Second ESCA/IEEE Workshop
on Speech Synthesis
Mohonk, New Paltz, NY, USA
September 12-15, 1994

TRAINING INTONATIONAL PHRASING RULES AUTOMATICALLY FOR ENGLISH AND SPANISH TEXT-TO-SPEECH

Julia Hirschberg
Pilar Prieto
AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill NJ
USA, 07974–0636
{julia,prieto}@research.att.com

Abstract

We describe a procedure for acquiring intonational phrasing rules for text-to-speech synthesis automatically, from annotated text, and some evaluation of this procedure for English and Mexican Spanish. The procedure employs decision trees generated automatically, using Classification and Regression Tree techniques, from text corpora which have been hand-labeled with likely locations of intonational boundaries by native speakers, in conjunction with information available about the text via simple text analysis techniques.

Rules generated by this method have been implemented in the English version of the Bell Laboratories Text-to-Speech System and have been developed for the Mexican Spanish version of that system. These rules currently achieve better than 95% accuracy for English and better than 94% for Spanish.

Intonational Phrasing

Assigning appropriate phrasing in text-to-speech systems is important both for naturalness and for intelligibility, particularly in longer sentences and longer texts (Silverman et al., 1993). This paper describes the automatic acquisition of methods of assigning these boundaries in real-time unrestricted text-to-speech.

Intuitively, intonational phrases divide utterances into meaningful 'chunks' of information (Bolinger, 1989). Variation in phrasing can change the meaning hearers assign to tokens of a given sentence. For example, the interpretation of a sentence like '*Bill doesn't drink because he's unhappy*' will vary, depending upon whether it is uttered as one phrase or two. Uttered as a single phrase, this sentence is commonly interpreted as conveying that Bill does indeed drink — but the cause of his drinking is **not** his unhappiness. Uttered as two phrases, with an intonational boundary between '*drink*' and '*because*', it is more likely to convey that Bill does *not* drink — and that the reason for his abstinence is his unhappiness.

To characterize this phenomenon phonologically, we adopt Pierrehumbert's theory of intonational description for English (Pierrehumbert, 1980; Beckman and Pierrehumbert, 1986). In this theory, there are two levels of phrasing in English. An INTERMEDIATE PHRASE consists of one or more PITCH ACCENTS (local f0 minima or maxima) plus a PHRASE ACCENT (a simple high or low tone which controls the pitch from the last pitch accent of one intermediate phrase to the beginning of the next intermediate phrase or the end of the utterance). INTONATIONAL PHRASES consist of one or more intermediate phrases plus a final BOUNDARY TONE, which may also be high or low, and which occurs at the end of the phrase. Thus, an intonational phrase boundary necessarily coincides with an intermediate phrase boundary, but not vice versa. We employ Pierrehumbert's system also for our Spanish corpus, with modifications that do not affect the description of phrasing levels.

While we assume phrase boundaries to be perceptual categories, these have been found to be associated with certain physical characteristics of the speech signal. In addition to the tonal features described above, phrases may be identified by one of more of the following features: pauses (which may be filled or not), changes in amplitude, and lengthening of the final syllable in the phrase (sometimes accompanied by glottalization of that syllable and perhaps preceding syllables). In general, major phrase boundaries tend to be associated with longer pauses, greater tonal changes, and more final lengthening than minor boundaries. These generalizations appear to hold for both English and Spanish. In the Bell Laboratories Text-to-Speech system (TTS), intonational boundaries are realized by the manipulation of all of these features. However, currently, only intonational phrase boundaries are modelled in TTS, so this is the only level of phrasing we will discuss below.

Phrasing Prediction for Text-to-Speech

Most text-to-speech systems that handle unrestricted text rely upon simple phrasing algorithms based upon orthographic indicators, keyword or part-of-speech spotting, and simple timing information to assign phrase boundaries (O'Shaughnessy, 1989; Larreur et al., 1989; Schnabel and Roth, 1990). More sophisticated rule-based systems have so far been implemented primarily for message-to-speech systems, where syntactic and semantic information is available during the generation process (Young and Fallside, 1979; Danlos et al., 1986). However, general proposals have been made which assume the availability of more sophisticated syntactic and semantic information to use in boundary prediction (Altenberg, 1987; Bachenko and Fitzpatrick, 1990; Monaghan, 1991; Quené and Kager, 1992; Bruce et al., 1993), although no current proposal integrating such information into the process has been shown to work well even from hand-labeled input. And, even if such information could be obtained automatically and in real time for text-to-speech, such hand-crafted rules systems are notoriously difficult to build and to maintain.

