The Boston University Radio News Corpus

M. Ostendorf* P. J. Price† S. Shattuck-Hufnagel‡
Boston University †SRI International ‡MIT

3 February 1995

Abstract

We describe a corpus of professionally read radio news data, including speech and accompanying annotations, suitable for speech and language research. The corpus consists of over seven hours of speech recorded from seven radio announcers (4 male, 3 female). Subsets of the corpus are labeled with phonetic alignments, part-of-speech tags and prosodic markers. We motivate the uses of such a corpus and describe the data and labeling mechanisms.

1 Introduction

The last decade of speech research has seen tremendous gains in computer speech processing technology, as well as in our fundamental understanding of human speech communication. Many of these gains can be attributed directly or indirectly to the availability of large shared corpora, which facilitate knowledge acquisition in several ways. Sharing of resources makes more data available for analyses; this is particularly useful for the more costly portions of the corpus, such as hand labeling. Further, large corpora make possible the systematic development and testing of automatically trained algorithms. Of course, shared corpora also allow for more direct comparison of algorithm performance and other research results.

There are many examples of common speech corpora in speech recognition and understanding research today. For American English, such corpora include the TIMIT [14], Resource Management [23], ATIS [22], Switchboard [11], Wall Street Journal [2] and TRAINS corpora. Of these, only the ATIS corpus has even limited prosodic notation. The ATIS transcription conventions mark extra emphasis or lengthening and some phrasing, but because of the multiple transcribers at various sites, the inclusion of these markings is sporadic at best. A small subset of the ATIS and TRAINS corpora have also been transcribed more systematically with prosodic markers, but these labels are not generally available.

Other European and US sources of data are described by Edwards and Lampert in [9]; many of these sources are available through the ACL Data Collection Initiative, or through the Linguistics Data Consortium. Though many corpora are available, they are focussed on text; availability of the corresponding digitized speech is rare. Since many aspects of prosody can interact with syntax, having some syntactic information is very helpful. Edwards and Lampert also survey several sources of syntactic bracketing and labeled parts of speech, but only the Penn Treebank [17] seems to include speech data and only for some of the labeled data. The London-Lund corpus of spoken educated British English [1, 30] includes syntactic and prosodic markings, but in general the speech is not available. (Semantic information would of course also be helpful, but is currently not available in any common corpus.)

In sum, there is relatively little American English data available for research in prosody and/or speech synthesis.[1] In the public domain, there are several sources of large amounts of text data, fewer of transcribed and digitized speech, but very few sources of speech annotated with prosodic transcriptions. Rarer still are those speech corpora with accompanying syntactic or other analyses.

This paper describes a corpus designed to fill this gap, specifically by providing material for the study of prosodic patterns, with particular emphasis on synthesis applications. The corpus consists of radio newscast speech, as described in Section 2. Utterance annotation includes orthographic transcriptions, phonetic alignments, part-of-speech tags, and prosodic

[1]For other versions of English, the situation is somewhat better, with ongoing transcription efforts that should provide richer corpora in the future. Existing speech corpora include the HCRC Map Task (mainly Glaswegian English), the Lancaster Spoken English Corpus (described in [9]) and Australian English.

labels. Section 3 describes these annotations and assesses their accuracy and consistency. Finally, Section 4 describes potential uses of this corpus and other future directions.

