PARSERS, PROMINENCE, AND PAUSES

Nick Campbell, Tony Hebert, and Ezra Black

ATR Interpreting Telecommunications Research Laboratories
http://www.itl.atr.co.jp/chatr, e-mail: nick@itl.atr.co.jp

ABSTRACT

We present results of a comparison between two prosody prediction algorithms, showing that the incorporation of information from a parser results in significantly improved performance for our text-to-speech synthesiser. We used a stochastic tree-based parser to generate a tagged and bracketed representation of the input text, and then interpreted this higher-level information to produce a ToBI-type prosodic annotation of the text. From this annotation an intonation contour was predicted for use in synthesising the speech. Results show that prediction of prosodic phrasing and focal prominence are improved by 56% and 62% respectively over previous methods compared against a human reading of the same test utterances.

1. INTRODUCTION

The intelligibility of synthetic speech depends on a combination of clear voice quality and suitable prosody to portray the meaning of each utterance in context [1, 12]. The Chatr synthesis system [9, 4, 8] re-cycles segments of natural speech to ensure high definition in voice quality, but relies on prediction of an appropriate prosodic contour in order to select the segment sequence that most faithfully represents the intended meaning of a given utterance. In the case of machine-mediated speech, we can use the input prosody as a guide for the output synthesis, but when synthesising from plain text alone, we have to estimate the prosody by rule. The key decisions to be made in this case pertain to phrasal boundary position and semantic focus.

Previous methods of predicting prosodic phrasing and prominence have, in the absence of a reliable parser, had to rely upon a sparse analysis of the input text [3] or on heuristic devices such as length of utterance [5, 6] to determine where to insert a pause or to add prominence to a syllable.

Earlier work from this lab [15] resulted in multi-level intonation prediction systems to generate a basic fundamental frequency contour from part-of-speech information and syntactic constituent structure and then modify it according to higher-level discourse information, such as speech act type and scope of focus, when available. Chatr currently offers several methods for intonation prediction, they are all rule driven but the rules and parameters are derived automatically from naturally spoken dialogues.

The 'Hirschberg' method [14] assigns each word to one of four accentuation levels. The algorithm, based primarily on part of speech tags, distinguishes key words into four classes, though proper nouns, numbers and complex nominals form special cases, and there are special rules for specific words such as *"not"*, *"but"*, *"first"* etc. Our implementation conflates the *emphatic* and *accented* types to 'accented' and *de-accented* and *cliticised* types to 'unaccented'.

The 'Monaghan' algorithm [16] defines an explicit notion of prosodic phrase, and of its internal accent structure, in which each phrase must contain one and only one nuclear accent. No accents can follow within a phrase, but secondary accents may precede the nucleus. After initial accent assignment a *Rhythm Rule* ensures the well-formed-ness of accents by limiting accentuation of adjoining words. In addition to the general conditions there are a number of specific heuristics for certain words such as *"not"* and *"but"*.

The 'Decision Tree' method is automatically trained from word feature vectors, using classification and regression trees derived from CART [13] to predict accents from a window of 5 part-of-speech tags (including the current word and two on either side) plus the boundary type for the current word.

We previously tested the above three methods [15] and concluded that the decision trees were best for use as a default, because they model the contour well and offer ease of training and adaptation to new data. However, recent developments in parser and tagger technology have necessitated further tests as large data-based non-heuristic text analysis systems have become available.

The algorithm proposed in this paper uses the output of one such parser and is adapted for the ToBI system of prosodic labelling [?, ?]. It first checks whether a word is strongly linked with the word that follows it, and then, by extension, whether a group of words is strongly linked with the group of words that follows, to cluster similar constituents into 'phrases'. Finally, it marks the last content word of each phrase thus formed with a ToBI 'H*' accent to indicate prominence in the default case. Because of the semantic information available from the parser, such word grouping allows very natural-sounding intonation to be predicted.

2. A NEW PARSER

The new ATR General English Parser (SPATR) is a grammar–based probabilistic parser trained on a large, highly varied tree-bank of unrestricted English text [7]. Probabilistic decision trees are utilized as a means of prediction, and a grammar with about 3000 semantic–and–syntactic tags, and 1100 non–terminal node labels supplies detailed linguistic information. Further such data is supplied for prediction purposes by questions about "raw" words, expressions, and the sentence as a whole. The questions are created in the first place by a grammarian utilizing a flexible special–purpose language, then the system is trained by exposure to a very large tree-bank of parsed texts. The rich information base used for parse prediction allows the system to parse in a domain–general, open–vocabulary setting, and to output detailed semantic as well as syntactic information for each sentence processed.

