreover, the ToBI system allows one to label prosodic events that are conspicuous in spontaneous speech. For example, significant lengthening of a final syllable, without the comcommittant intonational cues associated with prosodic phrasing can be labeled with a diacritic. If these events serve a communicative purpose in informal speech (marking focus, holding the floor) then future models may make use of these labels.

Applying Bayes' Law to Equation 3:

$$\mathrm{p}(\mathbf{x}|\mathcal{I}) = \sum_{\mathbf{a}} \mathrm{p}(\mathbf{x}/\mathbf{a})\mathrm{p}(\mathbf{a}/\mathcal{I}) \quad (4)$$

This form of the equation more clearly reflects the use of prosody as an intermediate representation, relating the acoustic sequence to the information structure by modeling $\mathrm{p}(\mathbf{x}|\mathcal{I})$ in terms of the conditional probability of acoustic sequence given the prosodic structure ($\mathrm{p}(\mathbf{x}|\mathbf{a})$) and the conditional probability of prosody given the information structure ($\mathrm{p}(\mathbf{a}|\mathcal{I})$). The model parameters can be estimated using statistical methods: in this work, using an automatically derived binary decision trees.

In speech synthesis applications, each prosodic event can be realized by manipulating the acoustic signal according to a set of context-based f0

and duration rules (vanSanten, 1993). For example, a syllable labeled with a high pitch accent (H* or simply +prominence from this work), would be given a rise/fall f0 contour, adjusted according to e.g. the duration of the vowel, etc. In such applications, $p(\mathbf{x}/\mathbf{a})$, the probability of the acoustic parameters given the prosodic labels, is fully determined by these rules alone. Although a stochastic model of prosody and acoustic features could be useful in more specifically determining the correlates of prosody in f0, duration and vowel quality, it is left to future work to incorporate a probabilistic acoustic model of prosody in the synthesis algorithm. Therefore, with this simplification, the model equation becomes:

$$argmax_{\mathbf{x}} p(\mathbf{x}|\mathcal{I}) = argmax_{\mathbf{a}} p(\mathbf{a}|\mathcal{I}) \quad (5)$$

$$p(\mathbf{a}|\mathcal{I}) = \prod_{i=1}^{n} p(a_i/\mathcal{I}) \quad (6)$$

assuming prosodic labels are independent.
Or,

$$= p(a_1|\mathcal{I}) \prod_{i=2}^{n} p(a_i/a_1 ... a_{i-1}, \mathcal{I}) \quad (7)$$

assuming a Markov dependency.

Here $[a_1 ... a_n]$ is the sequence of n prosodic labels in the utterance. The task remains, therefore, to find the sequence of prosodic labels with the highest probability given the underlying concept to be conveyed $\mathcal{I}$.

2.2 Decoding Algorithms

Several decoding algorithms have been proposed to find the best sequence **a**. If prosodic labels are assumed to be independent, as in Equation 6, the highest probability sequence will be the sequence of highest probability elements $p(a_i|\mathcal{I})$.

Note that, until this point, no assumptions have been made about the dependence or independence of the sequences of acoustic, prosodic or informational events. Such assumptions are certainly incorrect. For example, both Prevost(Prevost, 1996) and Selkirk(Selkirk, 1997) propose prosodic structure that involves a combination of prominence and phrase boundary placement to cue meaning-specific speech renditions. Furthermore, though useful to simplify the decoding problem, independence assumptions are probably not viable once demands on spoken language systems become more sophisticated.

Several prosodic models have been developed that relax this assumption. For example, work in predicting prominence by (Ross et.al., 1992) makes use of a Markov assumption. Also, a hierarchical model (Ostendorf and Veilleux, 1993), makes use of the strict layer hypothesis of prosodic phrase structure, assuming that a well-formed utterance is comprised of major phrases, which are in turn comprised of minor phrases. In that work, a dynamic programming algorithm (see Figure 2.2) proposes a major phrase within an utterance, uses hypothetical minor phrases within that major phrase to estimate its probability and then, finally, choses the most likely sequence of major phrases. The most likely hypothesis of minor phrases within each major phrase is determined by using probability estimates from a previously derived binary decision tree. As we will see in the next section, the binary decision tree provides estimates of $(p(a_k|T(\mathcal{W}_i, m_{prev})))$, that is, the probability of a particular prosodic event (phrase break indices here), given the word sequence and the previous minor phrase.