2 Speech Corpus

The Boston University Radio News Corpus includes speech from FM radio news announcers associated with WBUR, a public radio station. The corpus was collected primarily to support research in text-to-speech synthesis, particularly generation of prosodic patterns. The FM radio newscasting style is appropriate for research and development in speech synthesis and computational modeling of prosody for a number of reasons. First, the FM newscaster faces a task similar to that of a text-to-speech synthesis system, in that text should be read out in a pleasant way (i.e., the prosody should not be too monotonous) without misleading the listener (i.e. the risks of ambiguity may be less than those of choosing the wrong prosodic pattern). Second, the particular strategy for prosodic marking used by radio announcers is well-suited to analysis for synthesis applications: there is evidence that these newscasters use more clear and consistent indications of prosodic structure than non-professional read speech [20], which should facilitate analyses and modeling of the style; and there is evidence that the radio news style may facilitate comprehension of spoken information by using F0 to delimit semantic units and to mark sentential focus [10]. Perhaps for the above reasons, the radio news style has also been the basis for synthesis and other studies in other languages (e.g., see Sorin *et al.* 1987 for French, Strangert 1991 for Swedish, and Vihanta 1991 for Finnish, Fujisaki *et al.* for Japanese) [28, 29, 32, 10]. Finally, the radio newscasting style provides a natural but controlled style, combining some of the advantages of read speech (greater control in that the newscasters can be given material to read), with some of the advantages of non-read speech (it is an accepted, naturally-occurring form of read speech).

The main *radio news* portion of the corpus consists of news stories recorded in the WBUR radio studio during broadcast. Speech from each of seven speakers (three female

Speaker	F1A	F2B	F3A	M1B	M2B	M3B	M4B
Minutes	52	49	107	48	58	32	91
Stories	43	34	340	36	35	21	62
Clean Paragraphs	276	124	341	161	214	126	236
Noisy Paragraphs	1	40	51	108	102	32	41
Words (*times* 1000)	11.9	12.2	28.6	15.7	18.4	10.5	25.6

Table 1: Duration in minutes of speech, and other statistics about the radio news stories, compiled for each of the speakers in the radio news corpus. 'F' or 'M' in the speaker identifiers indicate a female or male speaker, respectively; the number is a unique identifier within the male or female speakers; and the 'A' or 'B' indicates the speaker type, as described above.

and four male) were dubbed from these tapes to new tapes for compiling speaker dependent data. Two types of announcers were recorded; speakers are indexed by (A) if their job is normally to read news live and (B) if they normally pre-record and edit their stories. The stories read by type-B announcers are typically longer, more in-depth feature stories than the short news updates read by the type-A speakers. Type-B speakers are also more familiar with the material, since they wrote it. The distribution of the approximately seven hours of speech across speakers is indicated in Table 1. Additional data from other speakers (other radio announcers and interviewees) is available on audio tape from the original recordings, but this data has not been digitized or transcribed.

Since the radio news stories were recorded during actual news broadcasts, some utterances include background sound effects. Such sound effects include, for example, rushing water, traffic noise, music, and low-level conversations. This noise occurs in different percentages of the paragraphs, depending on the style of the announcers. In particular, it is used often in stories read by the type B announcers, but almost never in the speech recorded from the type A announcers, as indicated in Table 1. The noise appears to be at

Speaker	CJ	CP	TP	SR
Paragraphs	6	4	7	7
Sentences	23	22	28	36
Words	445	388	577	713

Table 2: Characteristics of the news stories recorded in the lab for multiple speakers. above.

a high enough level to affect automatic phonetic alignment but does not seem to hurt F0 contour analysis, based on examining several paragraphs of noisy data. As a consequence, this subset of the data is useful for some types of prosodic analysis, but would not be appropriate for generating diphone elements for waveform concatenation in synthesis. Speech without such sound effects is referred to as 'clean speech' below. (File naming conventions and documentation available with the corpus indicates which files are considered 'clean' vs. 'noisy'.)

In addition to the recordings made in the radio studio, we have recorded six of the announcers reading the same four type-B news stories in our laboratory, referred to as the *lab news* portion of the corpus. The multiple versions of each story provide insight into the amount of variability in prosodic patterns across speakers that is acceptable for a given sentence, as explored in [25]. For this reason, this subset is designated as the 'test' portion of the corpus and annotation for this subset will be hand-corrected. The announcers were asked first to read the stories in their non-radio style and then, 30 minutes later, to read the same stories in their radio style. Thus, the corpus includes examples of the same speaker reading the same story in different styles. Although we have not examined this style factor in detail, there appear to be clear differences in the styles for most speakers. However, many of the announcers seemed to slip into the radio style at times during the non-radio recording, so the data may not consistently represent non-professionally read speech.