Recently, efforts have been made to acquire phrasing rules for text-to-speech automatically, by training self-organizing procedures on large prosodically labeled corpora (Wang and Hirschberg, 1991a; Wang and Hirschberg, 1991b; Hirschberg, 1991; Wang and Hirschberg, 1992; Veilleux and Ostendorf, 1992). Such methods were used to train a phrasing module for the Bell Laboratories Text-to-Speech system from labeled speech from the DARPA ATIS corpus (Wang and Hirschberg, 1991a; Wang and Hirschberg, 1991b; Hirschberg, 1991; Wang and Hirschberg, 1992), which predicted intonational phrase boundaries correctly in just over 90% of cases, where data points were defined at the end of every orthographic word.

For this module, Classification and Regression Tree (CART) analysis (Breiman et al., 1984) was used to construct decision trees automatically from sets of continuous and discrete variables. In this case, these sets included values for all variables which appeared potentially useful in predicting phrasing decisions, and which could be acquired automatically from text analysis in real time.

To produce a decision tree, CART accepts as input a vector of all such independent variable values plus a dependent variable for each data point, and generates a decision tree for the dependent variable. At each node in the generated tree, CART selects the variable which best minimizes prediction error for the remaining unclassified data. In the implementation of CART used in this study (Riley, 1989), all of these decisions are binary, based upon consideration of each possible binary split of values of categorical variables and consideration of different cut-points for values of continuous variables. CART's cross-validated estimates of the generalizability of the trees it produces have proven quite accurate for the current task, when compared with tests on separate data sets; in every case CART predictions for a given prediction tree and that tree's performance on a hand-separated test set fall within a 95% confidence interval.[1]

This procedure performed fairly well, with results reported in Wang and Hirschberg 1992 of a CART cross-validated success rate of 90% correct classification of intonational phrase boundaries for trees grown using only information available automatically and in real time from text analysis. However, the hand-labeling required for the training data is enormously time-consuming and expensive, requiring well over one person-year to accomplish for the phrasing procedure described here. But automatic labeling of prosodic features does not appear to be reliable enough yet to serve as a substitute, despite some progress made in this area in recent years (Ostendorf et al., 1990).

Training Phrasing Procedures on Annotated Text

The current English version of Bell Labs TTS contains a phrasing module which was produced automatically, using procedures similar to those used in (Wang and Hirschberg, 1991a; Wang and Hirschberg, 1991b; Hirschberg, 1991; Wang and Hirschberg, 1992) to train phrasing procedures on hand-labeled speech. However, the prediction tree in this module was itself trained not on prosodically labeled speech but upon a hand-annotated corpus of approximately 87,000 words of text taken from the AP newswire, and labeled for likely prosodic boundaries by a native speaker of standard American English. The use of such text training data cuts the time needed to train a new phrasing module from well over a year to just two or three days, by eliminating the costly hand-labeling of speech. Thus it is possible to retrain the existing TTS phrasing procedure quickly, as deficiencies are uncovered, by the simple addition of exemplars of the (corrected) behavior to the training set. It is also possible to produce phrasing procedures easily for new domains or languages without recording or labeling a large corpus. The Spanish phrasing procedure recently developed is a demonstration of this technique's versatility: a baseline version of this model which performed at about 90% correct was produced in only about a one and one-half person weeks.