Perceptual experiments have been performed to try to investigate whether the hierarchical model described above improves the intelligibility of synthesized speech. Evaluations of this sort are notably difficult, but necessary. Instead of trying to evaluate the "naturalness" of synthetic speech, Silverman et.al. (Silverman, 1993; Boogaartand Silverman, 1992) have suggested a transcription or response-type task to evaluate comprehensibility. Improved comprehensibility should manifest itself as improved transcription performance. However, in the perceptual experiment used to evaluate the hierarchical model, subjects performed similarly well on a transcription task designed to compare three different prosodic phrase break models on an AT&T TTS system (the AT&T default, the hierarchical model and a random generator)[2](.62-.67 correct, $\sigma^2 \approx 0.5$), including randomly placed breaks. Informal discussion with human subjects revealed that the task was considered very difficult. Although several subjects claimed to have understood the sentences, they said that they didn't have enough time to transcribe the sentence.

In addition to transcribing the synthetic sentences, subjects were also asked to check which, of five adjectives ("choppy", "okay", had "not enough pauses", or "unnatural"), best described the phrasing of the sentences. The results of this experiment for twenty subjects are tabulated on Table 1. Overall, more hierarchical model sentences were judged to be "okay". However, hierarchical model sentences were judged to be "choppy" more often than AT&T

[2]Details of the hierarchical model or this experiment appear in (Veilleux, 1994), also available as a postscript file via http://raven.bu.edu/~mv)

Dynamic Programming Routine for Prosodic Parse Prediction

For each word t in unit U_i $(t = 1, \ldots, l_i)$:
Compute $\log p_t(u_{i1}(1,t)|\mathcal{W}_i, U_{i-1})$.
For each n-length sequence of subunits spanning $[1,t]$ $(n = 2, \ldots, t)$:

$$\log p_t(u_{i1} \ldots u_{in}|\mathcal{W}_i, U_{i-1}) = \max_{s<t} \log p_s(u_{i1}, \ldots, u_{i,n-1}|\mathcal{W}_i, U_{i-1}) + \log p(u_{in}(s+1,t)|\mathcal{W}_i, u_{i(n-1)})$$

(Computing $\log p(u_{in}(s+1,t)|\mathcal{W}_i, u_{i(n-1)})$ with a recursive call to this routine.)
Save pointers to best previous break location s.

To find the most likely sequence,

$$p(U_i|\mathcal{W}_i, U_{i-1}) = \max_n \log p_{l_i}(u_{i1}, \ldots, u_{in}|\mathcal{W}_i, U_{i-1}) + \log q(n|l_i).$$

Here, the probability $q(n|l_i)$ provides a length constraint.

sentences, which in turn were more often considered to have "not enough pauses". Interestingly, the hierarchical model sentences were not considered more choppy than the random model's sentences, despite the minor phrase breaks generated by the hierarchical model in addition to major breaks. Moreover, the hierarchical model was given significantly more "Okay" ratings than the other two models.

Based on other evaluation metrics, the performance of the hierarchical model is believed to be of higher quality in general text-to-speech tasks. Although there are specific syntactic structures that are consistently problematic (e.g. particles and prepositions), improved POS labeling and additional rules in text-processing can easily reduce these problems. Furthermore, this work did not take into consideration discourse or non-syntax semantic information, and did not alter the AT&T defaults for placing prominences (pitch accents).