An advantage of working with professional radio announcers is that they tend to be more fluent than most non-professional speakers, e.g. they produce fewer disfluencies like "er" and "um", and fewer prosodic errors in the sense of using inappropriate constituent boundaries or accent placements. Type B announcers have the option of editing their stories before broadcast, further reducing the number of disfluencies in the data. Of the type A announcers, F1A was almost never disfluent, and F3A had a low rate of disfluencies. To obtain similar quality recordings for the laboratory news speech, we allowed the announcers to reread paragraphs if they made a mistake (i.e., a prosodic error, disfluency, or misreading). A few disfluencies were not caught at the time of recording, and these have been edited out of the speech where possible without affecting the naturalness of the utterance. For example, words or word fragments associated with restarts could be edited out, but mispronounced words could not be. An ascii documentation file associated with each speaker indicates where disfluencies were edited out.

We also recorded additional materials from the radio announcers in order to study specific questions about prosody. These corpora are documented elsewhere, and are mentioned only briefly here. The *ambiguous sentence corpus* [21] included 35 pairs of phonetically identical and syntactically ambiguous sentences read in disambiguating contexts by four of the announcers. These sentences were used primarily to study the role of prosody in disambiguating different syntactic structures, but the corpus has also been useful in analyzing the relationship between duration changes and the different prosodic markers in our prosodic labeling system [34].

3 Utterance Annotation

Each story read by an announcer has been digitized in paragraph size units, which typically include several sentences. The files are digitized at a 16kHz sample rate using a 16 bit A/D. The paragraphs are annotated with the orthographic transcription, phonetic alignments, part-of-speech tags, and prosodic labels. The orthographic transcriptions were generated by hand and include indication of where the speaker took a breath. The phonetic alignments are generated automatically using constrained speech recognition, as described below, for the subset of data considered 'clean'. The part-of-speech tags are also generated automatically, as described in this section. Both phonetic alignments and part-of-speech tags are hand-corrected for the designated test data, and in a subset of the main corpus as well. The prosodic labels are marked by hand and are available only for a subset of the corpus, though our goal is to annotate the entire corpus eventually. In addition, a subset of the corpus is included in the Penn Treebank and therefore also has syntactic bracketings available.

Phonetic Alignment

The phone labels and segmentations are generated automatically using a recognizer with a grammar constrained to the orthographic transcription of the sentence, specifically the Boston University recognition system based on the stochastic segment model, as described in [12]. The steps involved in phonetic alignment are outlined below, followed by an analysis of alignment errors. The phone labels are based on the TIMIT phonetic labeling system, and are described in documentation associated with the corpus. Separate models are used for lexically stressed vs. unstressed vowels; vowels are marked as 'lexically stressed' (indicated by '+1' associated with the vowel label) when allowed by the lexicon and recognized as such during segmentation, as described below. Segmentation times and phone durations are provided in units of 10-msec frames.

The first step in obtaining phonetic alignments involves generating a pronunciation network for each word in the paragraph to allow for multiple pronunciations of words. Baseform pronunciations are obtained from a large on-line dictionary, which is augmented to include new words as needed. The dictionary is a derivative of the commercially available MOBY dictionary, which has been augmented with the SRI dictionary and includes some corrections and other additions to the original. The dictionary currently contains over 115,000 single-word entries, and approximately 3-5 new words are added per story (that is, dictionary coverage of new radio text is 99%), with over half of the new words being proper names. For words that have different pronunciations corresponding to different part-of-speech usage, the dictionary flags the pronunciation by part-of-speech (noun, verb, adjective, interjection), and the appropriate pronunciation is chosen according to the part-of-speech annotation of the text.