To produce this phrasing procedure, or a phrasing procedure for a new application or domain or a new language, we proceed as follows: On-line text from an appropriate domain is first annotated with likely intonational boundaries by a native speaker of the language for which rules are desired. We are currently using newswire text from the English and Spanish AP

[1] CART estimates are derived in (roughly) the following way: CART separates input training data into training and test sets (90% and 10% of the input data in the implementation used here), grows a tree on the training data and tests on the test data, repeats this process a number of times (five, in the implementation used here), and computes an average result for each subtree.

for general TTS training purposes, but other text could be used for particular applications, for example. The unannotated version of the text is itself analyzed to extract values for features of each potential boundary site (defined as each position between two orthographic words $< w_i, w_j >$ in the input) which have been shown or appear likely to correlate with phrase boundary location — and which can be extracted automatically and in real time. For English, these features include:

- a part-of-speech window of four around the site, $< w_{i-1}, w_i, w_j, w_{j+1} >$;
- whether w_i and w_j are ACCENTED (intonationally prominent) or not;
- the total number of words in the utterance;
- the distance in words from the beginning and end of the utterance to $< w_i, w_j >$;
- the distance in syllables and in stressed syllables of $< w_i, w_j >$ from the beginning of the utterance;
- the total number of syllables in the utterance;
- whether the last syllable in w_i is strong or weak;
- the distance in words from the previous internal punctuation to w_i;
- the identity of any punctuation occurring at the boundary site;
- whether $< w_i, w_j >$ occurs within or adjacent to an NP;
- if $< w_i, w_j >$ occurs within an NP, the size of that NP in words, and the distance of $< w_i, w_j >$ from the start of the NP

For Spanish, the feature set currently includes only the part-of-speech window, whether or not w_i and w_j are accented or not, the total number of words in the sentence, the distance of the potential site from the beginning and end of the sentence in words and from the beginning of the sentence in syllables, the identity of any punctuation occurring at $< w_i, w_j >$, the distance of $< w_i, w_j >$ from the last punctuation mark, and whether or not vowel elision would occur across $< w_i, w_j >$.

Vectors of independent feature values plus the dependent "observed" value — is an intonational phrase boundary likely to occur at $< w_i, w_j >$ in the annotator's reading of the sentence or not — are then input to an implementation of CART (Riley, 1989). Note that the features described above represent only a subset of the features originally proposed to the automatic procedure; features which are not useful in prediction are simply ignored, and can be omitted from the final tree so that those feature values will not have to be obtained in text analysis in TTS. New features can be proposed to CART as readily as the requisite information can be obtained from the text. Features tested for English, which proved not to improve performance over those noted above, include: mutual information scores for words close to $< w_i, w_j >$ and structural syntactic information about constituents bordering on $< w_i, w_j >$ and immediately dominating that site.

For the phrasing module currently implemented in English TTS, a new matrix of feature vectors can be generated for new text simply by running TTS in training mode. The resulting tree is then compiled automatically into *c* code, which can be used for prediction itself in a stand-alone procedure, or which can be substituted for an existing decision tree module in the larger phrasing module.

Evaluation and Discussion

Decision trees produced for English using annotated text for training perform somewhat better than trees trained on prosodically labeled speech, probably due to the increased size of the training set. The best result for English annotated text is a cross-validated score of 95.4% correct predictions on an 89,103 word training corpus, compared to around 90% cross-validated accuracy for the best trees trained on labeled speech. For Spanish, the best cross-validated success rate is 94.2% correct predictions of intonational phrase boundaries for a 19,473 word corpus.

We have described a procedure for training intonational phrasing decisions for unrestricted text-to-speech on annotated text, using CART techniques to generate phrasing decision trees automatically. This procedure has been used to build a phrasing module which is incorporated in the English Bell Labs TTS system and has also been used to construct a stand-alone procedure for the Mexican Spanish version of this system, for eventual inclusion in the Mexican TTS system. The advantages of this procedure are several: It makes updating an existing phrasing procedure simple and rapid: One need only provide a new or additional set of annotated text. Observed phrasing errors can often be corrected simply by providing correctly annotated exemplars of the observed error. Phrasing for new domains can also be modeled easily, simply by annotating text for the new domain. And phrasing rules can be acquired for new languages easily, limited mainly by the tools for text analysis available for the new language.

REFERENCES

Bengt Altenberg. 1987. *Prosodic Patterns in Spoken English: Studies in the Correlation between Prosody and Grammar for Text-to-Speech Conversion*, volume 76 of *Lund Studies in English*. Lund University Press, Lund.

J. Bachenko and E. Fitzpatrick. 1990. A computational grammar of discourse-neutral prosodic phrasing in English. *Computational Linguistics*, 16(3):155–170.