2.3 Parameter Estimation Using Binary Decision Trees

For the hierarchical model, or any future model based on the more general acoustic/ prosody /syntax formalism presented here, we need to have estimates for $p(a_i/\mathcal{I})$ in order to decode the most probable prosodic sequence. Binary decision trees are used in this work to estimate $p(a_i/\mathcal{I})$ for several reasons, and are a main feature supporting the generality and adaptability of the overall model. First, binary trees can be used to map heterogeneous features to prosodic labels. Features can be continuous (e.g. degree of syntactic bracketing) or discrete (e.g. classes of function word types). Furthermore, the features can be inter-dependent, such as *location of the last predicted phrase boundary or prominence.* The automatic algorithm which determines the tree structure typically chooses features to minimize misclassification of prosodic labels, and as such, can indicate how relevant a feature is in the choice of a particular prosodic label. Finally, and most importantly from a theoretical standpoint, a decision tree presents a model of the relationship between one domain, e.g. information structure (or alternatively acoustics) and another domain, e.g. prosody. As shown below, the path from root node to leaf is unique for each token in the input sequence and describes a prosodic label as a function of the tree features. (Let $T(\mathcal{I})$ represent a function T of the information structure $\mathcal{I}$). This model therefore allows us to map information structure onto prosodic structure, or, alternatively, to map acoustic parameters onto prosodic structure.

Binary decision trees, like the one shown in Figure 1, are a series of binary questions about features (text-derived syntactic, grammatical and pragmatic properties in this case). Each data token (a data token would be a word pair in trees designed to predict prosodic phrase breaks and a single syllable to predict ± prominence), is "dropped" through the tree, starting with the root question. In the example shown here, a binary decision tree has been grown to predict the phrase break index between each word pair. The root node represents the question *Is the left word a content word and the right word a function word?* If the answer is *no*, the word pair is dropped to the lower right node, and examined in the light of that node's question (*What function word class is the right word?*). The process is repeated until the data token reaches a leaf (terminal node). In some cases (Wang and Hirschberg, 1992) the leaf node would be associated with a specific prediction, e.g. phrase break index 1, and the word pairs shunted to this leaf node would be pre-