The baseforms are expanded into a network according to a set of phonological rules operating within words and optional silences are allowed between words. In addition, alternate pronunciations listed in the dictionary are incorporated into the pronunciation network. The pronunciation rules for vowels include: optional stressed and unstressed models (parallel paths) for syllables marked with secondary lexical stress and for one-syllable words; optional reduction of unstressed vowels /ah/, /eh/ and /ih/ to schwa /ax/; optional syllabic consonant (e.g. /ax l/ in parallel with /el/); and different forms of /er/ allowed (/ah r/, /er/, /ax r/). The pronunciation rules for consonants include rules for unreleased stops, deleted stops and flapping; rules for nasal flapping and nasal place of articulation change; and optional voiced /hv/ when /hh/ occurs between sonorants. No rules were added to accommodate r-less dialects, as it was not necessary for this group of talkers. These rules were chosen to allow a variety of pronunciations without too much overgeneration, based in part on rules described for recognition in [5] and in part on errors observed in the initial segmentation of a subset of the data. In all but the final pass of segmentation, the phonological rules were not applied across word boundaries. In the final pass, cross-word boundary rules were used to expand the set of allowable pronunciations, and flagged in the resulting alignments when used. The set of cross-word boundary rules cover palatalization, flapping, and consonant mergers. Although it would be desirable to expand the rule set to account for more types of phonological variation, particularly for phonetic alignment of more general data, the current rule set provided high quality alignments for this task, as described further in the error analysis.

In order for the recognizer to handle the fairly long, paragraph-sized files of speech, we reduced computation by using automatic breath detection to break the speech into more manageable units. Breath detection was based on a 3-class model of breaths, speech and silence. Each of three classes was associated with a multivariate Gaussian model of a vector that included cepstra and derivative cepstral features, and the sequence of class labels was assumed to form a Markov chain. Speaker-dependent models were trained on a subset of the data that was hand-labeled for breath locations, with a heuristic for determining the locations of silences. Breath detection was implemented by using Viterbi decoding (dynamic programming) to find the most likely sequence of labels from the classes of 'breath', 'speech' and 'silence'. A run of at least four consecutive 10-msec frames labeled with 'breath' is recognized as a detected breath. (This threshold was chosen based on experiments with speaker F2B.) The number of detected breaths is compared with the number of breaths marked by a human listener during orthographic transcription, and the detected breaths are hand-corrected when the numbers do not match. The start and end times of the breath are used to constrain the recognition search.

High quality phonetic alignments are obtained for this corpus by using several iterations of training and resegmentation. Forty-dimensional feature vectors of cepstra and derivative cepstra are used, computed from a 20 ms window at 10 ms time intervals. The initial phonetic models are trained using the speaker-independent TIMIT corpus. (The TIMIT corpus does not annotate lexical stress, so an approximate annotation was developed based on a comparison between TIMIT pronunciations and the stress markings from dictionary baseforms.) Using the TIMIT models, all of the 'clean' speech from a single speaker is segmented with no constraints other than the breath endpoints and minimum/maximum phone duration constraints. Next, new speaker-dependent phone models are estimated from the segmented radio data. The segmentation step is now repeated with the new models, but with constraints on the new phone boundaries to fall within $\pm N$ frames. The phone boundary constraints greatly reduce the computation time required for resegmentation. The segmentation and retraining steps are iterated three or four times on the clean speech, with a reduction in the window size N for allowable phone times. In the final pass of segmentation, phonological rules are applied to allow for cross-word boundary pronunciation changes, as mentioned previously. Eventually, we hope to use the models trained on the clean speech to align the noisy speech, but our current algorithms do not provide sufficient accuracy for much of the noisy data.

The automatic segmentation algorithm used in this work is similar to other segmentation algorithms recently proposed, e.g. [15, 33, 31, 16], with the exception that our acoustic models use the stochastic segment model instead of a hidden Markov model. Like [33], the problem in our case is complicated by the length of the speech files, though the single-talker read speech in our case is less challenging than their casual conversational speech. Unlike most previous work, we are able to take advantage of the speaker-dependent nature of the corpus and iteratively re-estimate models to obtain higher quality alignments.