Given an input sentence such as the following:

> "We do charge a cancelation fee of three hundred and fifteen dollars if you cancel less than a week in advance"

the output from the parser will be:

[start [sprpd23 [sprime4 [sd1 [nbar6 We-PPIS2 nbar6] [vbar2 [o8 do-VDO o8] [v2 charge-VVICOMP-B [nbarq4 [nbar4 [d1 a-AT1 d1] [n4 [n1a cancelation-NN1 FUNCTION n1a] [n1a fee-NN1MONEY n1a] n4] nbar4] [i1e [p1 of-IIOF [nbar1 [n1c [multi-word4 three-MPRICEWRD51 hundred-MPRICEWRD52 and-MPRICEWRD53 fifteen-MPRICEWRD54 dollars-MPRICEWRD55 multiword4] n1c] nbar1] p1] i1e] nbarq4] v2] vbar2] sd1] [iebar11 [fa1 if-CSIF [sd1 [nbar6 you-PPY nbar6] [vbar1 [v2 cancel-VVOINCHOATIVE [nbar2 [d25 less-DAR [fc1 than-CSN [nbarq4 [nbar4 [d1 a-AT1 d1] [n1a week-NR1 n1a] nbar4] [i1e [p1 in-IIIN [nbar1 [n1a advance-NN1TIME n1a] nbar1] p1] i1e] nbarq4] fc1] d25] nbar2] v2] vbar1] sd1] fa1] iebar11] sprime4] sprpd23] start]

Every non-terminal node is labelled with the name of the ATR English Grammar rule[1] that generates the node; and each word is labelled with one of the tags in the grammar's tag-set (See [11]) Together, the bracket locations, rule names, and lexical tags of a Tree-bank parse specify a unique parse within the grammar. In the Grammar parse, rule names and lexical tags are replaced by bundles of feature/value pairs. Each node contains values for 66 features, and there are 12 values per feature, on average.

Figure 1 presents a detail of the beginning of this example to illustrate the structure of the parse.

Prediction in the parser is conditioned partially on questions about feature values of words and non-terminal nodes. For instance, to predict whether a

[1] There are currently 1155 rules in the Grammar.

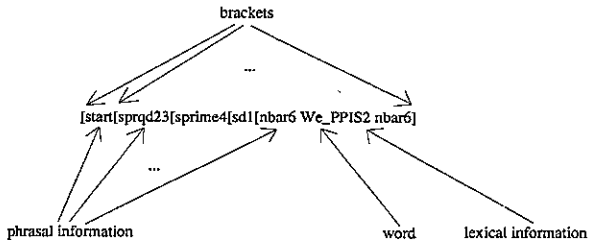

Figure 1: Components of the parser analysis

```
(Utterance PhonoWord
  (:D ()
    (:S ()
      (:C ()
        (We)
        (do)                    (:C ()
        (charge)                     (if)
        (a)                          (you)
        (cancelation)               (cancel))
        (fee (H*)))                (:C ()
      (:C ()                        (less)
        (of)                        (than)
        (three)                     (a)
        (hundred)                   (week)
        (and)                       (in)
        (fifteen)                   (advance (H*))))))
        (dollars (H*)))
```

Figure 2: Phonoword representation for Chatr

constituent has ended, it will count the number of words until the next finite verb; the next comma; the next noun; etc. In tagging, it will check whether the same word has already occurred in the sentence, and if so, determine its value in relation to the previous occurrence with respect to the various relevant features.

By labelling Tree-bank nodes with Grammar rule names, and not with phrasal and clausal names, as in other (non–grammar–based) tree-banks, the parser is able to gain access to all information provided by the Grammar regarding each Tree-bank node.

3. PROSODY FROM PARSE

Our synthesis algorithm takes the parser output as input and, by reducing the brackets, produces a PhonoWord [9] Utterance representation as output (see Fig 2). It then reduces unstressed words, predicts pauses based on strength of bracketing and phrasal information, and marks (currently) the last content word of each phrase with a ToBI H* to show prominence.