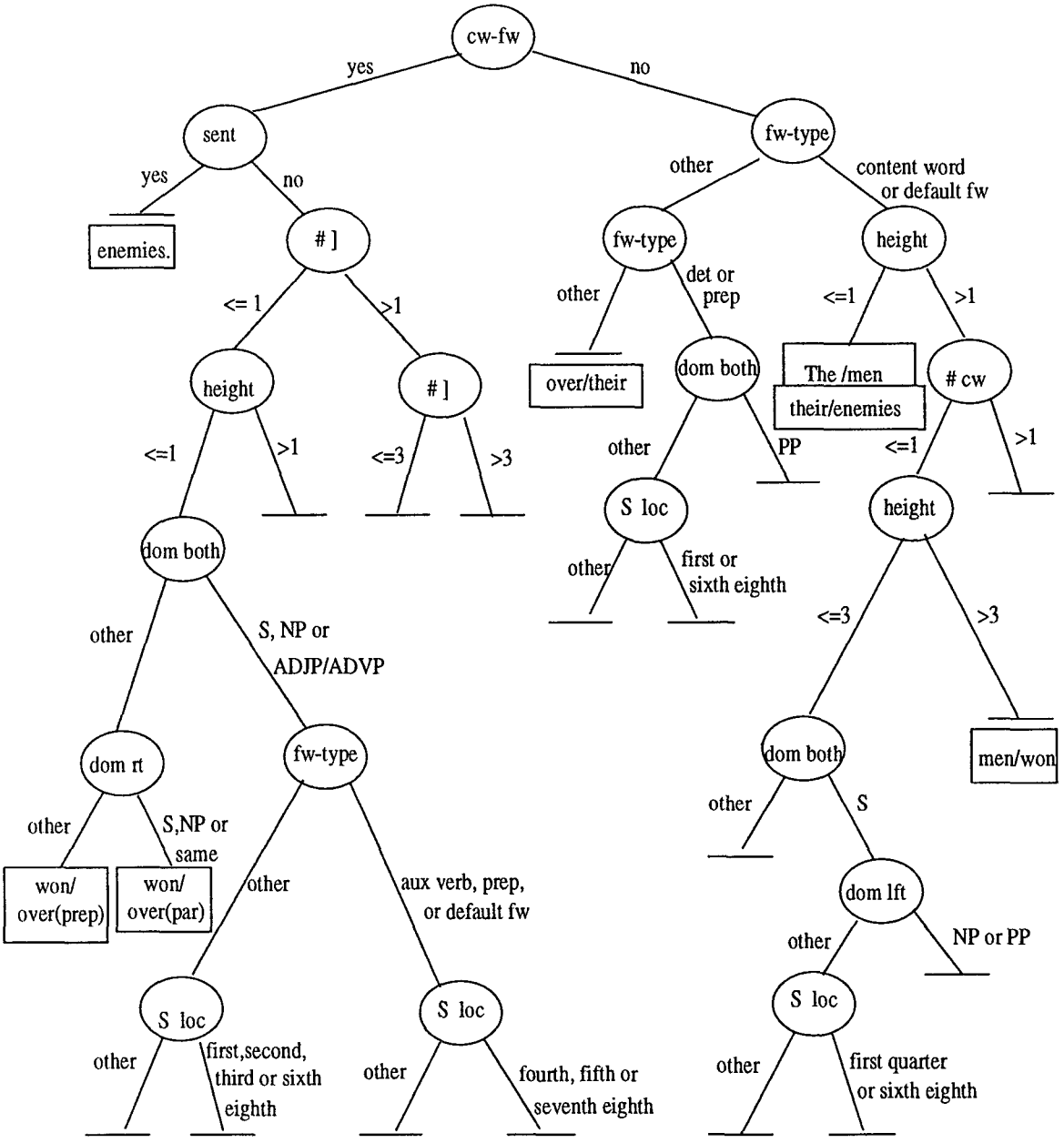

Figure 1: *Binary decision tree designed to generate the probability of break indices given syntax,* $p(b_i|syntax)$. *Each data token (a word pair) is shunted along a path from the root to a leaf node. Features include syntactic derivations (e.g. dom_lft means "what is the highest syntactic constituency that dominates the left word but not the right word"), word class type, degree of syntactic bracketing, relative position in the syntax tree and relative position in the sentence. The word pairs from the sentence* The men won over their enemies *are shown under their destination leaves.*

Table 1: Results for subjective judgments by 20 human subjects for nine sentences synthesized using each of three prosodic phrase break prediction algorithms (AT&T default, hierarchical model, and randomly placed breaks).

Phrase Break Model	choppy	okay	not enough pauses	unnatural
AT&T default	7	16	23	14
Hierarchical Model	12	23	8	17
Random	14	16	17	13

dicted to be separated by a typical intra-phrase word break. However, the approach taken here differs in that trees are not used directly to predict phrase breaks or prominences. Instead the trees are used to generate conditional probability distributions (in this example, p(break index$_i$|T(*syntax*))). The distributions are then used in the computational model described above to find the most likely sequence of prosodic labels. Again, the advantage of decoding the entire sequence is that one is able to explicitly make use of the inter-dependencies in the assignment of specific prosodic labels as in (Ostendorf and Veilleux, 1993). Furthermore, focus might be marked by a confluence of prominence and phrasing (Selkirk, 1997), requiring a relaxation of the assumption that these events are independent.

In order to see how a decision tree is used as a conditional probability estimator rather than as a predictor, recall how such a tree is originally constructed. Binary decision trees, like most statistical models, are derived using labeled training data. In this case, the training data would be hand-labeled prosodically (ToBI) and would be associated with features automatically extracted for each data token. The tree in the example above was derived to estimate the conditional probability of the prosodic break indices (b_i), given

- syntactic constituency and derivation rules,
- part-of-speech classes,
- and pragmatic rules such as relative location in the utterance.