Part-of-Speech Tags

The part-of-speech tags used in this corpus are the same as those used in the Penn Treebank [17]. This tag set includes 47 parts-of-speech: 22 open class categories, 14 closed class categories and 11 punctuation labels. The 36 word categories are summarized in Table 3 for reference.

Part-of-speech labeling is carried out automatically using the BBN tagger [24]. POST is similar to other probabilistic taggers (e.g. [6, 13]), but has an improved mechanism for handling unknown words. The tagger uses an bigram model[2] of the tag sequence and a probability of tag given word taken from either a dictionary or, in the case of an unknown word, based on features of the word related to endings, capitalization and hyphenation. The version of POST used in this work was trained on a set of Wall Street Journal sentences that formed part of the Penn Treebank corpus.

The part-of-speech tags on the lab news subset of the corpus were hand-corrected, in part to assess the accuracy of the algorithm on corpus. The hand-corrections were based

[2]Although the accuracy of the tagger can be improved slightly with a trigram model, the improvement is not warranted given the significantly higher computational cost.

CC	coordinating conjunction	PP$	possessive pronoun
CD	cardinal number	RB	adverb
DT	determiner	RBR	adverb, comparative
EX	existential *there*	RBS	adverb, superlative
FW	foreign word	RP	particle
IN	preposition/subordinating conjunction	SYM	mathematical or scientific symbol
JJ	adjective	TO	*to*
JJR	adjective, comparative	UH	interjection
JJS	adjective, superlative	VB	verb, base form
LS	list item marker	VBD	verb, past tense
MD	modal	VBG	verb, gerund or present participle
NN	noun, singular or mass	VBN	verb, past participle
NNS	noun, plural	VBP	verb, non-3rd person singular present
NP	proper noun, singular	VBZ	verb, 3rd person singular present
NPS	proper noun, plural	WDT	wh-determiner
PDT	pre-determiner	WP	wh-pronoun
POS	possessive ending	WP$	possessive wh-pronoun
PP	personal pronoun	WRB	wh-adverb

Table 3: Part-of-speech tags (excluding 11 punctuation labels) used in labeling this corpus, from the Penn Treebank set.

on the guidelines for Penn Treebank annotators. For the labnews stories, we found that 2% of the words were incorrectly labeled, out of a subset of 2133 words in the four stories from the test data that were hand-corrected. This error rate is relatively low relative to other reported results. For comparison, BBN reports error rates of 3-4% on known words and 15% on unknown words on Wall Street Journal sentences (outside of the training data), and an error rate of 8% in tagging data from a different domain [24]. The error rates for the Penn Treebank corpus are 7% for automatically generated labels (using Church's tagger [6] which was trained for a different tagset) and 4% for hand-corrected tags.

Prosodic Labels

The prosodic labeling system represents prosodic phrasing, phrasal prominence and boundary tones, using the ToBI system [27, 3, 19] for American English. The ToBI system represents prosodic phrase on a phrase break tier and tonal structure (accents and phrase tones) on a tone tier. Phrase break indices are used to express the degree of decoupling between each pair of words as follows: 0 - tighter connection than for a default word boundary, typically marked with phonetic modifications (e.g., palatalization), 1 - normal word boundary, 2 - boundary marking a lower-level perceived grouping of words that generally does not have an intonational boundary marker, 3 - intermediate phrase boundary, and 4 - intonational phrase boundary. Seven types of accent tones are labeled, corresponding to a simplified version of the system described in [4]: H*, !H*, L+H*, L+!H*, L*, L*+H and H+!H*, where H and L correspond to high and low targets, "!" indicates downstep and the asterisk indicates tone alignment. Intermediate phrase boundaries are marked with three tones (L-, !H- and H-), and intonational phrase boundaries are marked with with one of these tones plus a phrase boundary tone (L% or H%). Uncertainty markers are available to the labelers but are used rarely. For many of the files, we use an augmented version of the ToBI system that includes a sub-division of the intonational phrase category that distinguishes sentence boundaries (6) and within-sentence groupings of intonational phrases (5) from regular intonation phrase boundaries (4). A more extensive discussion of the mapping between the break labels and the prosodic constituents described in the literature is found in [34].