Mary Beckman and Janet Pierrehumbert. 1986. Intonational structure in Japanese and English. *Phonology Yearbook*, 3:15–70.

Dwight Bolinger. 1989. *Intonation and Its Uses: Melody in Grammar and Discourse.* Edward Arnold, London.

Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. 1984. *Classification and Regression Trees.* Wadsworth & Brooks, Pacific Grove CA.

Gösta Bruce, Bjorn Granström, Kjell Gustafson, and David House. 1993. Prosodic modelling of phrasing in Swedish. In David House and Paul Touati, editors, *Working Papers: Proceedings of an ESCA Workshop on Prosody*, volume 41, pages 180–183, Lund, September. Lund University Department of Linguistics.

Laurence Danlos, Eric LaPorte, and Francoise Emerard. 1986. Synthesis of spoken messages from semantic representations. In *Proceedings of the 11th International Conference on Computational Linguistics*, pages 599–604. International Conference on Computational Linguistics.

J. Hirschberg. 1991. Using text analysis to predict intonational boundaries. In *Proceedings of the Second European Conference on Speech Communication and Technology*, Genova. ESCA.

D. Larreur, F. Emerard, and F. Marty. 1989. Linguistics and prosodic processing for a text-to-speech synthesis system. In J. P. Tubach and J. J. Mariani, editors, *Proceedings of the European Conference on Speech Communication and Technology*, pages 510–513, Edinburgh, Vol. 1. CEP.

A. Monaghan. 1991. *Intonation in a Text-to-Speech Conversion System.* Ph.D. thesis, University of Edinburgh, Edinburgh.

D. O'Shaughnessy. 1989. Parsing with a small dictionary for applictions such as text to speech. *Computational Linguistics*, 15(2):97–108.

M. Ostendorf, P. Price, J. Bear, and C. W. Wightman. 1990. The use of relative duration in syntactic disambiguation. In *Proceedings of the Speech and Natural Language Workshop*, pages 26–31, Hidden Valley PA, June. DARPA, Morgan Kaufmann.

Janet B. Pierrehumbert. 1980. *The Phonology and Phonetics of English Intonation.* Ph.D. thesis, Massachusetts Institute of Technology, September. Distributed by the Indiana University Linguistics Club.

Hugo Quené and René Kager. 1992. The derivation of prosody for text-to-speech from prosodic sentence structure. *Computer Speech and Language*, 6:77–98.

Michael D. Riley. 1989. Some applications of tree-based modelling to speech and language. In *Proceedings of the Speech and Natural Language Workshop*, Cape Cod MA, October. DARPA, Morgan Kaufmann.

Betina Schnabel and Harald Roth. 1990. Automatic linguistic processing in a German text-to-speech synthesis system. In *Proceedings of the European Speech Communication Association Workshop on Speech Synthesis*, pages 121–124, Autrans. European Speech Communication Association.

Kim Silverman, Ashok Kalyanswamy, Julie Silverman, Sara Basson, and Dina Yashchin. 1993. Synthesiser intelligibility in the context of a name-and-address information service. In *Proceedings of the 3rd European Conference on Speech Communication and Technology*, volume 3, pages 2169–2172, Berlin. Eurospeech93.

N. M. Veilleux and M. Ostendorf. 1992. Probability parse scoring based on prosodic phrasing. In *Proceedings of the Speech and Natural Language Workshop*, pages 429–434, Harriman NY, February. DARPA, Morgan Kaufmann.

Michelle Q. Wang and J. Hirschberg. 1991a. Predicting intonational boundaries automatically from text: The ATIS domain. In *Proceedings of the Speech and Natural Language Workshop*, pages 378–383, Pacific Grove CA, February. DARPA, Morgan Kaufmann.

Michelle Q. Wang and J. Hirschberg. 1991b. Predicting intonational phrasing from text. In *Proceedings of the 29th Annual Meeting*, Berkeley.

Michelle Q. Wang and J. Hirschberg. 1992. Automatic classification of intonational phrase boundaries. *Computer Speech and Language*, 6:175–196.

S. J. Young and Frank Fallside. 1979. Speech synthesis from concept: A method for speech output from information systems. *Journal of the Acoustic Society of America*, 66(3):685–695, September.