To generate this p(b_i|T(*syntax*)) tree, data tokens (word pairs) were hand-labeled for prosodic break indices, analyzed by an automatic parser and other feature extraction programs equipped with e.g. function word tables and dictionaries, to produce a labeled database. A similar tree was also generated to estimate the probability of ± prominence, given similar syntactic/pragmatic features. To estimate the conditional probabilities between prosodic labels and acoustic signal, trees have also been derived using acoustic features such as normalized f0, duration, and vowel quality (Wightman, 1991).

After the database is fully labeled, an automatic tree-growing algorithm partitions this training data based on one of these extracted features (in Figure 1 the *cw-fw* feature) into two subsets, each more homogeneous with respect to break indices than the whole set [3]. The partitioning process is repeated on each of the subsets, using one of the extracted features and each time producing children subsets that are more homogeneous than their parents. Ideally, the final subsets, which share a particular syntactic/ pragmatic context, would contain data where all word pairs had been labeled with the same break index, prompting only a single prediction. However, this is unlikely, since all factors that determine prosody are either not yet understood, such as focus structure, or can not be known from text, such as speaking rate, and thus are not used in the partitioning process. Instead, each of the final subsets has a distribution of prosodic labels that can be used to estimate a probability distribution p($a|leaf_i$). In this example, each leaf represents a unique path which serves to describe a distribution of prosodic labels as a function of the syntax and related features. Therefore p($a|leaf_i$) = p(a|T(*syntax*)).

One interesting observation from the automatic design of the decision tree shown in Figure 1 is the selection of the cw-fw feature at the root. Sorin (Sorin et.al, 1987) has shown that prosodic phrase breaks in French tend to correspond with just cw-fw junctions. Although this rule over-generates phrase boundaries in English, the choice of this feature in the decision tree indicates a persistent correlation.

[3] In this case, words pairs spanning a content-function word boundary have been found to be associated with intonational phrase breaks (indices 3 and 4) and the data set that contains word pairs that are cw-fw pairs has a higher percentage of 3/4 break index labels than the whole data set.

3 Use of semantic features in the prosody/concept mapping

Previous work has made use of this probabilistic model of the relationships between prosody, the acoustic signal and information structure (Veilleux, 1996) but only insofar as information structure could be captured using syntax and related features. From the examples above, it is clear that prosody, even prosodic phrase structure, is not constrained by syntax alone. What remains to be investigated and incorporated are other factors that constrain prosody and are related to the concept the speaker intends to convey.

One such feature of the information structure is the placement of focussed constituents in an utterance. The literature presents a variety of definitions of semantic focus, some describing focus in terms of semantic intent (Rooth, 1994; Gussenhoven, 1994) and others more directly in relationship to givenness (Schwarzschild, 1997). Furthermore, some definitions of focus overlap with rheme (Prevost, 1996), while others do not. In any case, focus is generally agreed to be linked to pitch accent placement (e.g. (Selkirk, 1984)) and probably to phrase break placement as well (Selkirk, 1997)).

This focus/prosody relationship presents an opportunity to generate synthetic speech that has a more appropriately assigned prosodic structure, reflecting the underlying meaning to be conveyed. As the mapping between prosody and focus is investigated more fully, results can be incorporated into the computational model presented here by simply representing focus markings as labeled features in the binary tree.

Some promising work by Selkirk (Selkirk, 1997) describes the choice of prosodic phrase structure to be the outcome of an ordering of competing factors, including focus, syntax and pragmatic constraints. While some constraints may be violable (such as the alignment of major prosodic breaks with syntactic boundaries), the outcome is optimal, i.e. it conforms to the constraints of the highest ranked factor (e.g. the alignment of a major prosodic phrase boundary with the right edge of a focussed constituent).