The ToBI system has the advantage that it can be used consistently by labelers for a variety of styles [19]. In our own study of labeler consistency on a set of three stories containing 1002 words, we found agreement on presence versus absence for 91% of the words. On those 487 words that were marked by both labeling groups with an accent, there was 60% agreement on accent type with most of the disagreements occurring for the difficult L+H* versus H* distinction. When the H*'s were grouped together with the L+H*'s as in [19], there was 81% agreement on pitch accent type. Boundary tone agreement was 93% for the 207 words marked by both labelers with an intonational phrase boundary, and similarly there was 91% agreement for 280 phrase accents. Agreement for the five ToBI break index levels was within the uncertainty level[3] for 95% of 989 words (excluding trivial paragraph-final cases). These results are higher than that reported by Pitrelli *et al.*,[19] in a general study of the ToBI labeling system, in part because the radio style has more clearly marked prosodic structure.

It is possible that the human labelers marking break indices are biased by syntax in some cases, but we feel that this is not a serious problem. The break levels 4-6 are marked by tonal cues, which are typically fairly easy to spot. The break levels 1-4 correlate well with duration measures, having statistically significant differences in mean normalized duration in phrase final syllables [34]. In addition, the investigations by Collier et al. [7] in a similar labeling task in Dutch found that perceived boundary strengths for normal speech and speech that had been "delexicalized" (using signal processing that preserved the intonation contour but otherwise rendered the speech unintelligible) were similar when averaged across subjects. Finally, we note that there are some occurrences of phrase boundaries in our corpus that do not coincide with syntactic boundaries, so the labelers seem to be able to factor out syntax

[3]Agreement "within the uncertainty level" merges neighboring classes when the uncertainty diacritic is used: (3,4-) and (3,3-) for example.

at least some of the time.

Only a portion of the corpus is currently labeled with prosodic markers, including all of the clean F2B data, a quarter of the M1B data, and the labnews stories. However, we consider these labels to be preliminary as we are still in the process of learning the ToBI system. Preliminary analyses on these two speakers show significant differences in their strategies for prosodic marking, with F2B using phrase boundaries, pitch accents and low phrase tones much more frequently than M1B. For both speakers, the high pitch accent (H*) is used frequently, i.e. on over half of the accents. After H*, the most frequent accent types are down-stepped high (!H*) and bitonal high (L+H*) accents.

Hand labeling of prosodic markers is a fairly time-consuming and therefore costly procedure. As a consequence, it is our goal to eventually automate or partially automate this procedure. An algorithm for automatically recognizing this set of prosodic labels given phonetic alignments has been described by Wightman and Ostendorf [35]. Though the performance is not as good as for human labelers, it may improve prosodic labeling significantly to add the step of automatic detection, followed by hand-correction. (In the Penn Treebank part-of-speech tagging experiments, hand-corrections improved over fully hand-labeled data in speed, accuracy and consistency [17].) In addition, the prosodic labeling work is relatively new, and we therefore expect further improvements in accuracy in the future.

4 Summary

We have described a corpus of over seven hours of professionally read radio news speech. The corpus includes orthographic transcription, phonetic alignments, part-of-speech tags, and prosodic notation. We hope that the availability of this corpus will inspire further research on prosody, encourage the further development of automatically trained algorithms, and allow for more direct comparison of research results more generally, but for speech synthesis in particular.

Although we designed the corpus because of our interests in speech synthesis applications, it is suitable for many other applications and analyses. We have used the corpus ourselves in studies of prosody in syntactic disambiguation [21, 18], early accent placement [26], duration lengthening [34] and glottalization [8]. It is also suitable for general studies of prosody and its interactions with syntax and discourse.