Because the ATR English Grammar is detailed and comprehensive, complete syntactic and semantic analysis can be performed on nominal compounds (e.g. "the Heathrow Airport Long Term Car Park Courtesy Bus Pick-up Point", or "high definition

speech synthesis system") to allow more intelligible grouping of the sub-components. Further, the full range of attachment sites is available within the Grammar for sentential and phrasal modifiers, so that differences in meaning can be accurately reflected in parses. For instance, in "I couldn't come because I was talking, and didn't call for the same reason," the phrases "because I was talking" and "for the same reason" should probably post-modify their entire respective verb phrases, "couldn't come" and "didn't call", for maximum clarity.

4. EVALUATION

To compare the output of the algorithm with the phrasing produced by a native-speaker of English reading the same texts, we recorded readings of 8,500 words from documents taken from various different sources on the internet. After digitization of the speech waveforms thus produced, pauses and prominence were marked by hand by a trained human prosody-database labeller. The same texts were then synthesised by Chatr using the best of the previous and the current improved prosodic prediction algorithms. By comparing the two synthesised utterances against the labels of the human original, we were able to evaluate how closely to natural speech the pause and prominence predictions by Chatr were and to quantify the improvements gained by the proposed algorithm.

The texts used in the evaluation were:

baa304 (79 parsed sentences, 1078 words)
New York City Geographical Information

baa305 (99 parsed sentences, 1789 words)
Remarks of U.S. Secretary of Commerce Ronald H. Brown at the U.S.-GCC Economic Dialogue Riyadh, Saudi Arabia January 16, 1993

baa308 (117 parsed sentences, 2102 words)
Remarks by U.S. Secretary of Commerce Ronald H. Brown at the Martin Luther King, Jr. Holiday Event Amman, Jordan January 17, 1994

baa393 (131 parsed sentences, 3246 words)
Zen and the Art of Weight-lifting

5. RESULTS

The texts were processed by Chatr using the previous and new prosody modules, and the results were scored as follows:

If a pause predicted by the algorithm matched the position of a ToBI break-index label BI3 marked on the human reading, this was counted as a 'success', otherwise in any other position it would generate an error. The overall 'success rate' is defined by (successes)/(successes+errors)*100.

Table 1: Agreement with natural phrasing

baa304 : Breaks	missed	extra	correct	rate
NEW	17	10	213	88%
PREVIOUS	32	32	201	75%
baa304 : Prom	missed	extra	correct	rate
NEW	35	22	185	76%
PREVIOUS	1	105	159	60%
baa305 : Breaks	missed	extra	correct	rate
NEW	45	21	349	84%
PREVIOUS	73	68	453	76%
baa305 : Prom	missed	extra	correct	rate
NEW	74	33	308	74%
PREVIOUS	11	222	359	60%
baa308 : Breaks	missed	extra	correct	rate
NEW	27	15	223	84%
PREVIOUS	121	164	926	76%
baa308 : Prom	missed	extra	correct	rate
NEW	45	23	196	73%
PREVIOUS	32	445	734	60%
baa393 : Breaks	missed	extra	correct	rate
NEW	15	26	372	89%
PREVIOUS	72	130	619	75%
baa393 : Prom	missed	extra	correct	rate
NEW	63	41	306	73%
PREVIOUS	19	354	446	54%

If a labelled prominence matched with a prominence predicted by the algorithm (the last content word of each phrase) it was counted as a success, otherwise an error. As above, the success rate is defined by (successes)/(successes+errors)*100.

The results of the comparison are shown in Table 1. The details of performance are very similar, regardless of text type, so we can average them to obtain the results below, indicative of the general case. Figures in brackets show the overall percentage for missed and unnecessarily-inserted labels respectively.

Success rate:	Pauses		Prominence	
NEW	(-7% + 5%)	86%	(-14% + 8%)	74%
PREVIOUS	(-10% + 13%)	75%	(-1% + 39%)	58%

The new algorithm using information from the full parse clearly does better than the previous algorithm for both pause and prominence prediction, reducing the error rate for pause insertion by 56% and for prominence assignment by 62% .

6. DISCUSSION

The ATR parser is a probabilistic parser which uses decision–tree models. A parse is built up from a succession of states, each of which represents a partial parse tree. Transition between states is accomplished by one of the following steps: (1) assigning syntax to a word; (2) assigning semantics to a word; (3) deciding whether the current position is the end of a

constituent; (4) assigning a (rule) label to an internal node of the parse tree. Note that the first two steps together determine the tag for a word. Corresponding to each type of step is a model which estimates the probability of the outcome. For efficiency, the semantic model is represented by a set of models, one for each syntactic category. Each model uses as input the answers to a set of questions designed specifically for that model by a grammarian.