Previous use of the acoustic/prosody/syntax model has already established the function of syntactic edges to predict prosodic phrasing (note the use of e.g. *dom_lft* features in the tree given in Figure 1). Labeling the right edges of focussed constituents in training data and growing a binary decision tree based on this additional feature, will generate probability distributions as functions of the focus as well as syntactic and pragmatic structure. If the supposition about the relationship between focussed constituents and prosodic boundaries is represented in data, such a feature should be selected as useful in decreasing the mis-classification of prosodic phrase break indices between two words in the automatic design of a binary tree. Moreover, a ranking implies an interaction of factors, each of which can be encoded as binary tree features. The order in which the features are selected in the tree structure, as well as their co-occurance (or lack of) on a root-leaf path, can indicate potential areas of interaction or redundancy. In this way, binary decision trees not only generate conditional probabilities for synthesis models, but also test hypotheses about the relative use of a feature in predicting a prosodic label.

Work by Prevost explicitly addresses the relationship between theme-rheme and prosodic prominence and phrase placement in cases of explicit contrast. This work significantly extends previous heuristics concerning newness and pitch accent placement. Again, training data labeled with theme-rheme notation, and devising a feature for the decision tree growing algorithm to select, would incorporate this rule in the estimation of probabilities of prosodic structure.

Another active and related area of research that addresses the relationship between higher order linguistic structure and prosodic structure has been explored by (Terken and Hirshberg, 1994) and (Nakatani, 1993). The latter work examines the placement of accents, as constrained by the interaction of discourse, surface structure and lexical form. Pitch accent placement on pronouns as well as on explicit forms in the subject position motivate theory that describes new and givenness in terms of a hierarchical discourse structure (Grosz and Sidner 1986). Again, the implications of this theoretical framework can be extracted as features for generating conditional probabilities of prosodic events, with reference to the theory. One such feature could be an annotation of discourse segmentation in the input text. Using this annotation as a feature in the binary tree would also serve to allow the tree to chose limits on how far back in the history list to look for an antecedent. If the pronoun represents a new (re-introduced) item within this window, it may be more likely to be accented. Again, as a feature in a decision tree, this property would be a candidate for selection to minimize mis-classification error and generate conditional probabilities that are functions of the discourse environment.

In summary, the formalism for incorporating emerging linguistic theory in a joint model of the acoustic/prosody/concept relationships is described

here. It makes use of binary decision trees to estimate model parameters, the conditional probabilities. The binary decision trees themselves, make use of explicit linguistic information to partition data into more homogeneous prosodic contexts. In doing so, the model remains general, and can accommodate the results of our evolving understanding of the interaction between factors that determine prosody. While this model has been successful in both speech synthesis and analysis applications, it has made use of syntactic and pragmatic information alone. Extension of this model to map prosodic structure to other higher order linguistic structures that more fully describe the meaning that an utterance is to convey is straightforward. As hypotheses are developed in the ranking of competing constraints, including focus structure, and in the role of discourse history, they can be integrated into the model as features in the binary decision tree.

References

Joan Bachenko and Eileen Fitzpatrick. 1990. A Computational Grammar of Discourse-Neutral Prosodic Phrasing in English. *Computational Linguistics*, Vol. 16, No. 3, pp. 155–170, 1990.

T. Boogaart and K. Silverman. 1992. Evaluating the Overall Comprehensibility of Speech Synthesizers. *International Conference on Spoken Language Processing,* pp. 1207–1210, Banff, October 1992.

James Gee and Francois Grosjean. 1983. Performance Structures: A Psycholinguistic and Linguistic Appraisal *Cognitive Psychology*, Vol. 15, pp.411–458, 1983.

Barbara Grosz and Candance Sidner. 1986. Attention, Intentions, and the Structure of Discourse," *Computational Linguistics* Vol. 12, no. 3, pp.175–204, 1986.

Carl Gussenhoven. 1994. Focus and Sentence Accents in English *Focus and Natural Language Processing*, ed. Peter Bosch and Rob van der Sandt, IBM Deutschland Informations systems, Heidelberg, 1994.

Christine Nakatani 1993. Discourse Structural Constraints on Accent in Narrative May, 1993.