Acknowledgments

The authors gratefully acknowledge WBUR and all the radio announcers who cooperated in this study. We also thank several people who helped with data collection and labeling, including John Butzberger, Nanette Veilleux, Colin Wightman, Ken Ross, Cynthia Fong, Mark Paley, Doug Watson, Yeong Chang, Susan Zlotkin, John Kaufhold, Navneet Mathur, Suk Choi, Hatal Patel, Gay Ping Lau, Ahwat Schlosser, Loretta Hawks, Cameron Fordyce, Lisa Kwok, Liz Shriberg, Gay Baldwin, Laura Dilley, Luis Santiago, and Rachel Molenaar. The corpus collection, annotation and analyses were funded in part by NSF (grant number IRI-8805680), Apple Computer, and the Linguistics Data Consortium.

References

[1] B. Altenberg (1987) *Prosodic Patterns in Spoken English,* Lund University Press, Sweden.

[2] J. Baker, and D. Paul (1992) "The Design for the Wall Street Journal-Based CSR Corpus," *Proc. ARPA Speech and Natural Language Workshop,* 357-362.

[3] M. Beckman and G. Ayers (1994) "Guidelines for ToBI Labeling, version 2.0", Manuscript and accompanying speech materials [Obtain by writing to tobi@ling.ohio-state.edu].

[4] M. Beckman and J. Pierrehumbert (1986) "Intonational structure in Japanese and English," *Phonology Yearbook 3,* ed. J. Ohala, 255-309.

[5] M. Cohen (1989) *Phonological Structures for Speech Recognition,* Ph.D. Dissertation, U.C. Berkeley Department of Computer Science.

[6] K. Church (1988) "Stochastic Parts Program and Noun Phrase Parse for Unrestricted Text," *Proceedings of the Second Conference on Applied Natural Language Processing,* 136-143.

[7] R. Collier, J. R. de Pijper and A. Sanderman, (1993) "Perceived Prosodic Boundaries and their Phonetic Correlates," *Proceedings of the ARPA Human Language Technology Workshop*, L. Bates, Ed.

[8] L. Dilley, S. Shattuck-Hufnagel and M. Ostendorf, "Glottalization of Vowel-Initial Syllables as a Function of Prosodic Structure," manuscript submitted to *Journal of Phonetics.*

[9] J. Edwards and M. Lampert (1993) *Talking Data: Transcription and Coding in Discourse Research,* Erlbaum.

[10] H. Fujisaki, K. Hirose, N. Takahachi, and H. Morikawa (1986) "Acoustic characteristics and the underlying rules of intonation of the common Japanese used by radio and television announcers," *Proceedings of the International Conference on Acoustics, Speech and Signal Processing,* 2039-2042.

[11] J. J. Godfrey, E. C. Holliman and J. McDaniel (1992) "SWITCHBOARD: Telephone Speech Corpus for Research and Development," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing,* , San Francisco, March 23-26, Vol. 1, 517-520. Documentation also on the CDRom available through LDC.

[12] O. Kimball, M. Ostendorf and I. Bechwati (1992) "Context Modeling with the Stochastic Segment Model," *IEEE Transactions on Signal Processing,* 1584-1587.

[13] J. Kupiec (1989) "Augmenting a Hidden Markov Model for Phrase-Dependent Word Tagging," *Proceedings of the DARPA Speech and Natural Language Workshop,* 92-98.

[14] L. F. Lamel, R. H. Kassel and S. Seneff (1986) "Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus," *Proc. DARPA Speech Recognition Workshop,* Report No. SAIC-86/1546, 100-109.

[15] A. Ljolje and M. Riley (1991) "Automatic segmentation and labeling of speech," *Proceedings of the International Conference on Acoustics, Speech and Signal Processing,* 473-476.

[16] A. Ljolje, J. Hirschberg and J. P. H. van Santen (1994) "Automatic speech segmentation for concatenative inventory selection," *Conference Proc. of the ESCA/IEEE Workshop on Speech Synthesis,* pp. 93-96.

[17] M. P. Marcus and B. Santorini (1993) Building a Very Large Natural Language Corpora: The Penn Treebank. *Computational Linguistics,* 19(2) 313-330.