We attribute the improved performance to the fact that so much information is embedded in the parser regarding linguistic attributes of the words in the text. Previous parses based on minimal syntactic information were unable to disambiguate much of the bracketing and could only produce a simple default clustering of the text. Because the semantic information is also taken into account, the bracketing from the new parser is improved, and a simple prominence algorithm such as 'last word in phrase' can suffice.

However, because of processing constraints, the parser is not yet able to operate in real-time, and slows down the process of text-to-speech synthesis considerably. We are currently working on optimising the parser for speed as well as performance, and anticipate that it will be working in close to real time in the very near future.

7. CONCLUSION

In this paper, we reported improvements to the module which is used to predict intonation from text in a text-to-speech system. The improvements come largely from the incorporation of an improved parser and reduce the previous prediction error by 56% for prosodic boundaries, and 62% for marking of focus. The tests were performed using texts obtained from the internet, exemplifying a variety of information styles, and results were obtained by comparison with human readings of the same texts.

By incorporating the improved phrasal and lexical information provided by the ATR parser into the CHATR synthesis system, we have shown that it is possible to predict pitch accents, pauses, and phrasing to a higher degree than before.

Future work will involve generalising the parse-to-prosody algorithms so that the mapping to an intonation contour can be learnt directly from the labelled corpus without the need for an intermediate level of heuristic processing. If this is successful, then we will be able to model the speaker-specific variation in intonation that is necessary if different dialects are to be synthesised.

Acknowledgments

The authors are particularly grateful to S. Eubank, H. Kashioka, and D. Magerman for their assistance with this work.

REFERENCES

[1] K. Silverman, S. Basson, & S. Levas, "Evaluating synthesiser performance: is segmental intelligibility enough?", pp 147-150 in Proc ICSLP-90, 1992.

[2] A. W. Black & A. J. Hunt, "Generating F0 contours from ToBI labels using linear regression, Proc ICSLP-96, pp.1385-1388, 1996.

[3] A. W. Black, "Predicting the intonation of discourse segments from examples in dialogue speech", pp. 117-127 in **Computing Prosody** Sagisaka, Campbell, & Higuchi, eds, Springer-Verlag, N. Y., 1996.

[4] W. N. Campbell, " CHATR: A High-Definition Speech Re-Sequencing System", Proc 3rd ASA/ASJ Joint Meeting, 1223-1228, Hawaii, 1996(12).

[5] M. Wang & J. Hirschberg, **Automatic classification of intonational phrase boundaries**, pp 175-196 in Computer Speech & Language 6, 1992.

[6] M. Ostendorf & N. Veilleux, "A hierarchical stochastic model for automatic prediction of prosodic boundaries and ite relation to suprasegmental cues", JASA 90, 2275, 1991.

[7] E. Black, S. Eubank, H. Kashioka, R. Garside, G. Leech, & D. Magerman. "Beyond skeleton parsing: producing a comprehensive large–scale general–English treebank with full grammatical analysis". In *Proc 16th ACL*, pp.107–112, 1996.

[8] W. N. Campbell, Y. Itoh, W. Ding, & N. Higuchi, "Factors Affecting Perceived Quality and Intelligibility in the Chatr Concatenative Speech Synthesiser", this volume.

[9] W. N. Campbell & A. W. Black, "CHATR: a multilingual speech re-sequencing synthesis system", 45-52, SP96-7 Tech Rept IEICE, 1996.

[10] A. W. Black and P. Taylor. Assigning intonation elements and prosodic phrasing for English speech synthesis from high level linguistic input. In *ICSLP94*, Vol 2, pp 715–718, Yokohama, 1994.

[11] E. Black, S. Eubank, H. Kashioka, R. Garside, G. Leech, and D. Magerman. 1996. Beyond skeleton parsing: producing a comprehensive large–scale general–English treebank with full grammatical analysis. In *Proceedings of the 16th Annual Conference on Computational Languistics*, pages 107–112, Copenhagen.

[12] W. N. Campbell: "Synthesis Units for Natural English Speech", Transactions of the Institute of Electronics, Information and Communication Engineers, SP 91-129, pp 55 - 62.1992.