Scott Prevost 1996 Modeling Contrast in the Generation and Synthesis of Spoken Language *International Conference on Spoken Language Processing,* Philadelphia, Pennsylvania, 1996.

Patti. J. Price, Mari Ostendorf, Stefanie Shattuck-Hufnagel, and Cynthia Fong, 1991 The Use of Prosody in Syntactic Disambiguation. *Journal of the Acoustical Society of America,* Vol. 6, pp. 2956–2970, 1991.

Mari Ostendorf and Nanette M. Veilleux. 1993. A Hierarchical Stochastic Model for Automatic Prediction of Prosodic Boundary Location. *Computational Linguistics* December, 1993.

Janet Pierrehumbert. 1980. *The Phonology and Phonetics of English Intonation* Ph.D. Thesis, Massachusetts Institute of Technology, 1980.

Ken Ross, Mari Ostendorf and Stefanie Shattuck-Hufnagel. 1992. Factors Affecting Pitch Accent Placement. *International Conference on Spoken Language Processing,* pp. 365–368, Banff, October 1992.

Mats Rooth. 1994. A Theory of Focus Interpretation. *Handbook of Semantic Theory*, Shalom Lappin, ed., Blackwell, 1994.

Roger Schwarzschild 1996. Givenness and Optimal Focus

Elisabeth O. Selkirk 1984. Phonology and Syntax: The Relation between Sound and Structure MIT Press, Cambridge, Massachusetts, 1984.

Elisabeth O. Selkirk. 1997. The Interactions of Constraints on Prosodic Phrasing. manuscript, 1997.

Kim Silverman, 1993. On Customizing Prosody in Speech Synthesis: Names and Addresses as a Case in Point. *Proceedings of the ARPA Workshop on Human Language Technology,* pp. 317–322, Princeton, New Jersey, March 1993.

Kim Silverman, Mary Beckman, John Pitrelli, Mari Ostendorf, Colin Wightman, Patti Price, Janet Pierrehumbert, and Julia Hirschberg. 1992. ToBI: A Standard Scheme for Labeling Prosody. *International Conference on Spoken Language Processing,* pp. 867–870, Banff, October 1992.

C. Sorin, D. Larreur and R. Llorca. 1987. A Rhythm-based Prosodic Parser for Text-to-Speech Systems in French. *Proceedings of the International Congress of Phonetic Sciences,* Vol. 1, pp. 125–128, Tallinn, 1987.

Mark Steedman. 1991. Surface Structure, Intonation, and *Focus.* The Institute for Research in Cognitive Science, University of Pennsylvania, IRCS Report Number 91–31, September, 1991.

Jacques Terken and Julia Hirschberg. 1994. *Deaccentuation of Words Representing Given Information: Effects of Persistence of Grammatical Function and Surface Position. Language and Speech* Vol. 37, no. 2.

J. 't Hart, R. Collier and A. Cohen, 1990. *A Perceptual Study of Intonation.* Cambridge University Press, 1990.

Jan van Santen. 1993. Perceptual Experiments for Diagnostic Testing of Text-to-Speech Systems. *Computer Speech and Language* 1993, **7**, pp. 49–100.

Michele Wang and Julia Hirschberg. 1992. Automatic classification of Intonational Phrase Boundaries. *Computer Speech and Language* 6-2 April 1992, pp. 175–196.

Nanette M. Veilleux. 1994. Computational Models of the Prosody/Syntax Mapping for Spoken Language Systems Ph. D. Thesis, Department of Electrical, Computer and Systems Engineering, Boston University, 1994.

Nanette M. Veilleux. 1996. Stochastic Models of Prosody for Automatic Spoken Language Systems *Proceedings of the Acoustical Society of America* Honolulu, Hawaii, December, 1997

Colin W. Wightman. 1991 *Automatic Detection of Prosodic Constituents for Parsing* Ph.D. Thesis, Department of Electrical, Computer and Systems Engineering, Boston University, 1991.