[18] M. Ostendorf, C. W. Wightman, and N. M. Veilleux (1993) "Parse Scoring with Prosodic Information: An Analysis/Synthesis Approach," *Computer Speech and Language*, 193-210.

[19] J. F. Pitrelli, M. Beckman and J. Hirschberg (1994) "Evaluation of Prosodic Transcription Labeling Reliability in the ToBI Framework", *Proceedings of the International Conference on Spoken Language Processing.*

[20] P. Price, M. Ostendorf, S. Shattuck-Hufnagel, and N. and Veilleux (1988) "A Methodology for Analyzing Prosody," *J. Acoust. Soc. Am.,* 84, Suppl. 1, S99.

[21] P. Price, M. Ostendorf, S. Shattuck-Hufnagel, and C. Fong, (1991) "The use of prosody in syntactic disambiguation," *Journal of the Acoustical Society of America, 90,* 6, 2956-2970.

[22] MADCOW (L. Hirschman et al.) (1992) "Multi-Site Data Collection for a Spoken Language Corpus," in *Proceedings Speech and Natural Language Workshop*, Harriman, New York, 7-14 (Morgan Kaufman Publishers, Inc., San Mateo, California).

[23] P. J. Price, W. M. Fisher and J. Bernstein (1988) "The DARPA 1000-Word Resource Management Database for Continuous Speech Recognition," D. S. Pallett, *Proceedings of the International Conference on Acoustics, Speech and Signal Processing,* Database available on CD-ROM via the LDC.

[24] M. Meteer, R. Schwartz and R. Weischedel (1991) "POST: Using Probabilities in Language Processing," *Proceedings of the International Joint Conference on Artificial Intelligence.*

[25] K. Ross and M. Ostendorf, "Prediction of Abstract Prosodic Labels for Speech Synthesis," submitted manuscript.

[26] S. Shattuck-Hufnagel, M. Ostendorf and K. Ross, (1995) "Pitch Accent Placement within Lexical Items in American English," *Journal of Phonetics.*

[27] K. Silverman, *et al.* (1992) "TOBI: A standard for labeling prosody," *Proceedings of the International Conference on Spoken Language Processing,* 867-870, 1992.

[28] C. Sorin, D. Larreur, and R. Llorca (1987) "A rhythm-based prosodic parser for text-to-speech systems in French," in *Proceedings of the International Congress of Phonetic Sciences*, Vol. 1, 125-128, Tallinn.

[29] E. Strangert (1991) "Pausing in texts read aloud," in *Proceedings, XII International Congress of Phonetic Sciences*, Vol. 4, 238-241.

[30] J. Svartvik (Ed.) (1990) *The London-Lund Corpus of SPoken English: Description and Research*, Lund University Press.

[31] D. Talkin and C. W. Wightman (1994) "The Aligner: Text to speech alignment using Markov models and a pronunciation dictionary," *Conference Proc. of the ESCA/IEEE Workshop on Speech Synthesis,* pp. 89-92.

[32] V. Vihanta (1991) "Signalisation prosodique de la structure informationnelle dans le discours radio-phonique en Finnois et en Francais," in *Proceedings, XII International Congress of Phonetic Sciences*, Vol. 2, 422-425.

[33] B. Wheatley, G. Doddington, C. Hemphill, J. Godfrey, E. Holliman, J. McDaniel and D. Fisher (1992) "Robust Automatic Time Alignment of Orthographic Transcriptions with Unconstrained Speech," *Proceedings of the International Conference on Acoustics, Speech and Signal Processing,* I533-I536.

[34] C. W. Wightman, S. Shattuck-Hufnagel, M. Ostendorf, and P. Price (1992) "Segmental Durations in the Vicinity of Prosodic Phrase Boundaries," *Journal of the Acoustical Society of America,* Vol. 91, No. 3, 1707-1717.

[35] C. Wightman and M. Ostendorf (1994) "Automatic Labeling of Prosodic Patterns," *IEEE Trans. Speech and Audio Processing,* Vol. 2, No. 4, 469-481.