[13] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees.* newblock Wadsworth & Brooks, Pacific Grove, CA., 1984.

[14] J. Hirschberg. Using discourse content to guide pitch accent decisions in synthetic speech. In G. Bailly and C. Benoit, editors, *Talking Machines*, pages 367–376. North-Holland, 1992.

[15] A. W. Black, "Comparison of algorithms for predicting accent placement in English speech synthesis", Proc ASJ Spring meeting, 1995.

[16] A. Monaghan. *Intonation in a text-to-speech conversion system.* PhD thesis, University of Edinburgh, 1991.

Parsers, Prominence, and Pauses

Nick Campbell, Tony Hebert, and Ezra Black
ATR Interpreting Telecommunications Research Laboratories

Intelligibility of synthetic speech depends on a combination of clear voice quality and suitable prosody [5]. Previous methods of predicting prosodic phrasing and prominence have, in the absence of a reliable parser, relied upon sparse analysis of the input text [2] or on heuristic devices such as length of utterance [6, 4] to determine where to insert a pause.

The new ATR General English Parser is a grammar–based probabilistic parser trained on a large, highly varied treebank of unrestricted English text[1]. Probabilistic decision trees are utilized as a means of prediction, and a grammar with about 3000 semantic–and–syntactic tags, and 1000 non–terminal node labels supplies detailed linguistic information. Further such data is supplied for prediction purposes by questions about "raw" words, expressions, and the sentence as a whole. These questions are created by a grammarian utilizing a flexible special–purpose language. The rich information base used for parse prediction allows the system to parse in a domain–general, totally–open-vocabulary setting, and to output detailed semantic as well as syntactic information for sentences processed.

By incorporating the phrasal and lexical information provided by the ATR parser in the CHATR synthesis system [3], we have shown that it is possible to predict pitch accents, pauses, and phrasing to a higher degree than before. Given text input such as the following:

> "We do charge a cancellation fee of three hundred and fifteen dollars if you cancel less than a week in advance"

the output from the parser is:

> [start [sprpd23 [sprime4 [sd1 [nbar6 We-PPIS2 nbar6] [vbar2 [o8 do-VDO o8] [v2 charge-VVICOMP-B [nbarq4 [nbar4 [d1 a-AT1 d1] [n4 [n1a cancellation-NN1 FUNCTION n1a] [n1a fee-NN1MONEY n1a] n4] nbar4] [i1e [p1 of-IIOF [nbar1 [n1c [multiword4 three-MPRICEWORD51 hundred-MPRICEWORD52 and-MPRICEWORD53 fifteen-MPRICEWORD54 dollars-MPRICEWORD55 multiword4] n1c] nbar1] p1] i1e] nbarq4] v2] vbar2] sd1] [iebar11 [fa1 if-CSIF [sd1 [nbar6 you-PPY nbar6] [vbar1 [v2 cancel-VVOINCHOATIVE [nbar2 [d25 less-DAR [fc1 than-CSN [nbarq4 [nbar4 [d1 a-AT1 d1] [n1a week-NR1 n1a] nbar4] [i1e [p1 in-IIIN [nbar1 [n1a advance-NN1TIME n1a] nbar1] p1] i1e] nbarq4] fc1] d25] nbar2] v2] vbar1] sd1] fa1] iebar11] sprime4] sprpd23] start]

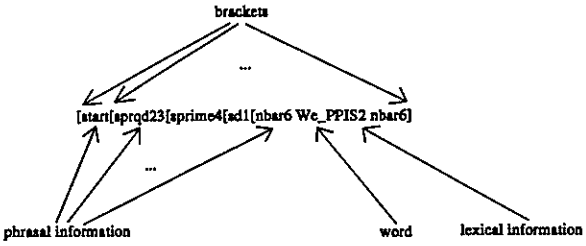

Fig. 1 Components of the parser analysis

```
(Utterance PhonoWord
  (:D ()
    (:S ()
      (:C ()
        (We)
        (do)                    (:C ()
        (charge)                    (if)
        (a)                         (you)
        (cancellation)              (cancel))
        (fee (H*)))               (:C ()
      (:C ()                        (less)
        (of)                        (than)
        (three)                     (a)
        (hundred)                   (week)
        (and)                       (in)
        (fifteen)                   (advance (H*))))))
        (dollars (H*)))
```

Fig. 2 Phonoword representation for Chatr

Figure 1 presents a detail of the beginning of this example to explain some of the structure. Our synthesis algorithm takes the parser output as input and, by reducing the brackets, produces a PhonoWord Utterance representation [3] as output (see Fig 2). It reduces unstressed words, predicts pauses based on strength of bracketing and phrasal information,

Table 1 Agreement with natural phrasing

baa304 : Breaks	missed	extra	correct	rate
NEW	17	10	213	88%
CHATR	32	32	201	75%
baa304 : Prom	missed	extra	correct	rate
NEW	35	22	185	76%
CHATR	1	105	159	60%
baa305 : Breaks	missed	extra	correct	rate
NEW	45	21	349	84%
CHATR	73	68	453	76%
baa305 : Prom	missed	extra	correct	rate
NEW	74	33	308	74%
CHATR	11	222	359	60%
baa308 : Breaks	missed	extra	correct	rate
NEW	27	15	223	84%
CHATR	121	164	926	76%
baa308 : Prom	missed	extra	correct	rate
NEW	45	23	196	73%
CHATR	32	445	734	60%
baa393 : Breaks	missed	extra	correct	rate
NEW	15	26	372	89%
CHATR	72	130	619	75%
baa393 : Prom	missed	extra	correct	rate
NEW	63	41	306	73%
CHATR	19	354	446	54%

and marks (currently) the last content word of each phrase with a ToBI H* to show prominence.

In essence, the algorithm checks if a word is strongly linked with the word that follows, and by extension, if a group of words or phrase is strongly linked with the group of words that follows it, and clusters similar constituents into a phrase.

To compare the output of the algorithm with phrasing produced by a native-speaker, we recorded readings of 8,500 words of text taken from different sources on the internet. From the speech waveforms thus produced, pauses and prominence were marked by hand. The same texts were then synthesised by Chatr using the previous and improved prosodic prediction algorithms. By comparing the two synthesised utterances against the human original, we were able to evaluate how closely to natural speech the pause and prominence predictions by Chatr and by the proposed algorithm were.

If a pause predicted by the algorithm matched a ToBI break-index label BI3 marked in the human reading, it was counted as a success, otherwise an error. The overall success rate is defined by (successes)/(successes+errors)*100.

If a labelled prominence matched with a prominence predicted by the algorithm (the last content word of each phrase) it was counted as a success, otherwise an error. As above, the success rate is defined by (successes)/(successes+errors)*100.

The results of the comparison are shown in Table 1, and a summary is presented Below. The details of performance are very similar, regardless of text type. So we can average them to obtain the results below, indicative of the general case.

Success rate:	Pauses	Prominence
NEW	(-7% +5%) 86%	(-14% +8%) 74%
CHATR	(-10% +13%) 75%	(-1% + 39%) 58%

The new algorithm using information from the full parse clearly does better than the previous algorithm for both pause and prominence prediction, reducing the error rate for pause insertion by 56% and for prominence assignment by 62% .

Acknowledgements

The authors are particularly grateful to S. Eubank, H. Kashioka, and D. Magerman for their assistance with this work.

Bibliography

[1] E. Black, S. Eubank, H. Kashioka, R. Garside, G. Leech, and D. Magerman. "Beyond skeleton parsing: producing a comprehensive large-scale general-English treebank with full grammatical analysis". In *Proceedings of the 16th Annual Conference on Computational Languistics*, pages 107–112, Copenhagen, 1996.

[2] A. W. Black, "Predicting the intonation of discourse segments from examples in dialogue speech", pp. 117-127 in **Computing Prosody** Y. Sagisaka, N. Campbell, & N. Higuchi, eds, Springer-Verlag, N. Y., 1996.

[3] W. N. Campbell & A. W. Black, "CHATR: a multilingual speech re-sequencing synthesis system", 45-52, SP96-7 Tech Rept IEICE, 1996.

[4] M. Ostendorf & N. Veilleux, "A hierarchical stochastic model for automatic prediction of prosodic boundaries and ite relation to suprasegmental cues", JASA 90, 2275, 1991.

[5] K. Silverman, S. Basson, & S. Levas, "Evaluating synthesiser performance: is segmental intelligibility enough?", pp 147-150 in Proc ICSLP-90, 1992.

[6] M. Wang & J. Hirschberg, **Automatic classification of intonational phrase boundaries**, pp 175-196 in Computer Speech & Language 6, 1992.